UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Department of Language and Culture (ISK)

When grammar can't be trusted -

Valency and semantic categories in North Sámi syntactic analysis and error detection

—

Linda Wiechetek

A dissertation for the degree of Philosophiae Doctor – December 2017

When grammar can't be trusted -

Valency and semantic categories in North Sámi syntactic analysis and error detection

Linda Wiechetek

A dissertation for the degree of Philosophiae Doctor
Department of Language and Culture (ISK)

December 2017

Für meine Eltern
&
buot sámegiela oahpahalliide

Contents

List of Figures

List of Tables

Glossary

Giella-sme North Sámi pipeline of morpho-syntactic analyzers. ix, 47, 49, 51–54, 128–130, 132, 135, 136, 138, 145, 154, 162, 171, 184, 295–297

GoDivvun the North Sámi grammar checker *Giellaoahpa Divvun*. 9, 10, 46, 48, 52, 127, 177–179, 184–188, 191, 261, 262, 265, 270, 275, 284, 285, 288, 295–297

SIKOR the Sámi International Corpus compiled by UiT The Arctic University of Norway and the Norwegian Sámi Parliament. xi, 46, 47, 49, 63, 64, 70–72, 75, 76, 101–103, 108, 110, 112, 114, 120, 123, 140, 142, 144, 157, 162, 163, 166, 169, 178, 188, 194, 206, 208, 210, 211, 213, 214, 217, 219, 222, 227, 229, 236, 243, 251–254, 262, 268, 283, 284, 296, 297

abbreviations.lexc the North Sámi lexc abbreviation lexicon. 51

acronyms.lexc the North Sámi lexc acronym lexicon. 51

adjectives.lexc the North Sámi lexc adjective lexicon. 51, 138, 147, 153

adverbs.lexc the North Sámi lexc adverb lexicon. 51, 153

analyser-gt-norm.hfstol the Divvun North Sámi normative Helsinki finite state compiler and morphological analyzer. 188

dependency.cg3 the Giellatekno North Sámi CG dependency grammar. 234

disambiguation.cg3 the Giellatekno North Sámi CG disambiguator and syntactic analyzer. 52, 201, 202

disambiguator.cg3 the Divvun North Sámi CG disambiguator for grammar checking. 51, 52, 184, 202, 231, 232, 261, 263, 270, 277, 285, 295, 297

generator-gt-norm.hfstol the Divvun North Sámi normative Helsinki finite state compiler and morphological generator. 185, 245, 261, 286

grammarchecker.cg3 the Divvun North Sámi CG grammar checker. 51, 184, 185, 190, 191, 204, 205, 207, 211, 234, 239–242, 246, 249, 251, 253, 260–262, 267, 268, 270, 284, 286, 297

mwe-dis.cg3 the Divvun North Sámi CG disambiguator of multi-word expressions. 185, 261, 285, 297

nouns.lexc the North Sámi lexc noun lexicon. 51, 127, 138, 145, 147, 154, 162, 163, 170

propernouns.lexc the North Sámi lexc proper noun lexicon. 51, 128, 138, 147, 153

root.lexc the North Sámi lexc root lexicon. 52

tokeniser-gramcheck-gt-desc.pmhfst the Divvun North Sámi descriptive Helsinki finite state compiler and morphological analyzer. 184, 261, 285

valency.cg3 the North Sámi CG valency annotation grammar. xi, 39, 46, 51, 63, 64, 68–72, 79–85, 88–92, 96–98, 100, 101, 103, 106, 107, 109, 114, 117–124, 206, 243, 261, 269, 270, 283–285, 296

verbs.lexc the North Sámi lexc verb lexicon. 33

Abbreviations

1du first person dual form.

1pl first person plural form.

1sg first person singular form.

2du second person dual form.

2pl second person plural form.

2sg second person singular form.

3du third person dual form.

3pl third person plural form.

3sg third person singular form.

abl ablative case.

abs absolutive case.

abs3pl auxiliary agreeing with an absolutive third person plural form.

abs3sg auxiliary agreeing with an absolutive third person singular form.

acc accusative case.

actio actio form.

all allative case.

attr attributive form.

aux auxiliary.

car caritative derivation.

caus causative derivation.

CG Constraint Grammar formalism.

com comitative case.

conneg connegative form.

connegII biblical connegative form.

CORR corrected.

dat dative case.

dat1sg auxiliary agreeing with a dative first person singular form.

denom denominal derivation.

du dual.

erg ergative case.

erg3sg auxiliary agreeing with an ergative third person singular form.

err/orth-nom-acc nominative form that should be an accusative form.

ess essive case.

ex. example.

foc focus.

freq frequentative derivation.

fst finite state transducer.

gen genitive case.

hfst Helsinki finite-state tool.

ill illative case.

imprt imperative.

inch inchoative derivation.

ine inessive case.

inf infinitive.

instr instrumental case.

L.W. Linda Wiechetek.

lexc lexicon formalism for machine-readable lexica designed by Xerox, and a compiler with the same name that turns the lexicon into a fst.

loc locative case.

lookup2cg a perl script that reformats the lookup output so that it can be interpreted as input to vislcg3.

mwv multi-word verb.

NLP natural language processing.

nom nominative case.

p.k. personal knowledge.

pass passive derivation.

pcle particle.

pl plural.

PoS part of speech.

pot potential.

prfprc past participle.

prs present tense.

prsprc present participle.

prt past tense.

pxdu1 first person dual possessive form.

pxdu2 second person dual possessive form.

pxdu3 third person dual possessive form.

pxpl1 first person plural possessive form.

pxpl2 second person plural possessive form.

pxpl3 third person plural possessive form.

pxsg1 first person singular possessive form.

pxsg2 second person singular possessive form.

pxsg3 third person singular possessive form.

q question particle.

recip reciprocal derivation.

refl reflexive derivation.

sb. somebody.

sg singular.

sth. something.

twolc two level compiler.

vislcg3 *visl* constraint grammar compiler (version 3).

xfst Xerox finite-state tool.

Acknowledgments

After submitting version 10171, the "final" version, of *phdlindawiechetek.pdf* to the svn, I would like to thank a few people that are co-responsible for there being a final version. In many ways, this dissertation has felt like an ultra-marathon, and although I have never run more than 20 km, I assume this must be how it feels, a looong way to go with physical challenges, but most of all psychological ones. During this race, my supervisor Trond Trosterud has been there at all supply stations to hand me the metaphorical water bottle with advice and support. His support never came at the cost of my freedom, and I particularly value that he always encouraged me to work independently. Hvis man snakker med Trond, så kan man være sikker på at han er 100% fokusert på det man sier og at man får gjennomtenkte svar uansett tema (muligens med 3 on-the-fly forkortelser i en setning ;)). Takk for å være en så takknemlig leser av det æ skriv. Æ vet ikkje om det er så mange som ville sagt at dem syns det er "spennandes" når dem leser et nytt PhD-kapittel. :D Etter en samtale med dæ hadde æ alltid lyst å skrive mer. Det er ikke en selvfølge, og æ setter stor pris på dæ som Doktorvater.

Secondly, I am grateful to my informants who went through hundreds, maybe thousands, of sentences to check valencies. We had long inspiring discussions about these and it was cool to see that someone else could be enthusiastic about the things I choose to work with.

I would also like to thank my colleagues from Giellatekno and Divvun (Biret Ánne, Biret Merete, Børre, Chiara, Ciprian, Duommá, Elena, Ilona, Inga, Kevin, Lene, Maja, Ritva, Saara, Sandra, Sjur, Trond and Tomi) who actively work(ed) with the language technology infrastructure I use in my dissertation and without whom there wouldn't even be a PhD position in Sámi language technology. They were welcoming when I first entered the group in 2005, patient when I messed up the compilation introducing semantic tags to the analyzers and they have been a pleasure to work with.

Sometimes people would ask me why I was laughing so much at work during my Skype meetings. This is partly Duommá's fault, who also thinks that grammatical errors, code switching and language innovations can be very amusing. Apart from contributing with his linguistic expertise and language intuitions about grammatical phenomena, he added hundreds of semantic tags to the lexicon. He also co-founded the life-saving activity group for nordic skiing, bouldering ("gupman" or if we go with Sjur's suggestion "kampesteine"),

climbing, running, and randonnee skiing.

Ritva Nystad was another life-saver in my writing process. Not only did we share an office, but also the daily struggles of a PhD student. Giitu oktasaš bottuid, lingvisttalaš ságastallamiid, movttiidahttima ja spontána meahccetuvrraid ovddas! Lene Antonsen was there for me with advice and emotional and practical support when I needed it. Thank you also for being so generous with your dinner invitations, ja giitu go mun ja bussát beasaimet orrut din luhtte nu ollu.

I'd also like to mention my "almost second author", Børre Gaup, who tried to commit changes to a subversion of my dissertation. (Unfortunately, I had to revert them :D) I owe him many thanks for his endless patience with my repetitive questions about certain Unix commands (where others might have just answered RTFM) and his help when calculating the numbers for some of my tables. I also want to thank Sjur Moshagen for being so flexible and supportive with regard to my PhD while I was working part time for Divvun. Also for approaching me with the grammar checking project just at the right time.

I am grateful for Jussi Ylikoski's thoroughness, expertise and valuable comments on the content and form of my dissertation. He had a sharp eye for errors and inconsistencies no one else saw and was ready to help whenever I had a question.

Molly Bechert's comments on idiomaticity, format, and English spelling and grammar were of huge help to me, and she even pointed out inconsistencies in Spanish, Basque and Sámi! I am grateful to Lars Borin for his comments on a draft of Chapter 4 and to Eckhard Bick for making the semantic analysis of his online analyzers available and being an inspiration within Constraint Grammar.

I could not have done without Tino Didriksen, who added features to the CG formalism on request and who I could count on for instant help with non-functioning rules. Francis Tyers was incredibly helpful and motivated when setting up Apertium for me and working on sme-smj machine translation with me. Although I didn't end up writing about it, this was an exciting exploration (and not my last one) into the field of machine translation accompanied by really cool conferences, exciting conversations and ska concerts. He also told me about the NILS mobility project from Universidad Complutense de Madrid, thanks to which I spent 3 months++ in Donostia. The stay was a huge inspiration to start developing the valency annotation grammar.

I would like to thank the members of Basque language technology group IXA at Euskal Herriko Unibertsitatea, who were extremely welcoming and helpful (Kepa Sarasola and Amaia Lorenzo were great with all the administrative work and the application process, and Ruben Urizar lent me his bike), and who were ready to discuss exciting things within Constraint Grammar (Jose Mari Arriola), valency/semantic role annotation (Ainara Estarrona and Izaskun Aldezabal), and dependency annotation (Maxux Aranzabe) with me. I really enjoyed our lunches with the people from Korta, and no one ever complained when I asked tons of questions about Basque grammar, history and culture :) Eskerrik asko

denei! Zorte ederra izan nuen zuekin euskera ikasteko abagunea izan nuenean. Eskerrik asko Antiguoko AEKko jendeei, hargatik! Not only did Basque morpho-syntax let me see North Sámi grammar in a different light, it also made me excited again about the immense variation in thinking (and expressing these thoughts in language) in this world. Giitu maid mu davvisámegiela oahpaheddjiide, Laila Susanne Oskarsson, Hans Herman Bartens ja Lene Antonsen!

It has been a pleasure to share this process with my fellow PhD students, with whom I shared many lunch breaks and who often kept my mind from agonizing. They would also kindly open the door for me when I had locked myself out in the evenings (especially Elli). I am also grateful to those that provided a balance to writing and reading, including TSI fjellgruppa, TSI Aikido, Laura Castor's yoga classes, Tromsø klatreklubb, and Aikido Amagoia.

Thanks to Jan Helge Bergheim for saving all the data on my computer twice, and thanks to Geir Tore Voktor for suggesting that I buy a huge external hard disk as a backup! Linda Nesby and Mayvi Johansen for their valuable tips (for eksempel angående sluttlesing og språkvask) and their administrative help. Giitu *Sámi fágajovkui* buriid kommentáraid ovddas ja *Giellagáldui* miellagiddevaš ságastallamiid ovddas! Thank you to Marwin for helping with a latex command that was driving me crazy.

Thanks also to a number of people who have believed in me and supported me during my studies. Amongst those are:

Herbert Gerke, my Latin and Spanish teacher, who taught me to be enthusiastic about languages and art! Howard Gregory, my supervisor in Göttingen, who was the first one that suggested I could publish an article and encouraged me to aspire a PhD.

Thank you to Raúl for reading through and commenting on some of my drafts and also for choosing Sállir over Uriellu. Thank you to Riddu and Ginttal for being there. Thank you to my friends for making my life richer. And thank you to my parents for your love and support throughout my life.

De giitu vel Sállirii čáppa ovdasiiddu ja ollu fiinna tuvrraid ovddas!

Part I

Beginning

A whole is what has a beginning and middle and end. (Aristotle, 1932)

Chapter 1

Introduction

As I like to learn new languages, I have taken part in a number of beginner's language courses. Second language learning includes challenges to both lexicon and grammar learning. But what I find most challenging are the phenomena that fall into the gap between lexicon and grammar, like verb valency, i.e. the number and form of the arguments of a verb. These phenomena are typically not grammatical enough to be taught by means of explicit grammar rules but cannot be inferred via common-sense semantics either. Without formalized valency competence, I – as a native German speaker – am likely to translate sentence (1-a) into (1-b) (North Sámi), (1-c) (Basque) and (1-d) (Polish) based on my German intuitions leaving my conversation partner either confused or amused. None of the examples follow the valency rules of the respective language. The correct realization of the argument in the respective language is given in brackets.

(1) a. Ich freue mich über das Geschenk.
I am.happy myself about the.ACC gift.ACC
'I am happy about the gift.'

b. Mun illudan *skeaŋkka birra (*correct:* skeaŋkkas/skeaŋkka dihte).
I am.happy gift about (*correct:* gift.LOC/gift.GEN because)

c. *Opariari buruz (*correct:* Opariak) ilusioa egin
gift.DAT about (*correct:* gift.ERG) happiness.ABS make
dit.
AUX.ABS3SG.DAT1SG.ERG3SG

d. Cieszę się *nad prezentem (*correct:* z prezentu).
be.happy me about gift.INSTR (*correct:* of gift.GEN)

In language teaching, valencies are typically either directly translated into the instruction language or explained by means of common-sense semantics. Literal translation only works if valency structures are parallel in the language taught and in the instruction language. If the instruction language is not your native language and you do not have strong intuitions, this method of instruction will not work for you. Common-sense semantics, on the other hand, are often inherent in the language and can only be applied when you have a certain competence in the language already.

While intuition is important when producing correct sentences in one's own language, formalized valency knowledge is necessary for the translation and production of correct sentences when learning a second language, cf. Tesnière (1959, Chapter 122, §8)[1] and Helbig and Schenkel (1973, p.11).[2] Both a language learner and a machine-readable grammar need access to formalized valency information to understand/analyze and produce a sentence in a foreign language.

In this dissertation, I discuss and develop natural language processing tools that use explicit grammar rules to model human grammatical knowledge. This dissertation has come into being in the context of the language technology groups *Giellatekno* and *Divvun* at *UiT (Norges arktiske universitet)*. Both groups work on linguistic and computational research in Sámi (e.g. North Sámi, South Sámi, Inari Sámi, etc.) and other morphologically-rich languages (e.g. Faroese, Icelandic, Iñupiaq, Romanian, Inuktitut, Somali, etc.). They also focus on the development of rule-based language technological tools, such as syntactic parsing, spell-checking and grammar checking, machine translation, pedagogical tools, electronic dictionaries and text-to-speech applications. These tools incorporate a rule-based (as opposed to statistic) analysis of the language in question, starting in a bottom-up manner with a morphological analyzer and lexicon followed by a syntactic analysis. They are designed to enable minority language societies to use their language in modern devices and in official contexts, which is essential for the survival of a language in modern society. The infrastructure and architecture of the North Sámi system including various analyzers will be referred to as *Giella-sme* here.

The Sámi languages belong to the Uralic languages and are spoken in the North of Norway, Finland, Sweden, and northwestern Russia. There are nine Sámi languages, of which North Sámi is the language with the largest group of speakers, found in Norway, Finland and Sweden – 25,700 speakers according to *Ethnologue* (Simons and Fennig, 2017). All Sámi languages are morphologically complex and different parts of their grammars show agglutinative and fusional characteristics.

Within this dissertation, I develop three machine-readable grammars for North Sámi: a valency annotation grammar, a grammar for morpho-syntactic analysis and disambiguation, and a grammar for dependency annotation, semantic role annotation and syntactic error detection. In addition, I enhance the North Sámi lexicon with semantic prototype tags. While there are syntactic tools for North Sámi that have been developed prior to this

[1]"But it should not be forgotten that if one wishes to master a foreign language and be capable of for[e]seeing the inversions of actants that must take place prior to the translation of one language into another, it is necessary to have in-depth knowledge of the actant structure of verbs, as much in the source language as in the target language."

[2]"Es handelt sich um spezielle Fehler bei Ausländern, da der Muttersprachler in solchen Fällen auf Grund seines Sprachgefühls – seiner sprachlichen Kompetenz – die richtige Entscheidung zu treffen vermag. Ein solches unmittelbares Sprachgefühl fehlt aber dem Ausländer, und der Lektor war bisher meist nur in der Lage, auf Grund seiner linguistischen Intuitionen (aber nicht auf Grund eines bestimmten Regelmechanismus) dem Ausländer zu verdeutlichen, wann er etwa „wissen" oder „kennen", wann er „sagen", „sprechen" oder „reden" verwenden muß."

dissertation, these tools are based on the assumption that the linguistic input is grammatically correct, and they only make use of morphological and syntactic information (including dependencies). However, when we process a sentence, we use knowledge on other linguistic levels as well, e.g. semantics, valency, cultural, and discourse knowledge. In this dissertation, I attempt to fill some of the linguistic gaps in the existing resources and make the linguistic context of an analyzed token linguistically denser. That way, ambiguous and erroneous sentences can be parsed and valency-specific tasks (e.g. valency error detection) can be performed. Within the development process, valency tags and grammatical rules that make reference to valency tags are created simultaneously, which has the advantage that valency tags are functional with regard to their tasks, and they can be tested while being developed.

The main contribution of this dissertation is the integration of exhaustive valencies in a complex rule-based grammar for error detection and the development of an approach to detect global grammatical errors when the grammaticality of the input sentence, and therefore the analysis as well, cannot be trusted. This dissertation consists of both a linguistic study of North Sámi valency variation and the construction of rule-based grammars for North Sámi and their evaluation.

The text is framed by an introduction (Chapter 1) and a conclusion (Chapter 6). The introduction is followed by a chapter on theoretical background and methodology (Chapter 2). Chapter 2 focuses on establishing a general understanding of valency, in terms of both syntax and semantics, previous research on North Sámi valency, and valency in natural language processing.

The main part consists of three chapters, the first of which describes a valency grammar (Chapter 3), the second of which describes a semantic prototype resource (Chapter 4), and the last of which describes a grammar checking module (Chapter 5). Chapter 3 consists of a study of the valency variation in North Sámi and describes the valency grammar. I discuss different types of governors, including multi-word governors, and their valencies, as well as the impact of morphological processes on their valency potential. In this context, I also discuss the annotation of several valency frames to one governor in the case of synonymy, polysemy and diathesis alternations. In addition, I present the internal structure of the valency tags used in the grammar checker *GoDivvun* and their reference to semantic roles, morpho-syntax, and selection restrictions. Finally, I describe the architecture of the valency grammar and its rules, which are evaluated with regard to their coverage.

Chapter 4 presents a system of semantic prototype tags for North Sámi. I start out with some theoretical background for semantic prototypes, the technical background for the existing linguistic resources, and the objectives for including semantic prototypes in natural language processing. I then present a set of semantic prototype categories for North Sámi and describe the principles behind this set. I also address issues regarding the

implementation, i.e. multiple tags for one entry in the case of polysemy and homonymy, and the annotation of compounds. Lastly, I evaluate the distribution of semantic prototypes in four syntactically ambiguous constructions.

Chapter 5 deals with the integration of the resources presented in Chapters 3 and 4 in a grammar checker for North Sámi, *GoDivvun*. The grammar checker consists of a module for disambiguating potentially ungrammatical input and a module for error detection and correction, which also performs partial dependency annotation and semantic role annotation. The grammatical rules refer to valencies and semantic prototypes. Firstly, I describe previous approaches to rule-based grammar checking, focusing on global error detection in particular, and present the North Sámi infrastructure and an error typology for North Sámi. Secondly, I present the North Sámi grammar checker *GoDivvun* and show how valencies and semantic prototypes are integrated. Here, I distinguish between local and global error detection rules and focus on the latter. These are described in detail and their precision, recall and accuracy are evaluated. Finally, Chapter 6 draws conclusions based on the evaluations in the previous chapters and gives an outlook on future research.

Chapter 2

Background and methodology

In this chapter, I take up the general theoretical background of valency theory, and discuss its role within Sámi research and language technology. The subsequent chapters (Chapters 3–5), on the other hand, take up specific theoretical background (e.g. of semantic prototypes, valency resources and grammar checking). Additionally, I address methodological issues, present the framework of the resulting natural language processing tools and define key terms.

In the first section, I describe the origins of valency theory and its different linguistic dimensions. I also try to define what an argument is, what belongs to the valency of a particular governor, and what a governor is. My focus here is on syntactic valency, semantic valency (i.e. semantic roles) and semantic selection restrictions. In my discussion of syntactic valency, I address the role of obligatoriness and syntactic tests of argumenthood. In my discussion of semantic valency, I look at the formal basis for semantic role sets and their distinction from syntactic valency and referential semantics. Apart from different restrictions to arguments I also address the semantic grouping of governors (i.e. verb classes) and restrictions to potential governors. Secondly, I discuss the role of valency theory and descriptions of related concepts in previous Sámi research. Lastly, I look at valency resources in natural language processing and their use in specific tasks.

In the second section, which deals with the methodology and framework, I address introspection and corpus search as means to construct a valency database and a grammar checker (*GoDivvun - GiellaoahpaDivvun*). With regard to the grammar checker, I also describe normative questions and measures to evaluate my tools. This section includes a description of *Constraint Grammar*, the framework for the valency grammar, the semantic prototype resource and the grammar checker, and introduces the functionalities used throughout this work. Lastly, key concepts that are used in this work are defined.

2.1 Theoretical background: Valency theory

"Valency" is derived from the Latin noun *valentia* 'power, might, strength' and the verb *valere* 'possess, or have predominance in' Glare (1983, pp.207–208). It was originally used in the field of chemistry to describe the capacity of the atom to combine with a specific number of atoms, and later picked up by Lucien Tesnière as a metaphor to describe the capacity of the verb to combine with a specific number and type of arguments:

> The verb may therefore be compared to a sort of atom, susceptible to attracting a greater or lesser number of arguments, according to the number of bonds the verb has available to keep them as dependents. The number of bonds a verb has constitutes what we call the verb's valency. (Tesnière, 1959, Chapter 97, §3)

Lucien Tesnière is considered to be the founder of valency theory. He first mentioned the term in works that were written in 1953 and published posthumously in 1959. However, the term "syntactic valency" had been mentioned earlier by A.W. de Groot (1892-1963) in his work *Structurale Syntaxis*, written in Dutch (Groot, 1949).

Valency theory has evolved as a central part of dependency grammar and has had an impact on both theoretical linguistics and computational linguistics. Helbig and Schenkel (1973) and Tesnière (1959) stress the importance of valency in second language learning and translation because of "metataxis", i.e. the "structural change occurring during the transition from one language to another" (Tesnière, 1959). Valency is considered to be the ability of any lexeme (prototypically verbs, but also nouns, adjectives and adverbs) to combine with/attract/govern other lexemes in the sentence. Tesnière's (1959) valency theory is based on the assumption of verb centrality and equality of the co-occurring lexemes. Instead of splitting the sentence into subject and predicate, both subject and object are seen as equal dependents of the verb. Typically some of the co-occurring lexemes in a sentence are considered to be part of the governor's valency, while others are not. Tesnière (1959, Chapter 48) applies a theater metaphor when distinguishing between "actants", which are part of the verb's valency, and "circumstants", which are not:

> §4 The **actants** are the beings or things, of whatever sort these might be, that participate in the process, even as simple extras or in the most passive way. [...]
> §6 **Actants** are always nouns or the equivalents of *nouns*. In return, nouns in principle always assume the function of actants in the sentence.
> §7 **Circumstants** express the circumstances of time, place, manner, etc. in which the process unravels. [...]
> §8 **Circumstants** are always **adverbs** (of time, of place, of manner, etc. [...]) or the equivalents of adverbs. In return, adverbs in principle always assume the function of circumstants.

According to Tesnière, the first actant performs the action, the second actant supports the action, and the third actant receives benefit or detriment from the action. However, in the sentence *Alfred change de veste* 'Alfred changes his jacket', the semantic and syntactic criteria of "actants" and "circumstants" do not coincide. While the prepositional complement *de veste* is closely connected to the verb semantically and therefore an "actant", because of its morpho-syntactic properties, i.e. it being a prepositional complement, Tesnière classifies it as a "circumstant" (Tesnière, 1959, Chapter 7, §6-§7).

I therefore count Tesnière's (1959) approach among the morpho-syntactic approaches to valency as opposed to approaches where semantic (or other) criteria are given priority when determining argumenthood. Tesnière's (1959) definition of valency is rather restricted, and I will use a wider definition of valency. On the other end of this spectrum, there are approaches like Čech et al.'s (2010) "full valency" approach within natural processing. As they see fundamental weaknesses in the introspective method that is used to define valency membership, the authors include both arguments and free modifications in the valency of the verb and reject a distinction between them. In ex. (1), *father, books, to*, and *yesterday* are all considered to be part of the valency of the verb *give* as "they are direct dependents of the verb" (Čech et al., 2010, p.294). However, in their approach, there are no qualitative distinctions between the dependents of a governor, which is why their approach is more a dependency theory than a valency theory, given that valency is the government of dependents that are specific to the governor as opposed to those that are unspecific to it (Fischer, 1997, p.43). Čech et al. (2010) establish syntactic relations between governors and their dependents, i.e. dependencies, like Tesnière (1959, Chapter 2), but do not specify their semantic roles, obligatoriness, ability to appear with specific verbs and not with others, etc. However, the latter aspects of valency are relevant to my research, which is why I will also discard Čech et al.'s (2010) approach to valency in this work.

(1) **My father** gave **four books to Mary yesterday evening** (Čech et al., 2010)

In the following I will distinguish between several non-isomorphous, i.e. related but autonomous, levels of valency: syntactic valency, semantic valency and selection restrictions. Panevová (1994, p.224) distinguishes between "(1) morphemic case, (2) the meaning (function) of case (verbal valency), and (3) the cognitive roles of verbal participants", and Helbig and Schenkel (1973, p.65) distinguish between syntactic, semantic and logic valency. Helbig's (1992) extended valency model, on the other hand, consists of 6 levels: I - quantitative semantic valency structure, II - inherent semantic features of the verb, III - qualitative semantic roles, IV - inherent referential-semantic features of the arguments, V - syntactic functions and morphological realizations of the arguments, VI - quantitative representation of actants distinguishing between obligatory and facultative arguments (Helbig, 1992, pp.153–155). In addition, I will discuss the semantic categorization of

verbs according to their inherent features (cf. Helbig's (1992) level II) and restrictions to potential governors.

2.1.1 Syntactic valency

In this work, I will use the term *syntactic valency* to describe the morpho-syntactic realization of obligatory and facultative actants, cf. Helbig and Schenkel's (1973) syntactic valency and Helbig's (1992) levels V and VI. There is typically a distinction between arguments (cf. Tesnière's "actants"), which can be either obligatory or facultative (i.e. implicit), on the one hand, and free modifications (cf. Tesnière's "circumstants"), which cannot be obligatory and are always facultative, on the other hand. Morphologically, there are many ways in which arguments can be realized, i.e. as nouns, prepositional phrases, adjectives or adverbs, cf. Helbig and Schenkel (1973, p.26). Obligatory arguments are necessary for the sentence to be grammatical, while facultative arguments can be omitted under certain circumstances, cf. also Tarvainen (2011, p.9). However, facultative arguments cannot be freely added to any verbal context as free modifications. Both obligatory (cf. *an der Spree* in ex. (2-b)) and facultative arguments (cf. *dem Kind* in ex. (2-a)) are part of the lexeme's valency, and determinable in number and kind, while free modifications (cf. *am Vormittag* in ex. (2-d)) are unrestricted in number and can be deleted and added arbitrarily (Helbig and Schenkel, 1973, pp.33–34). The omission of the obligatory argument *an der Spree* produces an ungrammatical sentence, cf. ex. (2-c).

(2) a. Er wäscht **dem Kind** die Hände.
he washes the child.DAT the hand.ACC.PL
'He washes the child's hands.' (Helbig and Schenkel, 1973, p.47)

b. Berlin liegt **an der Spree**.
Berlin lies by the Spree
'Berlin lies by the Spree.' (Ibid.)

c. *Berlin liegt.
Berlin lies
'*Berlin lies.' (Ibid.)

d. Er besuchte uns **am Vormittag**.
he visited us in morning
'He visited us in the morning.' (Ibid.)

2.1.1.1 Obligatoriness

The notion of obligatoriness is used to define arguments syntactically. However, the obligatoriness of an argument does not imply its morpho-syntactic realization in a sentence under any circumstances. Obligatory arguments can be omitted under certain circumstances, i.e. ellipsis, polysemy/homonymy, alternations and pragmatic omissions. Even though the object *Eier* 'eggs' in ex. (3) is considered to be an obligatory argument it can be ommitted, as the sentence with the object *Eier* 'eggs' is synonymous to the object-less version (ellipsis) (Tarvainen, 2011, p.33).

(3) Die Henne legt (**Eier**).
the hen lays (egg.ACC.PL)
'The hen lays eggs.' (Tarvainen, 2011, p.33)

When a form is homonymous (i.e. based on two unrelated lexemes which are written/spelled the same way) or polysemous (i.e. they have different but related meanings), different senses typically have different valencies. The polysemous verb *leitet* has a facultative accusative argument in ex. (4-a), where *leitet* means 'conduct (electricity)', and an obligatory one in ex. (4-b) where it means 'lead (a meeting)', cf. Tarvainen (2011, p.8).

(4) a. Kupfer leitet (**den Strom**).
copper conducts (the electricity.ACC)
'Copper conducts the electricity.' (Tarvainen, 2011, p.8)

b. Der Dekan leitet **die Versammlung**.
the dean chairs the convention.ACC
'The dean chairs the convention.' (Ibid.)

More systematic changes in the valency structure of a verb affecting the syntactic realization of an argument are diathesis alternations, cf. Levin (1993, p.2). They are alternations in the morpho-syntactic expression of a governor's argument, typically either reducing or enhancing a valency, cf. Helbig and Schenkel (1973). Lopatková et al. (2006, p.1730) specify further that alternations can have at least one of the following effects: a change in the verbform (i.e. derivation) or a qualitative or quantitative change in the valency frame. Qualitative changes involve the obligatoriness, morphological relaizations and lexical meaning of a particular argument.

In ex. (5-a), the valency of the verb *essen* 'eat' is reduced to express the progress rather than the execution of an action (cf. also *Unspecified Object Alternation* (Levin, 1993, p.33)). Valency can also be incremented, as in ex. (5-b) where otherwise intransitive verbs such as *regnen* 'rain' appear with a restricted number of objects.

(5) a. Er aß (**Brot**).
he ate (bread.ACC)
'He ate bread' (Tarvainen, 2011, p.31)

b. Es regnet (**dicke Tropfen**).
it rains (thick drops.ACC.PL)
'It rains thick drops.' (Ibid., p.34)

Panevová (1994, p.238) describes alternations that shift the direct object into subject position as with *the door*. The verb *open*, on the other hand, takes part in an a *causative/inchoative alternation* (Levin, 1993, pp.27–30), where the direct object *door* of (6-a) moves into subject position in ex. (6-b). Syntactically, the verb appears both with an obligatory subject and object, and only with a subject.

(6) a. Mary opens **the door** with a key. (Panevová, 1994, p.238)
b. **The door** opens (with a key). (Ibid.)

Both facultative and obligatory arguments can also be omitted for pragmatic reasons, as they can be text-obligatory instead of sentence-obligatory. In ex. (7-b), the obligatory argument referring to 'the dog', *dem Hund*, is missing. However, it appears as a direct object, *den Hund*, in the previous sentence, ex. (7-a). Tarvainen (2011, p.33) points out that the rules for omission are language-specific. While the direct object *kirjan* 'book' of the verb *antoi* 'gave' can be omitted in the answer in the Finnish example (7-c), it cannot be omitted in the German counterpart in ex. (7-d). The object is required even with the context available and can only be replaced with a pronoun (cf. ex. (7-e)). (Tarvainen, 2011, p.33)

(7) a. Fritz will **den Hund** füttern.
Fritz wants the dog.ACC feed
'Fritz wants to feed the dog.' (Tarvainen, 2011, p.32)
b. Er bringt das Fleisch.
he brings the meat.ACC
'He brings the meat' (Ibid.)
c. Hän antoi minulle **kirjan**. – Antoiko (hän) sinullekin?
s/he gave I.ALL book – gave.Q (s/he) you.ALL.FOC
'S/he gave me the book. – Did s/he give it to you too? (Ibid., p.33)
d. Er gab mir **ein Buch**. – *Gab er auch Dir?
he gave me.DAT a book.ACC – gave he also you.DAT
'He gave me a book. – *Did he give to you too?' (Ibid.)
e. Gab er auch Dir eins?
gave he also you.DAT one
'Did he give one to you too?' (Ibid.)

2.1.1.2 Syntactic tests

The formal basis of a syntactic valency definition are syntactic tests. Obligatoriness, as seen before, is an insufficient criterion to distinguish arguments from free modifications. It is tested by an elimination test, where a clause is removed, testing if the remaining part is still grammatical, cf. Helbig and Schenkel (1973, p.33) and Tarvainen (2011, p.25). However, it only distinguishes obligatory arguments from both free modifications and facultative arguments.

While Čech et al. (2010) criticize the absence of reliable formal criteria for a distinction between arguments and free modifications, Panevová (1994, p.239) and Helbig (1992, p.83) have more confidence in the existence of testable criteria for such a distinction. Helbig (1992, p.83) uses two criteria that characterize arguments: their inability to freely attach to any governor, and the impossibility to use two arguments of the same type in a sentence. Panevová (1994, p.226) formulates two questions to distinguish an argument (if both questions are answered negatively) from a free modification (if both questions are answered positively):

> (a) Do the rules of the language described allow for the occurrence of the given modification with every verb?
> (b) Can the modification occur more than once depending on a single verb token?

The first question is about the interchangeability of free modifications, and the non-interchangeability of arguments. The time adverbial *eine ganze Woche* 'a whole week' can be used both in the context of *half* 'helped' in ex. (8-a) and *unterstützte* 'supported' in ex. (8-b), suggesting its interchangeability and its status as a free modification. The place adverbial *am anderen Ort* 'at another place', on the other hand, can only appear with the verb *wohnt* 'lives' in ex. (8-c), but not with the verb *bewohnt* 'inhabits' in ex. (8-d), suggesting its argument-status.

(8) a. Er half ihm **eine ganze Woche**.
he helped him a whole week
'He helped him for a whole week' (Helbig, 1992, p.82)

b. Er unterstützte ihn **eine ganze Woche**.
he supported him a whole week
'He supported him for a whole week.' (Ibid.)

c. Er wohnt **am anderen Ort**.
he lives at another place
'He lives at another place.' (Ibid.)

d. *Er bewohnt **am anderen Ort**.
he inhabits at another place
'*He inhabits at another place' (Ibid.)

While the test works well for the previous examples, it has its limitations, e.g. in the case of lesser used types of adverbials as examples can be hard to find, cf. Panevová (1994, p.227). While arguments are not interchangable, non-interchangability is not necessarily a sign of argumenthood. According to Panevová (1994, p.227), PURPOSE-adverbials may not combine with just any type of verb for logical reasons and not because they are arguments. In ex. (9-a), 'fall ill' with a PURPOSE-adverbial sounds strange because 'falling ill' is not intentional. When it comes to the second question testing the repetitivity of free modifications, again there are certain types which are not frequent for logical reasons, e.g. free modifications denoting cause in ex. (9-b).

(9) a. ?John fell ill [**in order to be punished for his sins**]. (Panevová, 1994, p.227)
b. [**Due to poverty**] many people died of tuberculosis, [**since its treatment was expensive**]. (Panevová, 1994, p.228)

Apart from the two original questions to test argumenthood, there are assumptions that arguments and free modifications behave differently syntactically, and can therefore be tested by reduction, permutation, etc. The reduction test reformulates free modifications as subclauses or separate main clauses. The clause *hinter dem Hause* 'behind the house' in ex. (10-a) can appear as a separate clause in ex. (10-b), suggesting it is a free modification. In ex. (10-c), on the other hand, it cannot be transformed into two separate predications as in ex. (10-d), suggesting it is an argument, cf. Tarvainen (2011, p.26). However, the test cannot be applied to distinguish between free modifications and facultative arguments as the latter can be traced back to two separate predications as well, cf. Tarvainen (2011, p.27).

(10) a. Die Kinder spielen **hinter dem Hause**.
the children play behind the house (Tarvainen, 2011, p.26)
b. Die Kinder spielen. Das Spielen ist (geschieht) hinter dem Hause. (Ibid.)
the children play. the playing is (happens) behind the house
c. Der Obstgarten liegt **hinter dem Hause**.
the orchard lies behind the house (Ibid.)
d. *Der Obstgarten liegt. Das Liegen ist (geschieht) hinter dem Hause.
the orchard lies. the lying is (happens) behind the house

(Ibid.)

The permutation transformation tests argumenthood by reordering negation adverbials or temporal adverbials assuming that their position is flexible with free modifications and fixed with arguments.

The negation adverb *nicht* 'not' can appear before the clause *in Berlin* in ex. (11-a), but not after the negation verb, cf. ex. (11-b), leading to the assumption that it is an argument of *wohnen* 'live'. However, it can appear before and after the same clause in ex.

(11-c)–(11-d), leading to the assumption that it is a free modification of *treffen* 'meet', cf. also the permutation of time adverbials in Helbig and Schenkel (1973, p.47).

(11) a. Er wohnte nicht **in Berlin**.
he lived not in Berlin
'He did not live in Berlin.' (Tarvainen, 2011, p.30)

b. *Er wohnte **in Berlin** nicht.
he lived in Berlin not
'*He lived not in Berlin.' (Ibid.)

c. Er traf sie nicht **in Berlin**.
he met her not in Berlin
'He did not meet her in Berlin.' (Ibid.)

d. Er traf sie **in Berlin** nicht.
he met her in Berlin not
'He did not meet her in Berlin.' (Ibid.)

Testing argumenthood syntactically has its practical limitations. However, I agree with Tarvainen (2011, pp.30–31) that it seems to be more a theoretical than a practical problem, and intuitively the distinction between arguments and free modifications is clear. I will follow Panevová's (1994) syntactic criteria to distinguish between arguments and free modifications. In addition, I will use other criteria that are usefull in natural language processing tasks.

2.1.2 Selection restrictions and semantic prototypes

Selection restrictions describe the referential semantic properties of the arguments of a governor, cf. level IV in Helbig's (1992) 6 level system (Helbig, 1992, pp.153–155). According to Faulhaber (2011, p.212), governors do not only specify morpho-syntactic restrictions to their arguments, "they also seem to establish restrictions on the possible semantic 'cast' of such participants". Typical selection restrictions refer to humanity, animacy, locality, etc. as opposed to semantic roles, e.g. AGENT, PATIENT, etc. Semantic roles do not refer to the referential and inherent properties of an argument, but rather describe the relation between a governor and its arguments. While the verb *fahren* 'drive' in ex. (12-a) asks for a LOCATION-role referring to a location (*München*), the verb *zerstören* 'destroy' asks for a PATIENT, which can but does not have to be a location (*die Stadt* 'the city') (Helbig, 1992, p.165).

(12) a. Der Zug fuhr **nach München**.
the train drove to Munich (Helbig, 1992, p.165)

b. Die Bomben zerstörten **die Stadt**.
the bombs destroyed the city (Ibid.)

Selection restrictions provide an important link between semantics and syntax, which can-

Syntax	**CONCRETE**	**ANIMATE**	**HUMAN**
subject	possess have get	save obtain believe	own buy hold
object	disturb excite say to	annoy bother persuade	belong to worry embarrass

Table 2.1: English verbs with the selection restrictions *concrete, animate, human* to their subjects/objects (Gruber 1976, p.235)

not be made by "simply focusing on the number and semantic role of a verb's participant" (Faulhaber, 2011, p.223).

However, they differ substantially from syntactic valency in their importance for grammaticality. While a selection restriction violation can influence grammaticality, it can also be a conscious means to change the meaning of an expression. According to Helbig and Schenkel (1973, pp.52–53), the verb *schießen* 'shoot' requires an argument in accusative case with the selection restrictions +animate and -human. Therefore ex. (13-a) is grammatical, and ex. (13-b) is ungrammatical, as the selection restriction is violated, i.e. *Menschen* is +human.

(13) a. Er schießt **Rehe**.
he shoots deer (Helbig and Schenkel, 1973, p.52)

b. *Er schießt **Menschen**.
he shoots people (Ibid.)

Faulhaber (2011, p.213), on the other hand, speaks about "likelihood" rather than grammaticality regarding selection restrictions. She defines the selection restrictions for the object role of the verb *murder* as [+alive at the outset, –alive afterwards, +human]. However, corpus material provides examples with a number of inanimate objects such as thing, music, and hope, which she describes as instances of metonymy or metaphor (vs. grammaticality violations).

Selection restrictions are claimed to be universal (Helbig and Schenkel, 1973, p.65) and syntactically relevant. They can be both general and specific. Helbig and Schenkel (1973, p.52) use, for example, selection restrictions such as material, liquid and vehicle. Common selection restrictions are concrete, animate, human, place, mass, personal, male, female, cf. Gruber (1965, p.233) and Table 2.1. They can be conceptualized as binary features or prototypes, cf. also Bick (2000) and Chapter 4.

2.1.3 Semantic valency

I will use the term "semantic valency" to describe the specification of semantic roles and their constellations with regard to a particular governor, cf. Helbig's (1992) valency level III. Semantic roles are considered to be universal abstractions of language-specific syntactic surface forms on a deeper semantic level, cf. Fillmore (1968, p.1).[1] Theories on semantic roles go back to the Indian grammarian Pāṇini (ca. 500 BC). Pāṇini's theory describes a four-level module of language of which the semantic role level is the deepest and most abstract level. According to Pāṇini, there are six semantic roles ("apadana 'source', sampradana 'receiver', karana 'instrument', adhikarana 'location', karman 'patient' and katr 'agent'" (Keidan, 2011, p.276)) holding a one-role-to-many-morphological-realizations relation (Keidan, 2011, p.279).

The concept of semantic roles was reintroduced by Fillmore, who influenced by Tesnière's valency theory, proposed his (deep) case theory, which would later be known as semantic role theory, in 1966. Fillmore's (1968) original 'case theory' treats six (Agentive, Instrumental, Dative, Factive, Locative, Objective) and later eight/nine deep cases (Fillmore (1971)). Now, there are many semantic role sets that differ in size and their approach to semantics and syntax. One can distinguish between semantics as an intralinguistic concept or "referring to aspects of the extralinguistic situation" (Panevová, 1994, p.225). While a small set of general semantic roles is desirable, corpus work often leads to the need for finer-grained distinctions and larger role-sets, cf. Lopatková and Panevová (2005, p.84) who later introduced the roles "OBST(acle) and MED(iator)" and Fillmore (1968), who anticipated that "additional cases will surely be needed". Helbig and Schenkel (1973, p.63) claim further that not all relations are realized in all languages (and not in the same way in all languages) or they may be obligatory in some languages but are free modifications in others. Again, the general criteria for semantic roles are:

> (1) each argument can modify only a more or less closed class of verbs (that can be listed),
> (2) each argument can modify a particular verb only once (except for the case of coordination)
> (Benešová et al., 2008) (reformulating Panevová (1974, p.11)[2])

Semantic roles are considered to be abstract argument slots, which should not just rename syntactic labels, on the one hand, cf. Helbig (1992, p.19), and for referential semantics, on the other hand, cf. Panevová (1994, pp.233–234).

[1] "A common assumption is that the universal base specifies the needed syntactic relations, but the assignment of sequential order to the constituents of base structures is language specific."

[2] "(1) Can the given type of participant depend on every verb? [...] (2) Can the given type of participant depend more than once on a single verb token [...]?"

2.1.3.1 Semantic roles vs. syntactic functions

Semantic roles can be realized in various morpho-syntactic forms, not only across languages but also within one language. In the synonymous sentences in ex. (14-a) and (14-b), the LOCATION can be realized both as a prepositional phrase *in das Klassenzimmer* 'into the classroom' with the verb *treten* 'enter' (cf. ex. (14-a)) and as a direct object *das Klassenzimmer* 'the classroom' with the verb *betreten* 'enter' (cf. ex. (14-b)) (Helbig, 1992, p.23).

(14) a. Der Lehrer trat **in das Klassenzimmer**.
the teacher went into the classroom

(Helbig and Schenkel, 1973, p.52)

b. Der Lehrer betrat **das Klassenzimmer**.
the teacher entered the classroom (Ibid.)

However, it can be difficult to distinguish semantic roles from syntactic functions if the formal basis for establishing a semantic role set and distinguishing between semantic roles are syntactic tests.

Panevová's (1994) set of five argument types, i.e. ACTOR, PATIENT, ADDRESSEE, ORIGIN, and EFFECT, is based on purely syntactic criteria for the first two arguments of a verb, ACTOR and PATIENT, and semantic criteria for the other roles (Panevová, 1994, p.229). The only argument of an intransitive verb "though it corresponds to different semantic (ontological) roles, such as Bearer, Processor, Stimulus etc." (Lopatková and Panevová, 2005, pp.83–84) is considered an ACTOR. The object of a transitive verb is considered a PATIENT. The system is later enhanced by two additional semantic roles, i.e. obstacle and mediator, cf. Lopatková and Panevová (2005, p.84). Panevová's (1994) main reason for adopting a default subject/object role is to stay clear of non-linguistic (i.e. referential semantic) distinctions, which she claims are the basis for AGENTIVE, EXPERIENCER, THEME distinctions for the subject. However, this makes their semantic role set syntactical.

Panevová (1994, p.228) uses a "dialogue test" for semantic argumenthood. Semantic roles are assigned to semantically obligatory participants, which do not need to be realized syntactically, cf. Panevová (1994, p.232). The "dialogue test" assumes that a semantically obligatory item, which is missing on the surface, is easily recoverable in a communicative situation. The speaker of ex. (15-a) needs to be able to give a satisfying answer to the question in ex. (15-b) about the locative, qualifying it as a semantic argument (with a semantic role), i.e. not answer *I don't know* as this would disrupt the dialogue structure and disqualify her as a speaker. However, the speaker does not need to be able to answer the question in ex. (15-c) about the time, which is considered to be a free modification.

(15) a. Charles arrived **by train**. (Panevová, 1994, p.229)

b. Where did he go? (Ibid.)
c. When did he arrive? (Ibid.)

2.1.3.2 Semantic roles vs. referential semantics

While Panevová (1994) explicitly uses syntactic criteria as the basis of part of her semantic role set, other more semantic theories tend to confuse semantic roles with referential semantics. Fillmore's early case theory (e.g. Fillmore (1968)) has been criticized for being based on cognitive content and factual knowledge instead of linguistic meaning, cf. Sgall (1980, p.526)[3], Helbig (1992, p.26) and Panevová (1994, pp.235–236).

According to Fillmore (1968, p.27), *the wind* in ex. (16-b) is an INSTRUMENT, while *John* in ex. (16-a) is an AGENT, i.e. he distinguishes between inanimate and animate subjects of the same verb. His role distinctions are based on referential semantic characteristics of nouns, i.e. the "Agentive [is] the case of the typically animate perceived instigator of the action identified by the verb" and the "*Instrumental* (I), the case of the inanimate force or object causally involved in the action or state identified by the verb." (Fillmore, 1968, p.24). Those, I will treat as a separate level of valency, i.e. selection restrictions, cf. also Panevová (1994, p.237), who notes that semantic roles are based here on the "lexical content of the given verbs and not directly grammatically relevant, while others can be treated as well by means of a reference to the semantic features of the respective NP's".

(16) a. **John** opened the door. (Fillmore, 1968, p.27)
b. **The wind** opened the door. (Ibid.)

Below I will distinguish between semantic roles, i.e. a relation between governor and argument, cf. Helbig (1992, p.29), and lexical selection restrictions to the arguments.

2.1.4 Semantic verb classes

Semantic verb classes are another approach to a formal basis for semantic roles, but are also considered to be a valency level in their own right by Helbig (1992, pp.153–155) (level II). Potential governors can either be classified decompositionally, cf. Gruber (1965) and Helbig (1992, p.29), or based on their potential to appear in specific frames, cf. Levin (1993). In Figure 2.1, verbs are characterized by means of inherent semantic features, some of which are valency-relevant, i.e. they affect the semantic roles constellations, and others are not, cf. Helbig (1992, p.162). However, semantic ontological systems containing these

[3]"the level including cases (or case roles, etc.) does not belong to the language system in the strict sense, but rather to the realm of cognitive content [...] That is, it has to do with a structuring of factual knowledge, perhaps based on some properties of the structure of human memory, rather than with specific structural properties of a language [...]"

features are connected to conceptual systems, which makes it difficult, if not impossible, to ensure the systems' linguistic validity and their completeness, cf. Helbig (1992, p.162)[4].

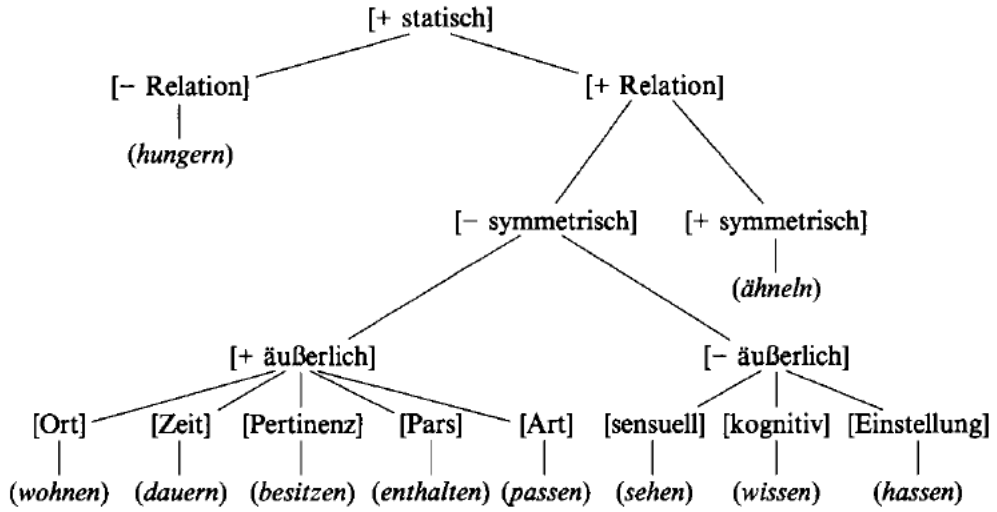

Figure 2.1: Lexical semantic features of German stative verbs, "Zustandsverben" in Helbig (1992)

Both Pinker (1989) and Levin (1993) discourage decompositional approaches of word meaning into atomic features. According to Pinker (1989, p.168) decompositions of verb meanings fail to translate back to the original verb or synonyms of it, i.e. "Chase is not the same as try to catch, for example, and kill is not the same as cause to die". Instead he suggests a syntactically oriented approach classifying verbs according to their ability to appear in the same set of syntactic frames in alternation classes, based on the assumption of semantic and syntactic coherence. This view is shared by Levin (1993, p.1), who assumes that "the behavior of a verb [...] is to a large extent determined by its meaning".

Levin's (1993) verb classes for English include those meaning components that distinguish verbs from each other and/or are syntactically relevant, i.e. participate in different alternations. Levin (1993) distinguishes between 47 alternations that affect the verb's transitivity or involve some diathesis alternation, and 57 verb classes. Verbs like *bake, eat, sing*, and *teach* can be involved in the *Unspecified Object Alternation* changing the verb's transitivity as in ex. (17-a) and ex. (17-b), cf. Levin (1993, p.33f.).

(17) a. Mike ate **the cake**. (Levin, 1993, p.33)
b. Mike ate. (Ibid.)

[4]"Man wird jedoch theoretisch in Rechnung stellen müssen, daß eine restfreie Zerlegung in semantische Merkmale und ein absolut hierarchischer Aufbau der Merkmale an deutliche Grenzen stößt, auch deshalb, weil semantische Kenntnissysteme wesentlich mit konzeptuellen Kenntnissystemen und Strukturen verbunden sind, die von anderem Typ und von anderer Struktur sind als die semantischen Repräsentationen (deren Extension sie sind) [...] "

While I find Levin's (1993) limited-size verb classes and their generalizations extremely useful for grammatical tasks and for constructing semantic role generalizations, I predict that a large-scale categorization of the verb lexicon will most probably result in many one-member sets rather than showing syntactic and semantic coherences between the majority of the verbs. Therefore, I will focus on syntactic and semantic valencies without deliberately constructing semantic verb classes.

2.1.5 Criteria for potential governors

Here I will use the term "governor" to mean a dominating lexeme attracting and requiring certain argument constellations, which are considered to be the valency of the governor. Valency theory is a verb-centered theory, in which the verb is considered to be the highest governor in the sentence. However, not only verbs can be governors and not all verbs can be governors. Additionally, lexemes can be multi-word expressions.

While Tesnière (1959) initially only focused on verbs as governors, most current descriptions assume that nouns, adjectives and adverbs can also have their own valency constellations. However, Helbig and Schenkel (1973, p.23) point out that noun valencies are never obligatory, but always facultative, cf. ex. (18-a). Adjectives, on the other hand, can have obligatory valencies, cf. ex. (18-b).

(18) a. der Besuch (**seines Freundes**)
the visit (his friend.GEN)
'The visit of his friend' (Helbig and Schenkel, 1973, p.23)

b. Der Mann ist **seiner Sorgen** ledig.
the man is his worry.GEN.PL free
'The man is free from his worries' (Ibid., p.22)

When it comes to verbs, only lexically full verbs as opposed to modal auxiliaries are considered to be potential governors, cf. Tesnière (1959). Tarvainen (2011, p.39) claims that modal verbs can be considered grammatical modal morphemes with little lexical content. In certain cases (elliptical constructions), the main verb can be missing and only the modal auxiliary is left with the argument, as *darf* 'may' with *ins Kino* 'into the cinema' in ex. (19-b). In ex. (19-a), on the other hand, the verb *gehen* 'go' is the explicit governor of the DESTINATION-argument *ins Kino* 'into the cinema'. Helbig and Schenkel (1973, p.57) do not consider the modal auxiliary to be the governor of *ins Kino* 'into the cinema'. Instead, the sentence is considered to be an elliptical reduction of the original version of the sentence containing a full lexical verb, and the meaning does not change. I consider the modal auxiliary a governor in the case of rule-based and restricted behavior, i.e. only certain types of arguments co-occur with the modal auxiliary, i.e. DESTINATION, but not SOURCE and can be found with a certain frequency in the corpus.

(19) a. Das Kind **darf** ins Kino **gehen**.
the child may into the cinema go
'The child may go to the cinema.' (Helbig and Schenkel, 1973, p.57)

b. Das Kind **darf** ins Kino.
the child may into the cinema
'*The child may to the cinema.' (Ibid.)

Single-token lemmata are not the only potential governors. Kettnerová and Lopatková (2015, p.191) also consider multi-word verbs that appear in "light verb constructions" potential governors, e.g. "'to make a request', 'to give a presentation', 'to get support', 'to take a shower"'. "Light verb constructions" can be combinations of light verbs and nouns as in ex. (20-a), adjectives as in ex. (20-b) or adverbs. The "light verb" is considered to be semantically incomplete and receives its full lexical-semantic properties in combination with the second part of the multi-word expression. Kettnerová and Lopatková (2015, p.192) assume that both syntactic elements function as a single governor as they have a single AGENT/EXPERIENCER-argument.

(20) a. Peter **won approval** from his boss to change the legal representative of the company. (Kettnerová and Lopatková, 2015, p.192)

b. John **is like** his father. (Ibid.)

2.2 Valency theory in Sámi research

Valency theory has also influenced Sámi research. While early grammars only present paradigms of distinct morphological cases (with supposedly different syntactic functions), later grammars point out relations between morphological case and semantic generalizations, and group verbs according to their potential to appear with specific cases and in specific VALENCY FRAMES. Recent Sámi research explicitly refers to the term "valency", and syntactic tests are suggested to distinguish between valency-bound items and non-valency-bound items.

2.2.1 Case and valency

Semantic roles generalize over alternative morpho-syntactic realizations of certain arguments. In Sámi linguistic descriptions, the association of morpho-syntax and meaning started out the other way around, i.e. as semantic generalizations of morphological cases. While early Sámi grammars categorize morpho-syntactic case semantically and assign meaning to morphology, later descriptions point out that there is no one-to-one correspondence between morphological case and semantic roles, cf. Helander (2001, p.21)[5].

[5]"Morfologalaš kásusis ii sáhte guorrasit njuolga semantihkalaš kásusii dahje temáhtalaš rollii, iige nuppegežiid."

The earliest North Sámi grammars include paradigms with morphological forms without explicitly pointing out syntactic or semantic implications or relations between particular verb classes and complements in a specific case, cf. Leem (1748, pp.1–25), Rask (1832, p.36,49).

Later case-based descriptions such as Stockfleth's (1840) grammar explicitly describe the relation between morphological case and syntactic function and/or semantic relation. However, Stockfleth (1840, p.9) assumes an isomorphy between morphological case and syntactic function/meaning. Friis (1856, p.142) was the first to describe both syntactic and semantic case use.[6] Semantic descriptions of cases typically try to map a particular case to one or several prototypical meanings, but do not refer to the valency of a particular governor. Comitative case, for example, is described by Beronka (1937, pp.63–66) as being used for the person one is accompanied by, speaks with, meets, the means that is used to execute an action, the circumstances of an action, and the causes/causer of an action.[7] However, Beronka (1937, p.65) mentions simultaneous valency changes in the verb *dadjat* 'say' in North Sámi (as opposed to the other Sámi languages). The verb appears with an argument in comitative instead of the synonymous illative case in ex. (21-a), parallel to the Norwegian dialect construction from Finnmark in ex. (21-b). Ruong (1970, p.165), on the other hand, distinguishes between CASE and SEMANTIC ROLE in ex. (21-c), where he describes a verb with two syntactic and semantic arguments, one of them an INSTRUMENT realized by a noun in comitative case (*biillain* 'by car') and the second a DESTINATION realized by a noun in illative case (*Gárasavvunii* 'to Gárasavvon').

(21) a. mon dadjen **iežainan**
I said myself.COM
'I said to myself' (Beronka, 1937, p.65)

b. æ sa de me han Jæns
I said it with him Jens
'I said it to Jens' (Ibid., p.34)

c. Áhčči vujii **biillain Gárasavvonii**
Dad drove car.COM Gárasavvon.ILL
'Dad drove by car to Gárasavvon' (Ruong, 1970, p.165)

[6] "Det Forhold, hvori et Substantiv eller et som Substantiv brugt Ord staar til de övrige Dele af Sætningen, betegnes ved dets Kasus (undertiden i Forbindelse med en Postposition). Substantiver, der staa i samme Forhold, sættes i samme Kasus [...] "

[7] "Die Grundbedeutung des Komitativs ist die des Zusammenseins. Er bezeichnet denjenigen, mit dem jemand zusammen ist, wirkt, spricht, vereinigt ist oder wird, dasjenige was man mit sich hat oder führt (womit jemand zusammen kommt)."

2.2.2 Rection and valency

An explicit approach to valency theory is made by introducing the term "rection" or "government", i.e. the requirement of a particular morpho-syntactic form of an argument with its governor, cf. Bartens (1972). I will use the term "rection" instead of "government" here to distinguish it from uses outside Sámi linguistic research.

The term comes from German and Russian traditional grammar (i.e. "Rektion des Verbs" and "upravlenie") where it denotes the governing of a certain morphological case, i.e. morphological cases of nouns being governed by the verb, cf. Pasierbsky (2003, p.812). According to Helbig and Schenkel (1973, p.44), rection differs from valency, as valency-bound elements do not need to be governed (i.e. covered by rection), but governed elements are always valency-bound, i.e. necessary/obligatory. In example (22-a), Helbig and Schenkel (1973, p.44) claim that the argument of *wohnt* 'lives' is not fixed. However, I do not agree, as the choice of adverbial is not entirely free, cf. ex. (22-b). In this work I will use the term GOVERNOR for all lexemes with valency-bound arguments.

(22) a. Er wohnt **in der Stadt/auf dem Lande/bei seinen Eltern**
he lives in the city/on the country/with his parents

(Helbig and Schenkel, 1973, p.44)

b. *Er wohnt **via Frankfurt**.
*he lives via Frankfurt [L.W., p.k.]

Helbig and Schenkel (1973, p.44) further define rection as both the governing of the morphological case of objects after prepositions, and of the nominal arguments of a verb.[8] Svonni's (2015) North Sámi grammar applies Helbig and Schenkel's (1973) rather broad view on rection including both verbal and prepositional governors, cf. Svonni (2015, pp.53–54).[9] Elsewhere in Sámi research, the term rection generally only applies to the verbal governing of a specific morphological case of their obligatory arguments, as in ex. (23), where *liikui* 'liked' governs the illative case of *niidii* 'the girl' (Svonni, 2015, p.53).

(23) Bárdni liikui **niidii**.
Boy liked girl.ILL
'The boy liked the girl.' (Svonni, 2015, p.53)

According to Bartens (1972, pp.14–15), Magga (1980, p.74), Sammallahti (2007, p.119),

[8]"Regierte Glieder sind immer valenzgebunden; aber - wie die notwendigen Adverbialbestimmungen zeigen – valenzgebundene Glieder sind nicht immer regiert. Valenz und Rektion – so eng sie zusammengehören – dürfen also nicht identifiziert werden, abgesehen von der Tatsache, daß sich die Rektion nur auf Objekte, nicht auf andere Glieder bezieht."

[9]"Dakkár gaskavuohta mii lea vearbba ja nomengihpu gaskkas, dego ovdamearkka dihte cealkagis (5:34) lávejit gohčodit *rekšuvdnan*. Rekšuvdna lea dat gihppu mii oažžu visses kásusa cealkagis go lea dakkár vearba mii gáibida visses kásusa[1] [1]Rekšuvdna lea maiddai nomengihppu mii lea genitiivva hámis pre- ja postposišuvdnagihpuin."

and Nickel and Sammallahti (2011, p.233), the term is used to describe the verbal governing of the morphological case of adverbials, i.e. objects in accusative case and subjects in nominative case are not included. Mikalsen (1993, p.25) explicitly restricts the use of the term to noun phrases, and thereby excludes adpositional arguments, which can be synonymous with morphological case. Nickel and Sammallahti (2011, pp.234–235) include verbal (e.g. *hilbošit* 'tease' in ex. (24-a)), adjectival (e.g. *áŋgir* 'keen' in ex. (24-b)) and nominal governors (e.g. *ráhkisvuohta* 'love' in ex. (24-c)).

(24) a. Ale *hilboš* **ádjáin**!
don't tease Grandpa.COM
'Don't tease Grandpa!' (Nickel and Sammallahti, 2011, p.233)

b. Olmmái lea *áŋgir* **dan bargui**.
man is keen the work.ILL
'The man is keen on the work.' (Ibid., p.235)

c. Son dovddai stuora *ráhkisvuođa* **Ipmilii**.
he felt strong love God.ILL
'He felt strong love towards God.' (Ibid.)

Bartens (1972, pp.14–15) additionally restricts rection to obligatory and semantically unpredictable adverbials.[10] Pope and Sárá (2004, p.251) restrict RECTION semantically to those arguments that do not express time, place, and reason. Semantic unpredictability is implicit in earlier grammars that omit semantic descriptions for rection verbs such as *liikot* 'like' and *luohttit* 'trust' while giving semantic descriptions for other case uses, cf. Ruong (1970, pp.40–41). Mikalsen (1993, p.93), on the other hand, includes examples with predictable semantics as in ex. (25), where the comitative is used to express company.

(25) Moai **ádjáin** oaidnaletne.
we.DU.NOM grandfather.COM see.PRS.2DU
'My grandfather and I meet/see each other.' (Mikalsen, 1993, p.93)

Instead of claiming semantic unpredictability, Nickel and Sammallahti (2011, p.234) and Svonni (2015, p.53) classify illative-rection verbs as a group of predominantly abstract verbs denoting emotions. Additionally, Svonni (2015, p.175) uses semantic roles for defining rection verbs in general. He argues that rection verbs have an EXPERIENCER in the subject position instead of the typical AGENT. An EXPERIENCER-subject is also typical for the group of emotion verbs, which is why Nickel and Sammallahti's (2011) and Svonni's (2015) approaches are based on the same notions.

[10] "Sentraaleihin lauseenjäseniin kuuluvat subjektin ja predikaatin lisäksi predikatiivi ja verbin obligatoriset komplementit: objekti ja verbiä täydentävät, verbin vaatimat adverbiaalit. Viime mainitussa tapauksessa on siis yleensä kysymys verbin rektiosta. Tässä esityksessä on kuitenkin rektiomääreeksi nimitetty vain silloin tällaista verbin määrämuotoon vaatimaa adverbiaalia, kun kaasukselle ei voi määritellä minkäänlaista merkitystä siitä syystä, että määre on ao. verbin määreenä ainoa mahdollinen; [...] "

2.2.3 Transitivity and valency

Not only rection, but also transitivity describes morpho-syntactic restrictions to an argument of a verb. Transitivity describes verbs that may have, but do not have to have, an accusative object, cf. Nickel (1994, p.409). Tesnière (1959) described the concepts of transitivity and valency as related, but valency is considered much broader than transitivity. The term is used only for verbs that govern objects in accusative case by Friis (1856, pp.27–28,143), Nielsen (1926-1929, p.318), Bergsland (1961, p.102) and Magga (1980, pp.74–75). Nickel (1994, pp.409–411), Sammallahti (2007, p.143) and Svonni (2015, p.174), on the other hand, include rection-adverbials in their definition of an object. Magga (2002, p.65) criticizes the latter use of the term based on syntactic criteria. He points out that particularly comitative arguments of rection-verbs can be used alongside rather than instead of accusative objects, suggesting that they do not occupy the same roles. He further mentions that, unlike rection adverbials, accusative objects are further involved in the passive transformation. Here I will use the term *object* only for objects in accusative case. In certain constructions, both obligatory objects and rection adverbials can remain unexpressed. In ex. (26-a), the locative argument of *ballat* 'fear' remains unexpressed, as does the object of *dohppii* 'grabbed' in ex. (26-b), which can be identified by means of its antecedent (*bártni* 'boy'), cf. Nickel (1994, p.410). The previous definition also defines many verbs with semantically limited accusative objects, which are predominantly used intransitively, as transitive verbs. Nielsen (1926-1929, p.319) and Sammallahti and Nickel (2006, p.727,149) define both *vázzit* 'walk' and *čohkkát* 'sit' as transitive verbs as they can have an accusative object as in ex. (26-c) and (26-d). Nielsen (1926-1929, p.318) observes that Sámi verbs that can have an object often correspond to Norwegian intransitive verbs.

(26) a. Mánná ballá.
child fear.PRS.3SG
'The child is afraid.' (Nickel, 1994, p.408)

b. Go stállu bođii sisa ja fuobmái, bártni, de dohppii gitta
when troll came inside and realized, boy.ACC, then took hold
(**su**)
(he.ACC)
'When the troll came inside and noticed the boy, s/he grabbed (him)' (Ibid., p.410)

c. Bohccuid, sávzzaid **vázzit**.
reindeer.ACC.PL, sheep.ACC.PL walk
'Herd reindeer, sheep.' (Vuolab, 1996, p.49)

d. **Čohkkát** riebaniid.
sit fox.ACC.PL
'Hunt foxes.' (Nielsen, 1932-1960*a*, p.413)

Nickel (1994, p.411) and Helander (2001, p.38) also mention availability for passive

diathesis and the potential to form a so-called "actio essive" (i.e. a type of gerund) construction as criteria for transitivity. According to Helander (2001, p.69), certain passive constructions as in ex. (27-a) and essive constructions as in ex. (27-b) are strange/ungrammatical, even though in both cases the verbs *muitit* 'remember' and *gullat* 'hear' typically appear with an object.

(27) a. ?Muhtun uhca mátkefearánaš **muit[o]juvvu**
a small travel.situation.NOM remembers.PASS.PRT.3SG
'A small travel situation gets remembered' (Helander, 2001, p.69)

b. *Son **lei gullamin** cizážiid.
s/he was hearing.ACTIO.ESS small.birds.ACC.PL
'S/he was hearing small birds.' (Ibid.)

Both Helander (2001, p.37) and Sammallahti (2007, p.143) include semantic criteria when defining transitivity. Helander (2001, p.37) distinguishes between objects that are affected THEMES and those that are regular THEMES, the second of which are typically governed by emotion and perception verbs. He uses a pro-verb (*dahkat* 'do') test to distinguish between affected THEMES as in ex. (28-b) governed by verbs like *huškut* 'hit' and regular THEMES as in ex. (28-a) governed by verbs like *gullat* 'hear'.

Sammallahti (2007, p.143) distinguishes betweeen a morphological and a semantic definition of transitivity. Semantically, any two-place verb distributing two semantic roles, independent of their morpho-syntactic realization, can be considered a transitive verb. Both adverbial complement constructions with adverbial case and adpositions in both ex. (28-c) and ex. (28-d) are included in his semantic definition of transitivity.

(28) a. Maid son dagai? – *Son **gulai** mu.
what did do.PRT.3SG – *s/he heard I.ACC
'What did s/he do? – S/he heard me.' (Helander, 2001, p.37)

b. Maid son dagai? – Son **huškkui** mu.
what did do.PRT.3SG – s/he hit I.ACC
'What did s/he do? – S/he hit me.' (Ibid.)

c. Máret guoskkai **girjái**.
Máret touched book.ILL
'Máret touched the book.' (Sammallahti, 2007, p.143)

d. Soai šiehtaiga **gávppi alde**.
they.DU.NOM agreed deal.GEN on
'They agreed on the deal.' (Ibid.)

In *verbs.lexc*[11], some verbs that can have objects, such as *vázzit* 'walk' and *čohkkát* 'sit', are annotated as intransitive verbs. Here, I will apply Nickel's (1994) definition of the term, with some caveats: all verbs that can have an object will be considered transitive

[11] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/verbs.lexc` (Accessed 2017-02-06)

verbs; however, unlike Nickel (1994) and like Magga (2002), I will only consider objects in accusative case *objects*. When describing different morpho-syntactic realizations of arguments, on the other hand, I will refer to their concrete valency.

2.2.4 Syntactic valency

Valency as a relation between the verb and certain types of nominal arguments producing certain types of meaning was addressed early on by Nielsen (1926-1929) and Lagercrantz (1929, p.89).[12] Nielsen (1926-1929, p.328) refers to semantically and syntactically coherent verb classes e.g. verbs denoting dying which appear with a CAUSE in illative case such as *jápmit* 'die' and *hávkat* 'suffocate'.

Mikalsen (1993, p.15) focuses on qualitative valency and groups Sámi verbs into avalent, cf. ex. (29-a), 1-, 2- and 3-place predicates, cf. ex. (29-b).

(29) a. Arvá.
rain.PRS.3SG
'It rains.' (Mikalsen, 1993, p.15)

b. Áhkku bijai **goikebierggu beavdái**.
grandmother put dried.meat.ACC table.ILL
'Grandmother put dried meat on the table.' (Ibid.)

Sammallahti (2007, p.146) defines valency as the quality of a word that decides which dependents it names or receives, cf. also Sammallahti (2005, p.39). When it comes to delimiting what belongs in the valency of a verb, Bartens (1972) distinguishes between central and peripheral parts of a sentence. While obligatory arguments are part of the verb's valency, peripheral ones are not. According to her, the central parts of the sentence are subject, predicative, and the verb's obligatory arguments, i.e. the object and obligatory (rection-)adverbials. Helander (2001, p.30) and Mikalsen (1993, p.15) include only obligatory arguments in their definition of valency.

Obligatoriness is a key concept within valency theory and considered to be a distinctive feature of valency-bound complements of a verb. However, obligatory arguments can be ommitted under certain circumstances, i.e. when they are inherent in the meaning of the governor (Bartens, 1972), or in diathesis alternations of, for example, optionally reciprocal verbs such as *háladit* 'talk a little; talk to each other' in ex. (30-a), where the object is missing. Furthermore, they can be omitted when accompanied by meaning changes as in ex. (30-b) where the past participle of *juhkat* is not used in its meaning 'drink' but 'be drunk'. Helander (2001) also mentions ellipsis, modalizing and contrast (cf. ex. (30-c)).

[12] "Es besteht ein Abhängigkeitsverhältnis mit Bezug auf den Wortsinn zwischen Prädikatsverb und Objekt von der Art, daß das Verb seine jeweilige aktuelle Bedeutung erst im Zusammenhang mit dem Objektswort erhält, wodurch es bestimmt wird."

(30) a. Ánde háladii **etniin**. – Moai háladeimme.
Ánde talked mother.COM – we.1DU.NOM were.talking
'Ánde talked to his mother. – We two were talking.' (Mikalsen, 1993, p.16)

b. Erke lea juhkan.
Erke has drink.PRFPRC
'Erke is drunk.' (Ibid.)

c. Ii son atte, muhto vuovdá.
not s/he give, but sell
'S/he doesn't give, but sell.' (Helander, 2001, p.20)

With regard to syntactic valency, Helander (2001), Nielsen (1926-1929, p.328), Bartens (1978), and Ylikoski (2006) focus further on different morpho-syntactic realizations of the same argument. Already Nielsen (1926-1929, p.328) mentions adpositional counterparts to morphological case arguments, e.g. as in arguments of the verb *suhttat* 'get angry', which can be used with an illative argument or with a postpositional phrase with *ala* 'on', cf. also Bartens (1978) and Ylikoski (2006). Kittilä et al. (2011, p.3) note that case and adposition are similar in their function, which is coding semantic roles. However, adpositions mostly code peripheral rather than core roles like AGENT or PATIENT, cf. Kittilä et al. (2011, p.9). Mikalsen (1993, p.38) mentions nominal arguments, non-finite arguments and subclause arguments for the group of rection verbs, where she classifies approximately 150 verbs according to their valency. Also Nickel and Sammallahti (2011, pp.526–533) show 26 different frames of syntactic valencies altogether including mostly nominal arguments in various morphological cases.

2.2.5 Governors

Verbal, nominal, adjectival and adverbial governors are all discussed in Sámi research. In addition to regular verbal governors, Nielsen (1926-1929, p.329) mentions copula–adjective verb constructions with illative arguments, e.g. *munnji læ al'ke* 'to me it is easy'. Nickel and Sammallahti (2011, p.235) mention both adjectival and nominal governors as in ex. (24-b)–(24-c).

In his school grammar, Ruong (1970, p.163) gives examples for different infinitive constructions, i.e. synthetic construction containing a modal auxiliary (*galgat* 'shall') and the infinitival main verbs (*čuoigat, vázzit, sukkat*) in ex. (31-a) and the main verb *gohčun* 'call (Prs. 1Sg.)' with its infinitival argument *bårrat* ex. (31-b). However, Ruong (1970, p.166)[13] does not point out the syntactic difference between these constructions.

(31) a. Son **galgá** čuoigat, vázzit, suhkat
s/he shall.PRS.3SG ski.INF, walk.INF, row.INF
'S/he shall ski, walk, row' (Ruong, 1970, p.164)

[13]"Märk att i de två sista exemplen uttrycker infinitiven **ändamålet** med (ändamålsorsaken till) resp **orsaken** (grundorsaken) till handlingen eller skeendet som uttrycks i predikatsverbet."

b. **Gohčun** Lásse borrat
call.PRS.1SG Lásse eat.INF
'I call Lásse to eat' (Ruong, 1970, p.164)

Sammallahti (2005, p.287) draws a semantic distinction between auxiliaries that express certain modalities and attitude, and content verbs that establish an event with various participants. In his in-depth study of infinitival constructions and modal verbs, Magga (1982) presents a number of formal criteria (morphological, diacronic, semantic and syntactic) to distinguish between governing verbs and auxiliaries and comes to the conclusion that the distinction between an auxiliary and a governing verb is a continuum, where some verbs are more prototypical auxiliaries than others. Ylikoski (2009, p.34) also points out the difficulty of making "a sharp distinction between auxiliaries and lexical verbs". The criteria for auxiliaries are reduced paradigms for periphrastic, potential (cf. ex. (32-a)) or imperative forms as in ex. (32-b) and availability for passive alternation. The prototypical auxiliary passive construction with auxiliaries as in *leat* 'be' changes the AGENT-subject into a modal PATIENT. However, in ex. (32-c), the PATIENT *mánaid* 'children' does not stay a modal PATIENT of *áigut* 'intend' in the passive construction. Instead, the children become the EXPERIENCER, which is the same role *eadni* 'mother' has in the active sentence. The active and passive versions of the sentence therefore do not imply a syntactic alternation with the same meaning. *Áigut* 'want' is therefore not a prototypical auxiliary with respect to diathesis alternations.

(32) a. ?Dáiddeš bat hal son gal boahtit?
could.POT.PRS.3SG it now s/he really come.INF
'Could it now really be possible that s/he comes?' (Magga, 1982, p.68)

b. *Sáhte / *Galgga / Geahčča vuolgit!
can.IMP / shall.IMP / try.IMP come.INF
'Can to/Shall to/Try to come!' (Ibid.)

c. Eadni **áiggui** gárvvuhit mánaid. – Mánat **áigo** gárvvuhuvvot.
mother intended dress children – children intended dress.PASS.INF
'Mother intended to dress the children. – The children intended to be dressed.' (Ibid., p.87)

2.2.6 Selection restrictions

Semantic selection restrictions made by governors to their arguments, i.e. syntactically relevant inherent semantic features, such as *abstract, concrete, countable*, were mentioned already by Nielsen (1926-1929, pp.303–304). He noted that morpho-syntactic changes in subject-finite verb agreement depending on the semantic features of the coordinated subject. While the finite verb tends to be plural in the case of two concrete nouns that are coordinated, cf. ex (33-a), it tends to be singular in the case of two abstract nouns, cf. ex (33-b).

(33) a. gákti ja boagán **ledje** **[leigga]** juo boahtán
costume and belt have.PRT.3PL have.PRT.3DU already come
'the costume and belt have already come' (Nielsen, 1926-1929, p.303)

b. roahpádus ja dorvvuhisvuohta **fertii** badjelii boahtit
misery and hopelessness have.PRT.3SG upon come
'misery and hopelessness had to come upon (one)' (Ibid., p.304)

Bartens (1978, pp.30–31) mentions semantic preferences to the gentitive complements of the postpositions *sisa* 'inside' typically expressing a DESTINATION and *siste* 'inside' expressing a LOCATION.[14]

Helander (2001) mentions that selection restriction violations differ from syntactic valency errors in that they do not imply ungrammaticality. Depending on the text domain (e.g. fiction) or metaphorical use, typical selection restrictions may be violated. In ex. (34-a), the selection restriction to the subject of *buohcat* 'be sick' is violated as the subject should be animate. According to Helander (2001, p.36), the ungrammaticality of ex. (34-b) is due to a valency error, i.e. an intransitive verb cannot have an object. However, *čohkkát* 'sit' can appear with an animate object (*riebaniid* 'foxes'), cf. ex. (34-c). Ex. (34-b) is really an example of violated selection restrictions leading to ungrammaticality.

(34) a. ?Min **biila** buohcá
our car is.sick
'Our car is sick' (Helander, 2001, p.34)

b. *Máret čohkká **girjji**.
Máret sits book.ACC
'Máret sits the book.' (Ibid., p.36)

c. Piehtár vulggii duoddara čuoigat rievssahiid vázzit, **riebaniid**
Piehtár left tundra ski, ptarmigan.ACC walk, fox.ACC
čohkkát.
sit
'Piehtár left to ski the tundra, hunt ptarmigans and hunt foxes.' (Lagercrantz, 1929, p.91)

[14]"Kysymykseen tulevat varsinkin kaikenlaisten onttojen, avautuvien esineiden, astioiden, säilytysesineiden, kuljetusneuvojen, syvennysten, kuoppien, reikien, rakennusten, kiinteiden, kappaleiden ja aineiden nimet. Kaasusilmausta ja postpositiorakennetta käytetään myös joidenkin maisemanosien nimistä: vuonon, laakson, lahden, metsän, järven, joen, meren nimityksistä."

2.2.7 Semantic valency

Semantic categorizations of specific types of verbal arguments appear as early as in Nielsen's (1926-1929) and Lagercrantz's (1929) grammars. However, Nielsen's (1926-1929) is morphological rather than semantic.

Lagercrantz's (1929) semantic categorizations, on the other hand, generalize over different morphological realizations. He distinguishes between "OBJECTS", which are directly affected as a result of the action expressed by the governing verb, cf. Lagercrantz (1929, p.89), and "OBJECTIVES", typicially a person that is indirectly affected by the action of the governing verb and has an advantage or disadvantage by the action, cf. Lagercrantz (1929, pp.93–94). OBJECTS can be realized as subclauses, postpositional phrases (cf. ex. (35-b)), or by means of noun phrases in accusative (cf. ex. (35-a)) or locative case (cf. ex. (35-c)). Lagercrantz (1929) further shows that both OBJECTS and OBJECTIVES can be realized in various morpho-syntactic forms. For OBJECTIVES, he mentions locative and illative case, cf. ex. (35-d), and postpositional phrases (i.e. with *ala* 'on') as possible realizations. Lagercrantz's (1929) semantic concepts resemble basic semantic roles, but his "role" distinctions are quite different from current descriptions. To name an example, the AGENT *hehpošii* 'horse (Ill.)' of the passive construction in ex. (35-e) is considered to have the same role as the EXPERIENCER *eamidii* 'wife (Ill.)' in ex. (35-d).

(35) a. Leahkas uvssa vai biegga jugista **suova** olggos goađis
open door so wind sucks smoke.ACC out hut.LOC
'Open the door so the wind sucks the smoke out of the hut' (Lagercrantz, 1929, p.91)

b. Hánsa ja Ivvár Piera leigga hállamin **rievssatbivddu** **birra**
Hánsa and Ivvár Piera were talking ptarmigan.hunting.GEN about
'Hánsa and Ivvár Piera were talking about ptarmigan hunting' (Ibid., p.90)

c. Itgo don bora **láibbis**, go gaccat liema?
don't you eat bread.LOC, when eat.with.spoon broth
'Don't you eat of the bread, when you eat broth?' (Ibid.)

d. Giitetgo don **eamidii** gáfe ovddas?
thanked you wife.ILL coffee for
'Did you thank your wife for the coffee?' (Ibid., p.96)

e. Heasttabiebmi čievččahalai **hehpošii**
horse.feeder kick.PASS.PRT.3SG horse.ILL
'The horse feeder got kicked by the horse' (Ibid., p.97)

Semantic role sets with a formal basis for a role definition and a distinction between single roles appear in Helander's (2001) (19 roles), Sammallahti's (2005) (24 roles) and Svonni's (2015) (8 roles) descriptions. Helander (2001, p.23) uses 19 roles for the purpose of showing the potential of the verbs *boahtit* 'come', *vuolgit* 'leave', *mannat* 'go', without commenting on their completeness. Sammallahti (2005, p.41), on the other hand, intends

to suggest a complete role set for North Sámi. The approaches also differ in their formal basis for assigning and distinguishing semantic roles and their use of either syntactic or referential semantic criteria.

While Svonni (2015) associates semantic roles mainly with syntactically obligatory arguments of the verb except for SOURCE/DESTINATION/LOCATION, both Sammallahti (2005) and Helander (2001) include facultative arguments of the verb. Sammallahti (2005, pp.61–62) refers to various diathesis alternations, i.e. causative, passive, reciprocal, and reflexive, that partially preserve semantic roles and change the morpho-syntactic realization of the arguments. This is discussed in detail in Section 3.2.3.3. Svonni (2015, p.166), on the other hand, distinguishes between verbs that ask for a subject that is an AGENT and ergative verbs that resemble passive verbs and have PATIENT-subjects as in ex. (36).

(36) **Lásе** cuovkanii.
window broke
'The window broke.' (Svonni, 2015, p.166)

The definition of semantic roles in Sammallahti (2007) and Nickel and Sammallahti (2011) is predominantly semantic, cf. also Sammallahti (2005, p.39).[15] However, Sammallahti (2007, p.127) specifies that only those dependents of the verb that denote entities (i.e. nouns, sentences, non-finite clauses) receive a semantic role,[16] which again is a syntactic criterion as syntactic argument-types are defined.

Nickel and Sammallahti (2011, p.378) distinguish between "rectives" (*rektiver*) and "oblicutives" (*oblikutiver*). "Rectives" are either "actuators" (*aktuatorer*) or "satellites" (*satellitter*). While "actuators", i.e. arguments, have a semantic role, "satellites", i.e. free modifications, do not have a semantic role. "Oblicutives", on the other hand, are either "statutives" (*statutiver*) or "predicatives" (*predicatives*). While "statutives" have a semantic role and a "semantic function", "predicatives" do not have a semantic role, only a "semantic function". Semantically, "satellites" are distinguished from "actuators" as they do not initiate a situation, but give it content or substance, or describe, comment or place the head, like *johtilit* 'quickly' in ex. (37-a). "Oblicutives", on the other hand, have a "semantic function" towards their syntactic head and their co-dependent, cf. ex. (37-a)–(37-c). The subject predicate *čeahppi* 'good at' in ex. (37-b) does not receive a semantic role, only a semantic function (i.e. modification). The same is true of the essive adjective *hágan* 'efficient' in ex. (37-c) according to Nickel and Sammallahti (2011, pp.367–368) based on the assumption that its semantic function is to modify its co-dependent. However, the sentence requires both an accusative and essive to be grammatical, which is why

[15] "Semantihkalaš rolla lea **dependeantta siskkáldas doaibma dan dilálašvuođas man oaivesátni ásaha**"

[16] "Entitehta doaibma dan dilálašvuođas man oaivesátni ásaha. Semantihkalaš rolla lea dakkár dependeanttain mat almmuhit entitehtaid (omd. substantiivvat, cealkagat, cealkkavástagat) muhto ii dakkár dependeanttain mat govvidit (omd. adjektiivvat, kvantifiserejeaddji sánit jna.)."

in *valency.cg3*, both *Máreha* and *hágan* 'efficient' are considered to have a semantic role.

(37) a. Juhán viegai **johtilit**.
Juhán ran quickly
'Juhán ran quickly.' (Nickel and Sammallahti, 2011, p.378)

b. Máhtte lea **čeahppi**.
Máhtte is skilled
'Máhtte is skilled.' (Ibid.)

c. Máhtte logai Máreha ikte **hágan**.
Máhtte said Máret.ACC yesterday efficient.ESS
'Máhtte said that Máret was efficient yesterday.' (Ibid., p.367)

With regard to assigning and distinguishing between semantic roles, Sammallahti (2005, p.25) and Svonni (2015, pp.164–165) refer to both the non-iterativity of an argument type,[17] cf. Panevová (1974, p.11), and the uniqueness of a semantic role for each argument,[18] which are also known as the THETA CRITERION in Generative Grammar.[19]

However, Nickel and Sammallahti (2011, p.588) mention that roles can appear more than once with respect to the same governor. In the curative (causative) constructions shown in ex. (38-a), both *Máret* and *Máhte* are considered AGENTS with respect to the verb *čuovuhit* 'let follow'. There can also be two roles of the same kind, i.e. two EXPERIENCERS, in Sammallahti's (2005) annotation of adversative passive constructions, due to the fact that any subject of this type is considered an EXPERIENCER, cf. Sammallahti (2005, p.62). In ex. (38-b), on the other hand, Nickel and Sammallahti (2011, p.577–578) consider *Máhtte* both an AGENT and a PATIENT with regard to the reflexive verb *basadit* 'wash oneself', i.e. the uniqueness principle is violated.

(38) a. **Máret** čuovuhii **Máhte/Máhttii** beatnaga
Máret follow.CAUS.PRT.3SG Máhtte.ACC/ILL dog.ACC
'Máret let Máhtte follow the dog' (Nickel and Sammallahti, 2011, p.588)

b. **Máhtte** basadii.
Máhtte washed.REFL.PRT.3SG
'Máhtte washed himself.' (Ibid., p.578)

Sammallahti (2005, p.25) and Helander (2001, p.56) use coordination tests to distinguish semantic roles from each other based on the assumption that roles of the same type can be coordinated. Helander (2001, p.56) notes that *mánát* 'children' and *spábba* 'ball' cannot be coordinated in ex. (39-a), as they are of different types, i.e. AGENT vs. INSTRUMENT. However, *mánat* 'children' and *spábba* 'ball' also differ in semantic prototype, i.e. *mánat* is animate and *spábba* is inanimate. In ex. (39-b) (Helander, 2001, p.56) two arguments

[17]"Lea boares jurdda, ahte guhtege semantihkalaš rolla sáhttá leat cealkagis dušše oktii" (Sammallahti, 2005, p.25)

[18]"Juohke argumeanta oažžu dušše ovtta temáhtalaš rolla" (Svonni, 2015, p.165)

[19]"Each argument bears one and only one θ-role, and each θ-role is assigned to one and only one argument." Chomsky (1981, p.36)

lávuin 'with the lávu' and *fatnasiin* 'with the/by boat', can only be coordinated if they have the same semantic role, i.e. *fatnasiin* is interpreted as an ACCOMPANYER '(together) with the boat' and not an INSTRUMENT of transportation 'by the boat'.

(39) a. ?**mánat ja spábba** cuvkejedje lása.
children and ball broke window
'The children and the ball broke the window.' (Helander, 2001, p.56)

b. ?Máhtte manai **lávuin** ja *fatnasiin* Unjárgii.
Máhtte went lavvu.COM and boat.COM Unjárga.ILL
'Máhtte went with the lavvu and by boat to Unjárga.' (Ibid.)

Both Sammallahti's (2005) and Helander's (2001) role sets are partly based on referential semantics. Sammallahti (2005, p.62) consistently distinguishes between controlled and uncontrolled situations.[20] However, a controlled situation can only be performed by an animate subject, which again is based on referential semantics, cf. Nickel and Sammallahti (2011, p.376). Helander (2001) also distinguishes between different semantic roles for subjects that differ in (referential) semantic category. In ex. (40-a)–(40-c), the subjects of the verb *boahtit* 'come' are considered AGENTS if animate, i.e. *moai* 'we two'. The inanimate object *páhkka* 'parcel' is considered a MOVER, and *biilageaidnu* 'road' is considered an ARRIVER based on a different meaning of the verb *boahtit*, cf. Helander (2001, p.66). While animacy is definitely a referential-semantic category, one can discuss its linguistic status as well. In their somewhat circular definition of linguistic animacy as an "entity's ability to act or instigate events volitionally", Kittilä et al. (2011, p.5) associate acting volitionally with the AGENT-role (Kittilä et al., 2011, p.8), and describe volition as "incompatible with inanimate entities" (Kittilä et al., 2011, p.13). Associating some semantic roles with animate entities and others with inanimate entities is therefore not only a natural tendency, but also a matter of definition. Some of the issues of this section will be discussed further in Chapter 3.

(40) a. **Moai** letne boahtán skuvlastohpui.
we.1DU.NOM have come school.building.ILL
'We two have come to the school.' (Helander, 2001, p.66)

b. **Páhkka** lea boahtán postii.
package has come post.ILL
'The package has arrived at the post office.' (Ibid.)

c. **Biilageaidnu** bođii Beaskáđđasii.
road came Beaskáđđas.ILL
'The road came to Beaskáđđas.' (Ibid.)

[20] "Dan ásahan dilálašvuohta lea maid juoga ládje kontrollerejuvvon, dasgo AUTOMÁHTAN ja VÁSIHEADDJIN leat álo sánit mat almmuhit olbmuid dahje olbmo doaimmaid (*vierut, giella, jurdagat* jna.), muhto *báinnahallat* (seamma go dan vuođđovearba *báidnit*) almmuha dáhpáhusa iige kontrollerejuvvon doaimma."

Style	Valency entry
Helbig and Schenkel (1973)	*beantworten* 'answer': Sn, Sa, (Sd)
Helander (2001)	A_{NOM} + *ŠADDAT* 'become' + B_{ILL} A olmmoš 'human', B báiki 'place' A olmmoš 'human', B doaibma 'activity'
Nickel and Sammallahti (2011)	$N_{nom} + V + N_{ill/lok/kom}$
Levin (1993)	Talk verbs (Class members: speak, talk)

Table 2.2: Valency entries in different human-readable valency resources

2.3 Human-readable and machine-readable valency resources

Valency theory has also influenced natural language processing. In their article about a rule-based approach to Czech valencies, Kettnerová et al. (2012, p.434) point out the key role of valency "in many rule-based NLP tasks such as machine translation, information retrieval, text summarization, question answering, etc.", cf. also Lopatková et al. (2005). But valency resources were valuable long before natural language processing applications had been introduced. Helbig and Schenkel (1973) emphasized their importance in foreign language learning and teaching as well as in translating which is in line with their use in grammar checking, iCall and machine translation.

2.3.1 Human-readable valency resources

Human-readable dictionaries such as Helbig and Schenkel's (1973) lexicon entries (cf. Table 2.2) are based on an active infinitive form specifying one or several sequences of arguments together with their morphological form and information on obligatoriness, i.e. *Sn* denotes an obligatory noun in nominative case, *Sa* an obligatory noun in accusative case, and *(Sd)* a facultative noun in dative case.

Helander (2001, p.50) and Nickel and Sammallahti (2011, p.529) also specify sequences of arguments together with morphological information about them in their grammar/linguistic description, in the form of case information, for example. Helander (2001) additionally specifies a list of possible combinations of selection restrictions, e.g. *olmmoš* 'human' and *báiki* 'place', and *olmmoš* 'human' and *doaibma* 'activity'.

Levin (1993, pp.111–276), on the other hand, specifies syntactic and semantic valencies via 57 verb classes. As opposed to other valency entries in Table 2.2, here verbs do not have single entries, but are listed together with other verb "class members" that share the same syntactic and semantic properties. The frames, i.e. argument constellations, are illustrated via examples and participation in specific alternations. In the case

of *Talk* verbs, participation in the "together reciprocal alternation", cf. ex. (41-e), and non-participation in the "with preposition drop alternation", cf. ex. (41-f), is specified. The comment section includes further explicit morpho-syntactic specifications of the arguments, i.e. sentential arguments, and prepositional phrases with *to* and *with* (Levin, 1993, p.208).

(41) a. Ellen talked.
b. Ellen talked to Helen.
c. Ellen talked to Helen about the problem.
d. Ellen talked with Helen (about the problem).
e. Ellen and Helen talked together.
f. *Ellen talked Helen.

2.3.2 Machine-readable valency resources

Below I will present and compare a number of machine-readable valency resources that are relevant for this research. *VALLEX* (Žabokrtský and Lopatková, 2007), *FrameNet* (Baker et al., 1998; Ruppenhofer et al., 2010), and *VerbNet* (Kipper Schuler, 2005)[21] are manually created machine-readable valency resources in their own right and/or are meant to be used in specific natural language processing tasks. Kettnerová and Lopatková (2013, p.159) mention that *VALLEX* can be used in "machine translation, tagging, word sense disambiguation". *FrameNet* "provide[s] a unique training dataset for semantic role labeling, used in applications such as information extraction, machine translation, event recognition, sentiment analysis, etc.".[22] *DeepDict* (Bick, 2009*a*), on the other hand, is automatically created.

VALLEX is closest to the previously discussed manual resources as it keeps distinctive entries for each lexical unit. *VerbNet* and *FrameNet*, on the other hand, direct both to a frame and a lexical entry (as the main objective is to produce syntactically and semantically coherent classes).

FrameNet is based on Frame Semantics, a theory that "is concerned with networks of meaning in which words participate" (Fillmore et al., 2003), and assigns a semantic frame to each meaning of a verb, e.g. the "*Experiencer_focus*"-frame to the verb *fear*. Semantic roles or "*FEs*" (frame elements), such as CIRCUMSTANCES, CONTENT, DEGREE, EXPERIENCER, EXPLANATION, STATE, TIME, their syntactic realization and their optionality are all specified. Semantic roles can be core, non-expressed core and non-core roles and do not need to be realized morpho-syntactically. A frame like "*Commerce goods-transfer*" can be evoked by verbs like *buy* and *sell*, which contain the same frame elements, but show a different perspective: "the first takes the Buyer's perspective and the second the Seller's

[21] `http://verbs.colorado.edu/$\sim$mpalmer/projects/verbnet.html` (Accessed 2017-02-06)
[22] `https://framenet.icsi.berkeley.edu/fndrupal/about` (Accessed 2017-02-06)

perspective" (Ruppenhofer et al., 2010, p.10). Frames are further related to each other within a frame hierarchy via inheritance relations and others. Phonological, morphological and etymological information is not provided by *FrameNet* according to Fillmore et al. (2003, p.248).

VerbNet and its second version *VerbNet2*, on the other hand, are based on Levin's (1993) classification of English verbs and involve semantic verb classes. As in *FrameNet*, semantic role constellations (i.e. frames), selection restrictions (e.g. *+animate*), and syntactic labels (e.g. *NP V NP*) are all specified for the entire verb class. The verb *fear* belongs to the semantic verb class *admire* with the semantic roles STIMULUS, EXPERIENCER [+animate] and ATTRIBUTE, some of which can be optional in particular frames. However, much fewer semantic roles are specified in in *VerbNet* than in *FrameNet*.

VALLEX is based on Functional Generative Description (FGD) (Lopatková et al., 2005) and maps one or multiple valency frames to a lexeme with all its senses. This also includes reflexive particles, aspect and light verb constructions if available. Each valency frame is made up of a sequence of coarse-grained semantic roles, e.g. 'ACTOR (ACT), PATIENT (PAT), ADDRESSEE (ADDR), ORIGIN (ORIG), and EFFECT (EFF), a list of their morpho-syntactic realizations, and obligatoriness specifications. There are various further attributes for diatheses, light verb combination, reflexivity and links to semantic verb classes based on *FrameNet*, cf. Kettnerová et al. (2012, p.25). Unlike the other resources mentioned, *VALLEX* has a rule-component in its lexicon specifying the potential of lexicon entries to enter specific diatheses alternations and morpho-syntactic changes related to that (Vernerová et al., 2014, p.2454). This rule-component allows generalizations over rule-based processes and economically stores only the unmarked lexicon entry (Lopatková et al., 2006). According to Vernerová et al. (2014, p.2452), the applicability of specific diatheses is often lexically conditioned (even if there are productive grammatical processes), which is why diatheses should be stored in the lexicon. While *VALLEX* is syntactically much denser than *FrameNet*, *FrameNet* includes more semantic information that is useful in generation, information retrieval and question answering tasks, which is why *FrameNet* data has been included in newer versions of *VALLEX*, according to Benešová et al. (2008, p.18). *VALLEX* and *FrameNet* also differ in their approach to animacy. While semantic roles in *VALLEX* generalize regarding animacy, *FrameNet*-roles are specific about animacy (Benešová et al., 2008).

Bick's (2009*a*) *DeepDict* is not a valency resource per se but a "multilingual co-occurrence lexicon automatically extracted from dependency parsed corpora" (Vernerová et al., 2014, p.2453). Bick (2010) uses it to extract selection restrictions for verbal arguments specifying both certain prototype classes and their probabilities, i.e. "a given verb has a 30% probability of a direct object (ACC) of the food class". By means of this information verbs can be classified and the animacy of their pronominal arguments can be determined. While *VALLEX*, *FrameNet* and *VerbNet* specify semantic roles and

	Helbig & Schenkel	**VALLEX 2.5**	**FrameNet**	**VerbNet**	**DeepDict**
Lexicon entries	500	6,460	>12,000	8,537	depends on the corpus
Semantic roles	no	8	>12,000	30	no
Selection restr.	ca. 12	no	no	36	?
Semantic classes	no	22	yes	273	no
Morpho-syntax	yes	yes	yes	yes	yes
Statistics	no	no	no	no	yes
Obligatoriness	yes	yes	no	no	no
Human readable	yes	yes	yes	yes	yes
Machine readable	no	yes	yes	yes	yes
Manually annotated	yes	yes	yes	yes	no

Table 2.3: A comparison of some machine-readable valency resources

syntactic realizations, *DeepDict* does not refer to semantic roles, or specify obligatoriness or other linguistic generalizations, but relies mostly on syntax, part of speech and statistics. *DeepDict* does not distinguish between arguments and free modifications or establish semantically and syntactically coherent classes of governors.

In terms of size, *FrameNet* (12,000 lexical units) is double the size of *VALLEX 2.51* (6,460 lexical units), which is just slightly smaller than *VerbNet* (8,537 lexical units), while Helbig and Schenkel's (1973) valency dictionary in its second edition is significantly smaller with 500 entries, cf. Table 2.3. Since *DeepDict* is not a dictionary per se, but searches for co-occurrences on the fly, it does not have a fixed size. In addition, *FrameNet* names its corpus of ∼123,000 sentences, which is meant to be used as "training data for semantic role annotation, information extraction, machine translation, event recognition, sentiment analysis, etc." but also as a human-readable valency dictionary.[23]

Although valency "belongs to the core information for any rule-based task of NLP (from lemmatization and morphological analysis through syntactic analysis to such complex tasks as e.g. machine translation)" according to Lopatková et al. (2006), few rule-based approaches actually use valency information in their concrete tasks. Most documented applications are machine learning systems, semantic role annotation systems such as Gildea

[23] `https://framenet.icsi.berkeley.edu/fndrupal/about` (Accessed 2017-02-06)

and Jurafsky's (2002) statistical system[24] and very few rule-based systems, such as e.g. Bick's (2000) syntactic parser for Portuguese (*PALAVRAS*). The incorporation of valency tags in specific natural language processing applications is discussed in the Chapters 3–5.

2.4 Methodology and framework

2.4.1 Methodology

The methodological questions in my research regard both linguistics and natural language processing as this study consists of both linguistic research and the devolpment of rule-based machine-readable grammars. As the resources are rule-based, I make use of both introspection and corpus when making decisions about grammaticality and constructing rules. Tesnière (1959, Chapter 18) dedicates a whole chapter on introspection, which is often considered to be subjective as it is based on intuitions. However, Tesnière (1959) argues for its objectivity as it is also based on internal experience and therefore an experimental method.[25] According to Tarvainen (2011, p.25) the subjectivity of one's own intuition can be reduced by the use of corpus material and additional informants.[26] While Tesnière stresses that introspection requires a native speaker of the language, Ylikoski (2009, p.23) points out that "the use of intuition is not limited to the study of one's native language; rather the possibilities and the limits of intuitive knowledge of language always accompany the study of more or less foreign languages as well". I am a native speaker of German and a second language speaker of North Sámi, which is why I make use of both well-represented grammatical structures in corpus material and native language speaker intuitions in this research where grammatical descriptions of the phenomena do not exist. In addition to using my own language intuitions as far as they exist, I mostly seek confirmation of them by means of using the language intuitions of predominantly two native speakers with a linguistic education, representing the eastern (referred to as *N*) and western (referred to as *H*) North Sámi dialects. In addition, I have had discussions about certain constructions with members of *Sámi fágájoavku* at *UiT The Arctic University of Norway* – i.e. Sámi philologists – and with the normative or-

[24] `https://framenet.icsi.berkeley.edu/fndrupal/ASRL` (Accessed 2017-02-06)

[25] "§2 This method can be accused in particular of being **difficult** to employ due to its **subjective** and consequently **dangerous** character. [...] §4 The introspective method will be criticized for its *subjective* character, for it appeals to **intuition**. §5 In this area, the grievance is even more questionable. The introspective method does indeed appeal to intuition. But it also appeals to **internal experience**. In this respect, it is an **experimental** method and as a consequence, it is **objective**. [...] §13 It follows that in principle the introspective method can only be used on the mother tongue of the user. Its use therefore requires that the **linguist also be the speaker**."

[26] "Man muß sich aber darüber im klaren sein, daß man [...] auf seine sprachliche Intuition, d.h. auf sein eigenes Urteil darüber, was in der eigenen Sprache üblich und möglich sei, angewiesen ist. Das subjektive Moment, welches diesem Verfahren innewohnt, ist selbstverständlich durch Herbeiziehung ergänzender Methoden, vor allem durch Informantenbefragung und durch Untersuchung von Textkorpora reduzierbar."

gan *Giellagáldu*. Unless marked otherwise, all native grammaticality judgments of North Sámi in this work are made by these two native speakers. Ungrammatical sentences are marked by an asterisk (<*>). It is commonly thought that grammaticality judgmenents are gradable. A question mark (<?>) marks a lesser degree of ungrammaticality, where a form is neither entirely accepted nor entirely discarded.

This research has three natural language processing products: a grammar for valency annotation, a set of referential-semantic tags to enhance the noun lexicon, and a grammar for error detection and correction, which is part of *GoDivvun*, the North Sámi grammar checker *Giellaoahpa Divvun*, all of which are manually constructed, but tested and evaluated on *SIKOR*. All resources are created manually. When developing the valency resources *VALLEX*, Kettnerová et al. (2012) note that manual annotation, despite being time consuming is "indispensable at this stage of research as it brings necessary insight into the problem", and according to Lopatková et al. (2005) "guarantees a significant rise of quality". Previous human-readable valency descriptions in various Sámi grammars and dictionaries, e.g. Nielsen (1926-1929) and Sammallahti and Nickel (2006), serve as the basis for developing the valency resource (*valency.cg3*), as does a verb frequency list in *SIKOR*. Examples, unless marked otherwise, are real examples taken from *SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection* (2016-12-08)/(2015-03-01). In a few cases, counter-examples are constructed by native speakers to test their grammaticality. *SIKOR* contains administrative, law, religious, non-fiction, fiction, and science texts. *SIKOR* (2015-03-01) consists of 21,108,052 tokens, annotated with morphological and syntactic, but not yet semantic information and can be searched using the corpus searching tool *Korp* developed by *Språkbanken*, cf. Borin et al. (2012) and Ahlberg et al. (2013). Although the corpus is growing and significantly bigger than a few years ago, it is small compared to the corpus of Swedish searchable by *Korp 6.0*, which has 11.52 billion tokens.[27] This means that certain grammatical phenomena are not represented or have very few examples, which again influences rule development and evaluation.

As regards grammatical terms describing North Sámi word forms, I primarily use terms from the Sámi linguistic tradition as those are applied in *Giella-sme*. If considered necessary, I mention their equivalent form in English, i.e. "the progressive form actio essive". Numerated example sentences that follow older North Sámi orthographic conventions, e.g. from Nielsen (1926-1929) or Beronka (1937, p.57), are standardized according to current conventions. Within larger quotes they are left in their original form. In some cases, minor spelling errors are corrected for better understanding; the corrections are marked by square brackets. In cases where examples illustrate grammatical or spelling errors, these are not corrected. Examples are glossed according to the "Leipzig Glossing Rules",[28]

[27] `https://spraakbanken.gu.se/korp/#?stats_reduce=word&cqp=%5B%5D` (Accessed 2017-08-31)
[28] `https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf` (Accessed 2017-03-23)

e.g. with regard to alignment, small caps for grammatical labels, one-to-many correspondences of words, etc. Glossing diverges from the Leipzig Glossing Rules in the following ways: Segmentable morphemes are not separated by a hyphen. As this research focuses on morpho-syntactic processes rather than identifying morpheme boundaries, glossing only provides a morphological analysis of the whole word. Only forms relevant to linguistic descriptions and those necessary for understanding are glossed morphologically. Relevant forms in examples are marked in bold letters and secondary forms are marked in italics. In the case of ambiguous forms, only relevant (and not necessarily all possible) analyses are glossed. Spelling follows the orthographical norm for North Sámi. Although forms like *nammalassii* (instead of *namalassii* 'namely') are commonly used, here they will be marked as errors as they do not follow the orthographical norm. The morpho-syntactic analysis follows the conventions used in *Giella-sme*. Some analyses can be disputed, e.g. *áddjáid* 'time-consuming' can be analyzed as an adverb or an accusative plural form of the adjective *áddjái*. However, these discussions fall outside the scope of this work and I will follow the analysis provided by *Giella-sme*. Sentences with errors are left uncorrected where the error is relevant as regards the analysis and error detection process. In their translation into English, I deliberately leave an incorrect construction if it serves as an explanation for a particular phenomenon in the Sámi original. In these cases the English translation is marked with an asterisk. Abbreviations for grammatical tags are taken from *Giella-sme* and Leipzig Glossing Rules and can be reproduced by the reader by means of the online morpho-syntactic analyzer.[29] Where necessary for the linguistic discussion, further grammatical abbreviations are used. A list of the grammatical tags used in *Giella-sme* is provided in Appendix C. Further abbreviations are listed in the glossary. A number of example sentences that are used for illustration are only glossed, not translated. Code from the machine-readable grammars or output from the analyzers is displayed in Verbatim font. When valency resources are incorporated in natural language processing tools, these serve as automated testing mechanisms that help to remove further inconsistencies and refine the system. Semantic prototypes are developed based on existing semantic sets within the disambiguation grammar (reference) and already existing semantic prototype hierarchies, cf. Bick (2000), and tested for their syntactic relevance within other North Sámi tools.

The error detection grammar is based on real errors from *SIKOR* and rules that make use of introspection and are tested on corpus material. The process of coming to a conclusion based on introspection is illustrated in Figure 2.2, which is an excerpt from a chat-discussion between informant *H* and me in the process of developing the grammar checker *GoDivvun*. When testing the grammaticality of ex. (42), i.e. why it should be *addet* 'give' not *áddet* 'understand', I reduce the context of the error to localize the relevant context cue for the error. *Dat sáhttet válljet áddet.* is correct. *Dat sáhttet válljet*

[29] `http://giellatekno.uit.no/cgi/d-sme.eng.html` (Accessed 2017-03-28)

áddet go Ealáhusdieđáhus is still correct although one expects a following context. *Dat sáhttet válljet áddet go Ealáhusdieđáhus 1 vai 2*, on the other hand, is not correct because of the coordinator *vai* 'or' (in questions).

(42) Dat sáhttet válljet ***áddet** go Ealáhusdieđáhus 1 vai 2.
they can choose understand Q industry.message.NOM 1 or 2
'They can choose to give either industry message 1 or 2.'

L: "Should this be áddet or addet?"
H: 'addet'
L: "Why?"
H: . . . [no anwer]
L: "Dat sáhttet válljet áddet go - is this wrong?"
H: "No."
L: "Dat sáhttet válljet áddet go Ealáhusdieđáhus 1 vai 2. - Is this wrong?"
H: "Yes."
L: "Why?"
H: "because of 'vai"'

Figure 2.2: The use of introspection when analyzing grammatical errors

While this work does not study and evaluate a norm, it describes and evaluates ways of modelling a norm. A grammar checker typically marks forms and constructions that are deviant from a norm as unacceptable. Normative acceptable is typically defined as something decided by an authorized organ, i.e. *Giellagáldu*.[30] In the case of the absence of approved norms, I will instead refer to written grammars, linguistically respected people in society, or the grammar intuitions of an individual in the case of an individual grammar. *Giellagáldu* is the Northern center for all Sámi languages across national borders (i.e. Norway, Sweden and Finland) and is answerable to the national Sámi Parliaments. Its responsibility is to preserve and promote the Sámi languages, decide on new terminology and work on a Sámi language norm. While its publication *Riektačállinrávvagat* (2015) mostly contains orthographic norms and norms about punctuation and formatting, its first publication *Čállinrávagirji* (2003) also contains a number of syntactic norms on the use of cases, passive, congruence, etc. Both competence and understandability are relevant in grammar checking. Users of a grammar checker typically expect their corrected writing to be normative and pass the revision of, for example, a teacher, an editor of a newspaper or a journal, or a committee receiving an application. Additionally, the text should be understandable for an audience.

Valency variation is generally not free and language intuitions can be very clear as to what is acceptable and what is not acceptable. The following chat-discussion with

[30] `http://www.giella.org/` (Accessed 2017-02-06)

informant *H* and informant *N* shows that informant *H* clearly finds *guoskat* 'touch' with an accusative object, i.e. *su* 'her, him' acceptable. Informant *N*, on the other hand, prefers an illative argument for *guoskat* 'touch'.

```
L: H can guoskat be with acc even if acc is not an object?
H: yes, Ašši guoská su is fine.
L: what do you think N?
N: for me guoskat needs to be Ill
```

The natural language processing tools are developed within the Constraint Grammar framework and tested by means of commonly used measures such as precision, recall and accuracy as regards the grammar checking tools and lexicon and corpus coverage (both for type and token) as regards valency tags and semantic prototype tags. Corpus coverage is tested on *SIKOR*. Recall is the number of correctly retrieved items divided by the number of items that should have been retrieved. Precision, on the other hand, calculates the number of correctly retrieved items divided by the number of actually retrieved items.

$$\textbf{Recall} = \frac{\text{number of items correctly retrieved}}{\text{number of items that should have been retrieved}}$$

$$\textbf{Precision} = \frac{\text{number of items correctly retrieved}}{\text{number of items actually retrieved}}$$

$$\textbf{Accuracy} = \frac{\text{number of items correctly retrieved and not retrieved}}{\text{number of items actually retrieved and not retrieved}}$$

2.4.2 Framework

Valency resources are integrated in the existing Divvun & Giellatekno infrastructure *Giella-sme*, cf. Antonsen et al. (2010) and Moshagen et al. (2013), in fst-lexica and compilers and various North Sámi constraint grammars.

2.4.2.1 The Constraint Grammar formalism

The Constraint Grammar formalism (CG) is a rule-based formalism for writing disambiguation and syntactic annotation grammars, originally introduced by Karlsson (1990), cf. also (Karlsson et al., 1995, p.57–63,70–71), and further enhanced with a number of features that, for example, allow for dependency annotation, etc. in its *visl* constraint grammar compiler (version 3) (vislcg3) version,[31] which is used for the compilation of constraint grammar rules (VISL-group, 2008), cf. also Bick and Didriksen (2015). The philosophy behind it relies on a bottom-up analysis of running text. Possible analyses

[31] `http://visl.sdu.dk/constraint_grammar.html` (Accessed 2017-02-06)

are discarded step by step with the help of morpho-syntactic context. The North Sámi constraint grammar analyzers take morphological ambiguous input, which is compiled with finite-state transducers and Xerox two level compiler (twolc) and lexicon formalism for machine-readable lexica designed by Xerox, and a compiler with the same name that turns the lexicon into a fst (lexc) (Beesley and Karttunen, 2003). The transducers may also be compiled from the same source code with the open source compiler *HFST*,[32] which is more in line with *Giellatekno & Divvun's* philosophy, cf. Trosterud (2006).[33] For grammar checking only the open source compiler is used.

Existing constraint grammar tools for North Sámi include, for example, a morphological disambiguator/syntactic parser (Trosterud, 2004), and a dependency analyzer (Antonsen et al., 2010). The analyzers include manually written rule sets, which select the correct analysis in case of homonymy, and add grammatical functions and dependency relations to the analysis. Constraint grammar rules typically specify domain, operator, target and context conditions. In the example below the error tag is mapped onto the third person indicative form (*Ind Prs Sg3*) under the following conditions. Firstly, there is an auxiliary to its left unless it is a form of the lemma *ii* 'not'. Secondly, tokens between the auxiliary and the third person indicative form should not include a noun phrase modifier or an adverb and act as a barrier to the constraint (*NPNHA*). Lastly, there should not be a habitive (*@HAB*) to the left of the auxiliary (*AUX*).

```
Operator Rule Domain      Target                              Context conditions

ADD errortag TARGET       (Ind Prs Sg3)  IF      (*-1 AUX - ("ii") BARRIER NPNHA
                                                 LINK NEGATE *-1 @HAB) ;
```

The contexts can be absolute (i.e. referring to a fixed position in the sentence) or relative (i.e. referring to a position to the left or right with a certain distance to a specific constraint). That way the full linguistic potential of a sentence is exploited and tedious pattern matching is avoided where only constructions within the imagination of the linguist/developer make their way into the system. Context conditions can be modified by means of barriers, i.e. tokens or combinations of tokens that stop the scanning of the sentence. In addition, contexts can also be linked to further contexts. This allows the system to work globally and refer to complex syntactic relations when performing tasks such as making e.g. disambiguation choices.

Linguistically, both phrase structures and dependency structures can be implemented in *Constraint Grammar*. However, it has been used predominantly for dependency grammars, which this work also relies on. There are rule types made specifically for dependency

[32] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/` (Accessed 2017-02-06)

[33] "A more serious question is the choice of Xerox tools vs. open source tools. In our project, we have no wish to modify the source code of the rule compilers themselves, but we notice that all binary files compiled by the xfst, lexc and twolc compilers are copyrighted property of the Xerox Corporation [now: Palo Alto Research Center, Inc]."

analysis, i.e. *SETPARENT* (mapping a specific token to its parent) and *SETCHILD* (mapping a specific token to its child). There are also a number of rule types useful for error correction, i.e. *ADDCOHORT* (adding a token and all potential analyses), *MOVE-COHORT* (moving a token and all possible analyses), and *DELETE* (deleting a token with its analyses). These are further described in Chapter 5. For a full overview of rule types cf. Didriksen (2010).

2.4.2.2 North Sámi constraint grammars

Three separate grammars were developed within this PhD project: an annotation grammar for valency tags and a few grammatical rules for alternations (*valency.cg3*[34]), a disambiguation grammar for the grammar checker (*disambiguator.cg3*[35]) and an error annotation/correction grammar (*grammarchecker.cg3*[36]). All grammars developed are compiled with vislcg3.

The grammars use a set of part of speech tags, morphological tags, shallow syntactic tags, semantic prototype tags and dependency relations. Lexica are divided by part of speech, i.e. adjective, adverb, conjunction, interjection, noun, numeral, particle, adposition, pronoun, proper noun, verb, based on Nickel's (1994) and Nickel and Sammallahti's (2011) North Sámi reference grammars. Abbreviations and acronyms are treated separately.

Semantic prototype categories are accessible for the grammars via lexc, and are stored in the *Giella-sme*-lexica: *nouns.lexc*,[37] *propernouns.lexc*,[38] *adjectives.lexc*,[39] *adverbs.lexc*,[40] *acronyms.lexc*,[41] and *abbreviations.lexc*.[42]

The subsequent program in the *Giella-sme* pipeline, i.e. *lookup2cg*, makes the output of the morphological analyzers constraint grammar-compatible. Newer versions of *Giella-sme*, such as the grammar checker *GoDivvun*, use an extension to the *hfst-pmatch*-runtime (Hardwick et al., 2015), developed by Kevin Unhammer, to analyze and segment a sentence in a constraint grammar-compatible way in one go. Part of speech and morpho-

[34] `https://victorio.uit.no/langtech/trunk/langs/sme/src/syntax/valency.cg3` (Accessed 2017-02-06)

[35] `https://victorio.uit.no/langtech/trunk/langs/sme/tools/grammarcheckers/disambiguator.cg3` (Accessed 2017-02-06)

[36] `https://victorio.uit.no/langtech/trunk/langs/sme/tools/grammarcheckers/grammarchecker.cg3` (Accessed 2017-02-06)

[37] `http://www.divvun.no/doc/lang/sme/nouns-stems.html` (Accessed 2017-02-06)

[38] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/sme-propernouns.lexc` (Accessed 2017-02-06)

[39] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/adjectives.lexc` (Accessed 2017-02-06)

[40] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/adverbs.lexc` (Accessed 2017-02-06)

[41] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/acronyms.lexc` (Accessed 2017-06-16)

[42] `https://victorio.uit.no/langtech/trunk/langs/sme/src/morphology/stems/abbreviations.lexc` (Accessed 2017-06-16)

logical tags can be found in the root lexicon *root.lexc*.[43] Syntactic tags[44] typically indicate a syntactic function (cf. Nickel (1994)), i.e. subject, object, adverbial, apposition, etc. prefixed by the @-symbol and indicating the direction of the dependency relation by an arrow to the left or to the right; however, without explicitly naming the head. *@ADVL>* for example indicates that the given token is the head of an adverbial with its mother (typically a finite or infinite verb) to its right. *@>N*, on the other hand, can be part of a noun phrase, modifying a nominal mother to its right. Verbal predicate tags typically do not have an arrow as they are considered to be at the topmost dependency level. *@-FMAINV* is an infinite main verb typically the daughter of a *@+FAUXV* (a finite auxiliary).

The distinction between auxiliaries and main verbs is made within the disambiguation grammars *disambiguation.cg3*[45] and *disambiguator.cg3* for grammar checking via separate sets for copulas (verbs that appear with predicatives), i.e. realcopulas (verbs that appear with perfect participle arguments), modal-aspectual auxiliaries, and verbs that can appear with objects and without another main verb.

A constraint grammar rule such as the one below then maps the finite auxiliary syntactic function tag (*@+FAUX*) to this group of verbs (*AUX*) if they are finite forms (*VFIN*).

```
MAP (@+FAUXV) TARGET VFIN IF (0 AUX);
```

When mapping dependencies, syntactic tags are slightly adapted to suit explicit dependencies.[46] Dependency tags are expressed in the following way, *#5->2*, the first number indicating the absolute position of the token in the sentence (i.e. 5) pointing to the absolute position of the token representing the mother (i.e. 2). The finite verb is typically pointing to 0, indicating its sentential head status. The infinite main verb typically points to a finite verb, which is either an auxiliary (cf. Figure 2.3 and ex. (43-a)) or a finite main verb (cf. Figure 2.4 and ex. (43-b)). While subjects point to the finite verb, objects and adverbials typically point to the main verb, i.e. *son* 's/he' points to *divttii* 'let (Prt. 3Sg.)' and *skuvlabargguid* 'schoolwork (Acc.)' points to *bargat* 'work' (cf. Figure 2.4). Within the Basque constraint grammar dependency analysis, Aranzabe (2008, p.89), on the other hand, annotates both subjects (*Mikelek*), objects (*bazkaria*), adverbials and auxiliaries (*du*) as daughters of the main verb, cf. ex. (43-c) and Figure 2.5, although both subject and direct/indirect object agree with the auxiliary in Basque. In her dependency annotation, the valency of the main verb is given predominance over the syntactic agreement.

[43] `http://www.divvun.no/doc/lang/sme/root-morphology.html` (Accessed 2017-02-06)

[44] `http://www.divvun.no/doc/lang/sme/docu-sme-syntaxtags.html` (Accessed 2017-02-06)

[45] `https://victorio.uit.no/langtech/trunk/langs/sme/src/syntax/disambiguation.cg3` (Accessed 2017-03-28)

[46] `http://www.divvun.no/doc/lang/common/docu-deptags.html` (Accessed 2017-02-06)

(43) a. Mun lean boahtán.
I have come
'I have come'

b. Son divttii Liná bargat skuvlabargguid.
s/he let Liná do schoolwork.ACC.PL
'S/he let Liná do her schoolwork.' [H, p.k.]

c. Mikelek bazkaria prestatu du.
Mikel.ERG lunch.ABS prepare AUX.ABS3SG.ERG3SG
'Mikel prepared lunch.' (Aranzabe, 2008, p.89)

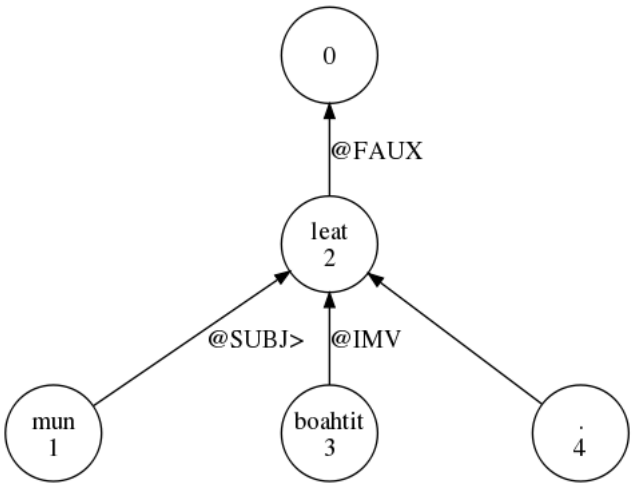

Figure 2.3: The dependency structure of *Mun lean boahtán.* in *Giella-sme*, cf. ex. (43-a), @FAUX = finite auxiliary, @SUBJ> = subject, @IMV = infinite main verb

2.5 Definition of key concepts

Here, I define a set of key concepts which will be used in the discussion to follow.

Key term 1 (Valency) Valency *is the potential of a governor (i.e. verb, noun, adverb) to syntactically/semantically combine with a specific number and type of arguments.*

Key term 2 (Governor) *A* governor *is a lexeme determining another lexeme (its argument) syntactically and/or semantically. Syntactically, a governor demands a certain morphological form or position of its arguments. Semantically, a governor demands a specific semantic role for its arguments. Unless otherwise specified, the term* governor *is used for semantic government.*

Key term 3 (Lexeme) *A* lexeme *comprises all forms within one unique paradigm. It is represented by a randomly chosen form, i.e. the infinitive of a verb or the nominative form of a noun to which a semantic prototype category or valency information is added.*

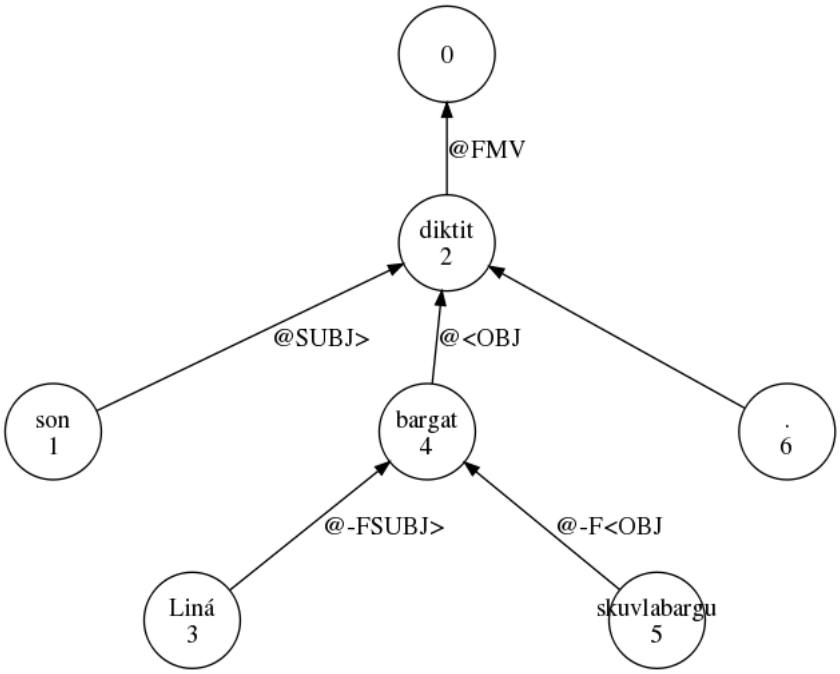

Figure 2.4: The dependency structure of *Son divttii Liná bargat skuvlabargguid.* in *Giella-sme*, cf. ex. (43-b),
@FMV = finite main verb, @SUBJ> = subject, @-FSUBJ> = subject of a non-finite verb, @<OBJ = object, @-F<OBJ = object of a non-finite verb

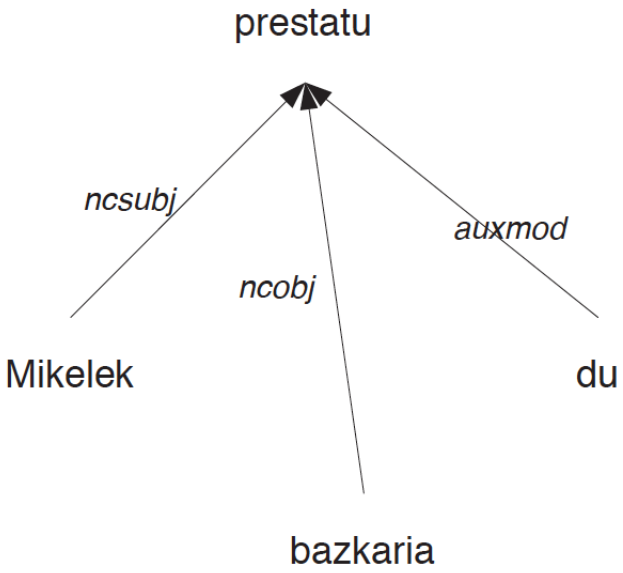

Figure 2.5: The dependency structure of *Mikelek bazkaria prestatu du.* (Aranzabe, 2008, p.89), cf. (43-c)
ncsubj = non-clausal subject, ncobj = non-clausal object, auxmod = auxiliary

Key term 4 (Lemma) *A* lemma *is the citation form of an entry in the North Sámi lexc dictionary. The lemma is similar to the lexeme. However, the lemma is not used as a morphological concept, but its definition depends on the objectives of the Giella North Sámi transducer. It may comprise several orthographic variations (i.e. virtet vs. firtet '(to) become nice weather'), multi-words separated by a space (e.g. dan botta go) and lexicalized versions of dynamic derivations, i.e. basadit (as opposed to bassat Der/d).*

Key term 5 (Auxiliary) *An* auxiliary *is a verb that does not govern any semantic arguments (↔* governing verb*).*

Key term 6 (Argument) *An* argument *is an obligatory or facultative dependent of a governor that receives a semantic role from its governor (↔* free modification*). An argument is furthermore specific for this particular governor and distinguishes it from other governors.*

Key term 7 (Free modification) *A* free modification *or an* adjunct *is a facultative dependent of the governor of a clause, which is unspecific to its governor and can be added freely. It does not belong to the valency of a governor.*

Key term 8 (Dependent) *A* dependent *is a lexeme which gets its morphological form from its governor. A syntactical dependent and its governor stand in a daughter-mother relationship to each other.*

Key term 9 (Semantic role) *A* semantic role *is a semantic relation between a governor and an argument in a certain context. Except for specific constructions (i.e. identity, causativity) each semantic role appears only once with respect to a specific governor and each argument receives only one semantic role.*

Key term 10 (Selection restriction) *A* selection restriction *is a restriction made by a governor to the inherent semantic features of its argument.*

Key term 11 (Inherent semantic feature) *An* inherent semantic feature *is a basic conceptual component of a lexeme's intension and can be either binary or a prototype. It is inherent in the lexeme because it does not depend on its use in a particular relation (↔* semantic role*).*

Key term 12 (Semantic prototype) *A* semantic prototype *bundles inherent semantic features and serves as a common denominator for a group of lexemes. While some lexemes can be more central members of the semantic prototype (i.e. sharing many distinctive features) others can be peripheral members of a prototype (i.e. sharing less distinctive features).*

Key term 13 (Valency frame) *A* valency frame *is a sequence of arguments of a governor with semantic role/syntactic and selection restriction specifications. Several independent valency frames can show the potential of a lexeme.*

Key term 14 (Rection) Rection *refers to morpho-syntactic requirements of a governor to its arguments. Unless otherwise specified,* rection *is used only for verbal/nominal/adjectival governors with locative/comitative/illative requirements to their arguments and abstract semantics.*

Key term 15 (Transitivity) Transitivity *is the ability of a lexeme to have an object in accusative case. A* transitive verb *is a verb that can but does not have to have an object in accusative case. An* intransitive verb *is a verb that cannot have an object in accusative case.*

Key term 16 (Multi-word verb) *A* multi-word verb *is a verb that consists of several tokens, which make up a single governor with its own valency. By this definition the term also covers what have been referred to in the literature as* phrasal verbs*,* light verb constructions*, and* incorporated verbs*.*

Key term 17 (Case) Case *is the morpho-syntactic marking of a noun, adjective, numeral, or pronoun for its grammatical function with respect to a governor.*

Part II

Middle

Chapter 3

Valency annotation

In this chapter, I discuss different types of valencies in North Sámi and valency annotation in the North Sámi constraint grammar *valency.cg3*.

Typically, a governor can appear in a number of different argument constellations, some of which are specified in North Sámi dictionaries. In his dictionary of North Sámi, Nielsen (1932-1960*b*, p.379) lists the verb *jápmit* 'die' with the following valency options: it can have an argument in accusative case, restricted to a cognate object, i.e. the object is related to the verb, like *jápmima* 'death (Acc.)' in ex. (1-a). It can also have a REASON-argument in locative, illative or comitative case, cf. ex. (1-b)–(1-d). The LOCATION-argument can be realized by a noun phrase in illative or relative case or the respective adverb.

(1) a. jápmit lunddolaš **jápmima**
die natural death.ACC
'die a natural death' (Nielsen, 1926-1929, p.379)

b. jápmit **nealggis**
die hunger.LOC
'die of hunger' (Ibid.)

c. jápmit **nealgái**
die hunger.ILL
'die of hunger' (Ibid.)

d. daid **háviiguin** jámii
the wound.COM.PL die.PRT.3SG
's/he died of the wounds' (Ibid.)

However, regular dictionaries are seldom an exhaustive source of valencies, and corpus research is necessary to show a governor's full valency potential, unless a valency dictionary is available. In *SIKOR*, the verb *jápmit* 'die' also appears in other argument constellations besides those mentioned by Nielsen (1932-1960*b*). These include other realizations of the REASON-argument: e.g. 80 occurrences (2.39%) of adpositional phrases in *geažil* 'because of' (cf. ex. (2)) and 27 occurrences of adpositional phrases with *dihtii/dihte* 'because of' (0.81%).

(2) ...eanet sáhttet jápmit unnán biepmu **geažil**.
...more can die little food.GEN because.of
'...more may die because of little food.'

The chapter starts out with a presentation of valency tags in rule-based applications, in particular in Eckhard Bick's various constraint grammars for semantic role annotation and verb sense disambiguation. Secondly, I present the annotation grammar *valency.cg3*. There, I discuss potential governors, i.e. lexemes receiving valency tags, and different types of valency frames. I also illustrate the structure of the valency tags, specifying semantic roles, morpho-syntax, and selection restrictions. Additionally, I present concrete valency rules and resulting verb classes. Lastly, I evaluate the coverage of both tokens and types of the valency tags in *SIKOR*.

3.1 Background

While many researchers/developers agree that valency resources are essential for rule-based natural language processing tasks, including "syntactic disambiguation and language understanding, as well as for advanced applications such as question answering, machine translation, and text summarization", cf. Estarrona et al. (2016, p.3) in their article about semantic role annotation in Basque, to my knowledge there are very few rule-based tools that actually include valency resources. Generally, the valency database is constructed prior to and independently of the tool it is to be used in, cf. Estarrona et al. (2016, p.15).[1] Valencies therefore cannot be tested with respect to their functionality for specific tools while being developed. Alternatively, the valency resource can immediately be integrated in a specific module used by the tool, i.e. as secondary tags that are "not themselves intended for disambiguation", added to regular lexicon entries, cf. Bick (2013, p.442). The advantage is that valency tags can be tested immediately and changed quickly if they serve the purpose of a specific rule better. For the North Sámi system, the tools are therefore being developed simultaneously with the linguistic resources.

Valency tags are used in a number of rule-based Constraint Grammar tools with good results, for example, within automatic semantic role annotation for Portuguese (Bick, 2007*a*) and for Spanish (Bick and Valverde, 2009), within *FrameNet* conversion for Danish (Bick, 2011), and verb sense disambiguation (English) (Bick, 2012), cf. Table 3.1. Precision is between 75% and 90% and recall between 80% and 89%.

Bick's valency tags are clearly syntactic. Bick (2012) establishes links between syntax and semantics to draw conclusions about typical syntactic realizations of semantic roles in specific languages. He points out that, while PATIENTS, ACTIONS and RESULTS are realized as direct objects, RECIPIENTS and BENEFICIARIES are realized as indirect objects

[1] "Furthermore, we conclude that to secure satisfactory results, an essential step in the methodology is to edit each verb entry completely before beginning to annotate its specific instances."

in Danish.[2] Bick (2007*a*) further uses semantic roles to generalize over "different surface-syntactic functions, like subject and object [...] depending on the arguments-slots of the governing verb and whether the syntactic structure is active, passive, reflexive or attributive". Semantic role mapping relies on separate modules for syntactic analysis (Bick, 2000), dependency and semantic prototype annotation (Bick, 2006*a*), but valency tags are directly integrated into the grammars instead of being developed independently of them. Bick and Valverde (2009) stress the interdependency between good results of automatic role labeling and a previous syntactic analysis, on the one hand, and a valency lexicon, on the other hand. Valency information in the Danish *DanGram* parser dictionary is coded into secondary tags of the following type:

```
<vdt> (ditransitive verb)
<ve> (ergative verb)
<por^vp> (prepositional valency with por)
<vk> (copulaverb)
<vta> (transobjective adverbial valency)
```

A single valency tag refers to a single argument of a specific morphological form ($<por^{v}p>$) or to several arguments ($<vdt>$). Bick (2007*a*), on the other hand, uses about 80 sets with 1,100 verb lexemes to codify valencies for semantic role annotation. The set *VPEMTH*, for example, contains verbs with a THEME-argument realized by a prepositional phrase introduced by the preposition *em* 'on, in', e.g. *crer em* 'believe in'. While the previous secondary tags encode morpho-syntactic information, these sets encode both the semantic roles and morpho-syntactic properties of an argument. Typically, these sets define only one argument instead of a constellation of arguments. Bick and Valverde's (2009) set *V-SP-SUBJ*, defining verbs with a SPEAKER-subject (e.g. *contar* 'tell', *decir* 'say', *hablar* 'talk'), is used in the semantic role-annotating rule below. This rule annotates the SPEAKER *§SP* to the subject with a member of the *V-SP-SUBJ* set as its parent (*p*).

```
LIST V-SP-SUBJ = "contar" "decir" "hablar" ...

MAP (§SP) TARGET §ARG1& (p V-SP-SUBJ);
```

Rules like the one below make reference to semantic prototypes as well, i.e. *N-LOC* (semantic prototypes with locative meaning). This rule "assigns the role 'destination' (§DES) to a dependent of preposition (@P<) if its semantic prototype is in the set N-LOC (that contains the semantic prototypes related with a locative meaning) and its parent is in the set of prepositions PRP-DES (that contains prepositions that typically

[2]"Even in a case-poor language like Danish, we found some clear likelihood relations between thematic roles and syntactic functions (table 2). Thus, agents (§AG, §COG, §SP) are typical subject roles, while patients (§PAT), actions (§ACT) and results (§RES) are typical direct object roles, and recipients (§REC) and beneficiaries (§BEN) call for dative object function."

	Portuguese (Bick 2007)	Spanish (Bick 2007)	Danish (Bick 2011)	English (Bick 2012)
F-score	88.6%	81,6%	85,12%	79.91%
Precision	90.5%	75.4%	85.20%	79.32%
Recall	86.8%	89.0%	85.05%	80.49%

Table 3.1: The performance of Constraint Grammar tools that make use of valencies

introduce this role, like *hasta (till), en dirección a (towards)*, etc.)" (Bick and Valverde, 2009).

```
MAP (§DES) TARGET @P< (0 N-LOC LINK p PRP-DES);
```

Rules such as the one below also make reference to derivations, i.e. *NDEVERBAL* (deverbal nouns) (Bick, 2007*a*). This rule maps the PURPOSE-role *§FIN* to the preposition *para* 'for', if the argument (*@P<*) in question is a deverbal noun unless the complement of *para* 'for' is of the human or place prototype category.

```
MAP (§FIN) TARGET @P<        (0 NDEVERBAL)(NOT 0 N/PROPLOC OR N/PROPHUM)
                             (*1 PRP LINK 0 PRPPARA);
```

Valencies are also crucial in rule-based verb sense disambiguation or lexical selection as a part of machine translation, cf. Bick (2007*b*). Verb sense disambiguation can be performed via valency disambiguation as according to Bick (2012), verb senses can almost always be distinguished via their valency frames, by means of either their syntactic or semantic specifications. Bick (2007*b*) refers to valencies in his *Dan2eng* machine translation system by means of "contextual distinctors", which are similar to valency tags as they define separate arguments with respect to a governor. These distinctors are used to distinguish between the different senses of, for example, the polysemous Danish verb *regne* 'rain, calculate, consider, expect, convert, include' by means of defining particular types of daughters, i.e. dependents of the verb in question. Each argument sequence corresponds to a particular translation equivalent. For the selection of 'rain' as a translation of *regne*, there has to be a formal subject *@S-SUBJ*. The translation equivalent 'consider' is associated with two arguments/dependents, i.e. a human accusative argument *(<H> @ACC)* and another argument introduced by the preposition 'for' *("for" PRP)*. In the case of 'count', not only the direct dependent, i.e. the preposition, is defined, but also the dependent of the preposition, which needs to be of the semantic prototype human (*<H>*).

```
regne_V1
(a) D=(@S-SUBJ) :rain;
(b) D=(<H> @ACC) D=("for" PRP)_nil :consider;
(c) D=("med" PRP)_on GD=(<H>) :count;
(d) D=("med" PRP)_nil :expect;
(e) D=(@ACC) D=("med" ADV)_nil :include;
(f) D=(<H> @SUBJ) D?=("på" PRP)_nil :calculate;
```

Bick (2011) uses verb valencies for verb sense disambiguation when automatically constructing a Danish *FrameNet*. In a later publication Bick (2017*b*) extends his valency annotation to 741 nominal governors. The Danish *FrameNet* is built by means of a Constraint Grammar frame tagger. The rules are converted automatically from "frame distinctors" by means of a converter program (*framenet2cgrules.pl*). The converter rule below adds *FrameNet* frame tags (*<fn:consist>*) and "argument relation tags" (*<r:SUBJ:HOL> <r:PIV:PART/MAT>*), i.e. valency tags, to a verb (here *bestå av* 'consist of') in a particular syntactic/semantic context. Frame tags specify a verb class with syntactically/semantically coherent verbs. Relation tags, on the other hand, specify the arguments of the governors, i.e. in this case a subject of the role-type *HOL* and a prepositional object of the role-type part/material, and both their syntactic relation (i.e. subject or prepositional object) and their semantic role (i.e. whole, part/material). Semantic prototypes and morpho-syntactic constraints are not specified in the relation tags themselves, but rather in the context conditions (e.g. *(1 (*) LINK *-1 VFIN LINK c @SUBJ LINK 0 <cc>)*).

```
SUBSTITUTE (V) (<fn:consist> <r:SUBJ:HOL> <r:PIV:PART/MAT> V) TARGET
("bestå" <mv> V)(1 (*) LINK *-1 VFIN LINK c @SUBJ LINK 0 <cc>)
(c @PIV LINK 0 ("af") LINK c @P< LINK 0 <cc> OR <mat>) ;
```

Then, semantic roles *(§$1)* that match the relation tags are mapped to the verb's arguments by means of regular expressions (*<r:ACC:.∗>r*). Here any role specified in the accusative relation tag of a governor is annotated to the subject *@SUBJ* of a passive verb (*PAS*). The semantic role of the accusative argument of an active verb form is mapped to the subject of a passive verb form (Bick, 2012).

```
MAP KEEPORDER (VSTR:§\$1) TARGET @SUBJ
(*p V LINK -1 (*) LINK *1 (<r:.*>r) LINK 0 PAS LINK 0 (<r:ACC:\(.*\)>r)) ;
```

The frame tagger annotates *FrameNet* senses (e.g. *<fn:establish>* or *<fn:decrease>*) and simple valency tags (e.g. *<v:vt>*) to the verbs (e.g. *nedsætter* 'set up') and semantic role tags to the verb's arguments, i.e. the AGENT-role *§AG* to *regeringen* 'government' and the RESULT-role *§RES* to *kommission* 'commission', cf. ex. (3).

(3) Nu nedsætter **regeringen** en *kommission*.
now set.up.PRS.3SG government a commission
'Now the government is setting up a commission.'

```
Nu "nu" <atemp> ADV @ADVL> #5->6
nedsætter "nedsætte" <mv> <v:vt> <fn:establish>
PR AKT @FS-STA #6->0
regeringen "regering" <HH> N UTR S DEF NOM
@<SUBJ §AG #7->6
en "en" ART UTR S IDF @>N #8->9
kommission "kommission" <HH> N UTR S IDF
NOM @<ACC §RES #9->6
```

3.2 The valency annotation grammar *valency.cg3*

In this section, I present the valency annotation grammar *valency.cg3*. The grammar consists of *Constraint Grammar* rules that add valency tags to a target in a specific context. First, I present and discuss the targets of the annotation process, i.e. the potential governors that receive certain valency frames. Secondly, I discuss specific argument constellations represented in the corpus that make up the valency frames described in the valency tags. Thirdly, I present the structure of the valency tags, focusing on their functionality for specific rule-based applications and their ability to describe syntactic phenomena in the North Sámi grammar. While Bick's (2007*c*) valency tags consist of morpho-syntactic specifications, Bick's (2011) relation tags add semantic role specifications and selection restrictions are referred to in rules, I specify all three levels of valency in explicit valency tags, i.e. semantic roles, morpho-syntax and selection restrictions. Bick (2017*a*, p.209) notes that a combination of linguistic information on these three linguistic levels ("syntactic function, semantic ontology and semantic role") can also be encoded "implicitly through annotated data" as he shows in the automatic annotation of a VerbNet Corpus for Danish. Lastly, I describe the structure of the annotation grammar and specific annotation rules.

3.2.1 Governors

Valency tags are annotated to a token specified in the target of a valency rule in *valency.cg3*, i.e. a potential governor. The tokens are referred to by means of a lemma specified in *lexc* and additional morpho-syntactic tags, i.e. valency rules are explicit with regard to their targets. As the annotation is done by means of hand-written rules, the targets are deliberately chosen. This choice is based on linguistic considerations (cf. also Chapter 2), corpus coverage and peculiarities of the *Constraint Grammar* formalism.

Formalism-related issues involve the token-based nature of *Constraint Grammar* and the lexicon structure. However, in multi-word expressions, a lexeme is made up of more

than one token, and its properties are not derived from the properties of the individual tokens. In *Constraint Grammar*, valency tags are mapped onto one token, i.e. the lemma listed in the lexicon, which in the case of the multi-word verb *atnit árvvus* 'value (lit. consider of value)' is only *atnit* 'consider'. Typically, only multi-word expressions that cannot be interrupted by other tokens in a sentence, e.g. *iešguđet ládje* 'in different ways', are listed as single lemmata in the *lexc*-lexicon. Derived verbs can either be listed as underived lemmata with a derivational tags or directly as lexicalized derived verbs in the lexicon. This distinction is made deliberately and needs to be taken into account when constructing valency rules. The form *beaškalit* 'slam (transitive)' for example is only listed as a lexicalized lemma, and not as a derived form of the intransitive verb *beaškit* 'slam (intransitive)'. The verb *bidjalit* 'put (once or quickly)', on the other hand, is listed both as a derivation of *bidjat* 'put' and under the lemma *bidjalit*.

While valency annotation deliberately focuses on verbs, a number of nouns, adjectives and adverbs also receive valency tags in *valency.cg3*. Noun valencies are relevant for compound error detection. Compound error detection is based on lexicalized compounds in the lexicon, e.g. *atnudávvirat* 'use artifacts' and *atnuávnnasin* 'use fabric', and on nominal valency tags. In ex. (4-a) and (4-b) the compounds *atnu dávvirat* 'use artifacts' and *atnu ávnnasin* 'use fabric' are written as two words, which violates the norm. However, depending on the syntactic context, the potential compounds can also be interpreted as single nouns that stand in a semantic relation to each other, which is when they should be written apart. This relationship can be defined by valency tags, describing e.g. the illative valency of *atnu* 'use' with the tag *<TH-Ill-Any>*. When the word to the right of *atnu* 'use' is in illative case as in *lieggabiktasiidda* 'warm clothes (Ill. Pl.)' in ex. (4-c), a rule referring to the valency tag discards the compound-reading even if there is a lexicalized compound.

(4) a. Buot dát ledje **atnu dávvirat**, maidda beaivválaččat lei dárbu.
all these were use.NOM artifacts, which daily was need
'All these were commodities which were needed daily.'

b. Guollenáhkki lea árbevirolaš materiála mii geavahuvvo sihke
fish.skin is traditional material which use.PASS.PRS.3SG both
čikŋan **atnu ávnnasin**.
decoration.ESS use.NOM thing.ESS
'Fish skin is a traditional material that is used both for decoration and as a fabric.'

c. Ii leat **atnu lieggabiktasiidda** dáppe siste.
not is use.NOM warm.clothes.ILL.PL here inside
'There is no use for warm clothes in here.'

3.2.1.1 Coverage

In order to achieve good corpus coverage, initial valency annotation in *valency.cg3* is based on the 500 most frequent verbs from *SIKOR*,[3] cf. Table 3.2, which are annotated with their full valency potential. While good coverage is the main objective when annotating valencies, it is not the only one. I also annotated verbs that are confused with any of the 500 most frequent verbs (causing real word errors), and typical valency error candidates. Additionally, I aim to achieve good coverage of different types of valency frames found in linguistic descriptions, e.g. Mikalsen's (1993) classification of approximately 150 *rection*-verbs, Sammallahti (2005), Nickel and Sammallahti (2011), Nielsen (1926-1929), etc. As linguistic descriptions seldom provide the full valency potential of a lexeme, I use *SIKOR* to test represented valency frames. Valency frames in linguistic descriptions and corpus material can diverge, which is why informant introspection is necessary as a third source in valency annotation.

The most frequent verbs in Table 3.2 include many auxiliaries/copulas such as *leat* 'be', *ii* 'not', *sáhttit* 'can', *galgat* 'be, not, can, shall' on the upper end. The motion verb *boahtit* 'come' is the governing verb with the highest frequency (4,355 occurrences). On the lower end, there are verbs with just above 700 occurrences such as *neaktit* 'act' (736), *gávppašit* 'shop' (736), and *gudnejahttit* 'honor' (733). Not all verbs are governing verbs, some are auxiliaries and copulas. The frequency list is influenced by the distribution of texts in the corpus, i.e. verbs that appear in administrative texts such as *ovddidit* 'promote' (21,878), *hálddašit* 'administer' (7,276), *doaimmahit* 'execute', (4,241), and *gieđahallat* 'deal with' (3,993) are well represented in *SIKOR*. In this chapter, the approach to valencies in *SIKOR* is mostly descriptive. However, in the context of grammar checking (cf. Chapter 5), only normative valencies (or valencies where no clear norm could be identified) are annotated.

3.2.1.2 Lexicon and morphological processes

Derivational processes can change the valency structure of a verb, e.g. reduce or increase the number of arguments, as in the case of the causative *lávlluhit* 'make sing' compared to the non-causative *lávlut* 'sing', cf. ex. (5). While the non-causative verb only has one argument in nominative case (e.g. *oahppit* 'students'), the causative verb requires two arguments (e.g. *oahpaheaddji* 'teacher' and *ohppiid* 'student (Acc. Pl.)').

(5) Oahppit **lávlot**. – Oahpaheaddji **lávlluha** ohppiid.
students sing.PRS.3PL – teacher sing.CAUS.PRS.3SG student.ACC
'The students sing. – The teacher makes the students sing.' (Sara, 2002, p.44)

[3](Accessed 2012-06-01). At the time of access, it consisted of 23,603,053 tokens altogether. It is a previous version of *SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection* (2015-03-01).

Frequency	Verb	Frequency	Verb
1,088,872	*leat* 'be'	...	...
273,088	*ii* 'not'	753	*boradit* 'eat'
121,456	*galgat* 'shall'	751	*divodit* 'repair'
99,295	*sáhttit* 'can'	749	*einnostit* 'anticipate'
70,422	*oažžut* 'get'	744	*váibat* 'get tired'
62,462	*lohkat* 'read, claim'	744	*soabadit* 'agree on'
52,506	*fertet* 'must'	744	*ovdanboahtit* 'emerge'
48,625	*boahtit* 'come'	744	*logahallat* 'enumerate'
45,773	*šaddat* 'become'	742	*guoimmuhit* 'entertain'
37,702	*muitalit* 'tell'	738	*spiehkastit* 'deviate'
35,396	*dahkat* 'do'	738	*revideret* 'revise'
34,938	*bargat* 'work'	736	*neaktit* 'resemble'
34,499	*dadjat* 'say'	736	*gávppašit* 'shop'
34,311	*váldit* 'take'	733	*gudnejahttit* 'honor, respect'
34,006	*addit* 'give'	728	*jorrat* 'spin, turn'

Table 3.2: Some of the 500 most frequent verbs in *SIKOR*

The *valency.cg3* grammar is applied on top of a morphological analyzer and a lexicon, which match a given word form with its lemma and a tag sequence consisting of part of speech tags and other morphological tags. Morphological tags include both inflectional processes (e.g. case marking, number and tense) and derivational processes. Both part of speech-changing and non part of speech-changing derivations can affect the lemma's valency. Derived verb forms can sometimes receive more than one analysis, one of which is based on the same baseform as the non-derived verb and a derivational tag, the latter of which uses the derived form as its baseform, i.e. the derived verb is lexicalized. This is why valency rules must not only specify the lemma, but also make positive and/or negative constraints with regard to morphological tag combinations.[4] Both derivation- and inflectional tags, cf. Table 3.3, can change a verb's qualitative or quantitative valency. Part of speech-changing derivational tags like the deverbal noun tags *Actor*, *Der/NomAct*, and *Der/muš* or deverbal adjective tags such as *Der/ahtti* can change both qualitative and quantitative valency. Non part of speech-changing tags such as verbal derivational tags, passive *Der/PassL*, causative *Der/h*, and continuative *Der/d* and inflectional tags, for example *Actio Loc*, can also be valency changing. In ex. (6-a), *riidaleaddji* 'arguing' is a present participle form of *riidalit* 'argue', and is used in attributive position. Although a verb form, it is not used with a comitative argument like most other inflected forms of *riidalit* 'argue'. The non-finite actio locative form *riidaleames* 'arguing' in ex. (6-b) typically does not have a comitative argument either.

[4]For a full overview of tags in *lexc* cf. Appendix C.

Form	**Derived from**	**Tag combination in *lexc***
Part of speech-changing		
gohčču̇t 'commanders'	*gohčču̇t* '(to) command'	Actor N Pl Nom
ealihahtti 'worth-living'	*ealihit* '(to) sustain'	Der/ahtti Actor N Sg Nom
vuhtiiváldámušaid 'things to pay attention to'	*vuhtiiváldit* '(to) take into account, pay attention to'	Der/muš N Pl Gen
njiedjan 'the decrease, descent'	*njiedjat* '(to) decrease, descend'	Der/NomAct N Sg Nom
Non part of speech-changing		
vuhtiiváldin 'paying attention'	*vuhtiiváldit* '(to) take into account, pay attention to'	Actio Gen @>P
čilgejuvvojit 'be explained'	*čilget* '(to) explain'	Der/PassL IV Ind Prt Sg2
oaččohit 'at last get sb. to do sth.'	*oažžut* '(to) get'	Der/h V Inf
giccodit 'climb for a long time'	*gizzut* '(to) climb'	Der/d V Ind Prs Pl3
riidaleaddji 'arguing'	*riidalit* '(to) argue'	PrsPrc
riidaleames 'arguing'	*riidalit* '(to) argue'	Actio Loc

Table 3.3: Part of speech-changing and non part of speech-changing derivations and inflections in North Sámi that affect valencies

(6) a. ...gosa **riidaleaddji** báhpat gullet.
...where argue.PRSPRC priests belong
'...to where arguing priests belong.'

b. ...jus eai heaitte **riidaleames** dego beatnagat ...
...if not stop argue.ACTIO.LOC like dogs ...
'...if they don't stop arguing like dogs ...'

Nickel's (1994) description of 11 non part of speech-changing verbal derivations (p.221), cf. Table 3.4, includes both derivations that change the verb's valency and those that do not. While the obvious candidates for change in the valency structure are passives and causatives, even derivations that preserve a certain valency-frame can have different preferences than the underived form. Their annotation is discussed further in Section 3.2.3.3.

Derivation	**Tag in *verbs.lexc***	**Example**	**(Optional) valency change**
passive	Der/Pass	*borrat - borrojuvvot* 'be eaten'	− Acc. argument
passive	Der/halla	*borrat - borahallat* 'be eaten'	− Acc. argument, + (Ill. argument)
causative	Der/h, Der/ahtti, Der/d	*goarrut - goaruhit* 'cause to sew'	+ (Acc. or Ill. argument)
reflexive	Der/d, Der/alla, Der/adda	*bassat - basadit* 'wash oneself'	− Acc. argument
reciprocal	Der/d, Der/alla, Der/adda	*dovdat - dovddadit* 'know each other'	− Acc. argument
momentous	lexicalized	*doahput - dohppet* 'grab (once)'	no change
subitive	Der/l	*borrat - borralit* 'eat quickly'	− (Acc. argument)
frequentative	Der/alla, Der/d, Der/adda	*suokkardit - suokkardallat* 'investigate (several objects / several times)'	no change
continuative	Der/d	*borrat - boradit* 'have a meal'	− (Acc. argument)
diminutive	Der/st	*addit - attestit* 'give a little'	no change
conative	lexicalized	*oažžut - oččodit* 'try to get'	no change
inchoative	Der/InchL	*lohkat - lohkagoahtit* 'start to read'	no change

Table 3.4: North Sámi verbal derivations according to Nickel (1994, p.221) and their effects on valency

3.2.1.3 Lexicon and syntactic issues: Multi-word verbs

Governors can consist of several word forms, which govern their arguments collectively. These governors are referred to here as multi-word verbs. Multi-word verbs include a verb and one or several other word forms (nominal, adpositional, adverbial or adjectival), which make up a semantic unit. However, in *valency.cg3*, valency tags are annotated to a lemma typically consisting of one token, which is why multi-word verbs need a special treatment. Multi-word verbs are frequent in *SIKOR* and need to be annotated in an adequate way to achieve successful matching of governors and their arguments.

I distinguish between copula–adjective constructions as in ex. (7-a), where the adjective is considered to be the governor and receives the valency tag, cf. also Haugen (2013, p.36)[5], and multi-word verbs as in ex. (8), where the verb receives the valency tag, but it contains a reference to the other token. In ex. (7-a), the adjective *giitevaš* 'thankful' is annotated for its ability to appear with a REASON-argument in locative case, i.e. *<RS-Loc-Any>*. Since the potential to be a governor is restricted to adjectives in predicative form/position, one could think of the copula–adjective construction as a multi-word verb, cf. its attributive use in ex. (7-b).

(7) a. Ledjen giitevaš **dan ráhkisvuođas**, man bessen vásihit ...
was thankful this love.LOC, which got experience.INF ...
'I was thankful for this love, which I got to experience ...'

b. Bálkkašupmi addá doaivvu, movttiideami ja dehálaš duođaštusa,
prize gives hope, encouragement and important confirmation,
muitala **giitevaš** festiválajođiheaddji.
tells thankful.ATTR festival.leader
'The prize gives hope, encouragement and important confirmation, says a thankful festival leader.'

However, also in constructions where the copula is missing, the adjective keeps its valency frame, which is why I annotate the valency tag to the adjective in predicative form. In the case of adjectives which are ambiguous as to their predicative and attributive form, the valency is annotated without further constraint although it may require a copula in its right or left context. Multi-word verbs are also often referred to as "phrasal verbs" (in the case of verb-particle constructions), "incorporated verbs" (cf. Bick (2011)) or "light verb constructions" (cf. Kettnerová and Lopatková (2013)). In the case of the multi-word verb *bidjat johtui* 'launch' as in ex. (8), the valency is annotated directly to the verb containing a reference to the second part of the multi-word verb, e.g. *johtui* 'motion (Ill.)'. In the case of *bidjat* 'put', the noun *johtu* 'motion' in illative case is part of the complex governor requiring an argument in accusative case. The verb *bidjat* 'put' alone, on the other hand,

[5]"As already mentioned, research on valency has mainly been concerned with verbs, but it is clear that adjectives can, often in combination with copula verbs, play a similar role in determining the basic structure of a clause."

typically asks for a THEME-argument in accusative case and a DESTINATION-argument in illative case. While *bidjat* is translated as 'put', *bidjat johtui* is translated as 'start'. Syntactically, *johtui* 'motion (Ill.)' is analyzed as an adverbial, cf. Sammallahti (2005, p.154). Bick (2011) also assigns syntactic tags (e.g. an object tag) and dependencies to "noun incorporations", i.e. nominal forms that belong to a multi-word verb. In semantic role annotation, he distributes a special tag referring to the multi-word status of the noun ("§INC (incorporate)").

(8) Mii leat dál álggos **bidjan johtui** gulaskuddamiid ...
we have now beginning.LOC put motion.ILL hearing.ACC.PL ...
'To begin with, we have launched the hearings ...'

Like Bick (2011), I annotate the valency of a multi-word verb to the verb with a reference to the other parts of the governor. However, Bick's (2011) transitivity tags for Danish, cf. the examples below, do not refer to semantic roles. His "transitivity tags", i.e. valency tags, refer to one or several incorporated elements after a general transitivity specification. The tag *<vi-op>* describes an intransitive verb with an incorporated preposition, *-op*, as in *kaste op* 'vomit'. The tag *<vt-i=sinde>*, on the other hand, describes a transitive verb with an incorporated prepositional phrase consisting of a preposition, *-i*, and a noun, *sinde*, as in *have i sinde* 'have in mind'.

```
kaste op 'vomit' - <vi-op>
slå fra 'deactivate' - <vt-fra>
komme ind på 'discuss' - <på^vt-ind>

holde kæft 'shut up' - <vt-kæft>
have brug for 'need' - <for^vtp-brug>

have i sinde 'intend' - <vt-i=sinde>
være på færde 'be going on' - <vi-på=færde>
```

In *SIKOR*, several frequent multi-word verbs can be found, four of which I investigated with regard to their valency, i.e. *atnit árvvus* 'value', *bidjat johtui* 'launch', *váldit vuhtii* 'take into account' and *váldit vára* 'take care', cf. Table 3.5.

The multi-word verbs are combinations of verbs and inflected forms of nouns (*árvvus* 'value (Loc.)' and *johtui* 'motion (Ill.)'), some of which are lexicalized as adverbs (*vuhtii* 'into account' and *vára* 'care'). I consider them multi-word verbs as they behave differently syntactically and semantically from the simple verb construction. In North Sámi human-readable dictionaries, these constructions are stored under both noun- and verb-entries and do not contain complete valency information, cf. Sammallahti and Nickel (2006, p.31, pp.719–720).[6] Nielsen (1926-1929, p.729) mentions a number of multi-word verbs as

[6] **árvu** [...] **atnit** *vt* **árvvus** ehren *vt*, in Ehren halten, *vt unreg*, achten *vt* ..." and "**váldit** [...] **vára** (juogamas) sich an|nehmen [...] **vuhtii** beachten *vt*, berücksichtigen *vt*, in Betracht ziehen ..."

Multi-word verb	***SIKOR***	**Morpho-syntactic valency distribution**	
		Active	**Passive**
atnit árvvus 'value'	1,055	accusative (725), *locative (6), illative (5), *nominative (3), none (58), *go*-subclause 'that' (21), *ahte*-subclause 'that' (8), *jus*-subclause 'if' (1), question subclause (5), infinitive (2)	nominative (222)
bidjat johtui 'launch'	1,823	none (16), accusative (1,327)	none (480)
váldit vuhtii 'take into account'	2,991	accusative (222), *locative (13)	-
váldit vára 'take care'	597	locative (404), *accusative (80), none (15), *comitative (5)	nominative (1), locative (1)

Table 3.5: Four multi-word verbs and their THEME realizations in *SIKOR*

well, however, without referring to their valencies.[7] Sammallahti and Nickel (2006, p.31, pp.719–720) only mention the locative valency of *váldit vára*; subclauses with *ahte* 'that' are not described. These nouns and adverbs are also used in simple constructions. The noun form *árvvus* 'value (Loc.)' can be used as a complement of verbs with a locative valency, i.e. *beroštit* 'care' as in ex. (9-a) or *hupmat* 'speak', as in ex. (9-b). In ex. (9-c), on the other hand, *atnit árvvus* 'value' is a multi-word verb with an accusative valency. However, between 87% and 99% of the occurrences of these nouns/adverbs are part of a multi-word verb, cf. Table 3.6.

(9) a. ...eai ge beroš oahpahusa kvalitehtas ja dan **árvvus**
...not either worry teaching.GEN quality.LOC and its value.LOC
'...they do not care about the teaching quality and its value'

b. Easkka dalle sáhttá hupmat duohta dohkkeheamis ja **árvvus**.
first then can talk true acceptance.LOC and value.LOC
'First, then, one can talk about true acceptance and value.'

c. Vuorasolbmuid máhttu lea dehálaš sámeservodaga ovddideamis
adult.GEN.PL knowledge is important Sámi.society development.LOC
máid politihkkarat galget **atnit árvvus**
which.ACC politicians shall consider value.LOC
'Adult knowledge is important for the development of the Sámi society, which the politicians shall value'

The multi-word verb *váldit vuhtii* 'take into account' is with almost 3,000 occurrences the most frequent of the multi-word verbs described here, followed by *bidjat johtui* 'launch', with 1,823 occurrences. In *SIKOR*, each of the verbs appears in two valency

[7] e.g. "val'det [...] vūttii: ta i betraktning"

Noun/adverb	*SIKOR*	As a mwv		Verbal distribution
árvvus 'value (Loc.)'	1,221	1,155	94.6%	*atnit* 'consider' (1,055), *leat* 'be' (62), *doallat* 'hold' (26), other multi-word verbs (10)
johtui 'motion (Ill.)'	3,019	2,971	98.4%	*bidjat* 'put', *boahtit* 'come' (425), *oažžut* 'get' (357), *vuolgit* 'leave' (95), *beassat* 'get' (61), *čievččastit* 'kick' (17), other multi-word verbs (186)
vára 'care'	597	523	87.6%	*váldit* 'take' (520), *atnit* 'have, consider' (5)
vuhtii 'into consideration'	5,160	5,094	98.7%	*váldit* 'take' (5,094)

Table 3.6: The distribution of nouns/adverbs as part of multi-verb words in *SIKOR*

frames. The multi-word verb *atnit árvvus* 'value', on the other hand, appears with noun phrases in a number of different nominal cases (i.e. accusative, locative, illative, nominative), subclauses introduced by a subjunction (*go* 'when', *ahte* 'that', *jus* 'if') or a question pronoun, and non-finite arguments, some of which are considered ungrammatical uses by Informant *H* (marked by <*>). A number of constructions appear without any THEME whatsoever. Furthermore, *váldit vára* 'take care' appears with a number of different cases and in constructions without a THEME-argument.

Multi-word expressions are not necessarily as fixed in their combinations, i.e. some nominal forms/adverbials form multi-word verbs with several different verbs, cf. Table 3.6. I will not discuss these combinations critically with regard to their multi-word status here, and am fully aware of the fact that some of the combinations may be disputable. While the adverbs *vára* 'care' and *vuhtii* 'into consideration' appear almost exclusively with *váldit* 'take' in a multi-word verbal expression, *árvvus* 'value (Loc.)', is used with a number of rather unrelated verbs: *atnit*, *leat* 'be', *doallat* 'hold', *gahččat* 'fall', *doalahit* 'prevent', *manahit* 'lose', *massit* 'lose', *oažžut* 'get', and *goargŋut* 'climb'. The form *johtui* 'motion (Ill.)' is used mostly with transitive verbs like *bidjat* 'put' and *oažžut* 'get', and intransitive verbs like *boahtit* 'come' and *vuolgit* 'leave'. But there are also single occurrences with rather specific verbs like *beaškalit* 'slam', *spoahkkaluvvot* 'be tapped', *čavget* 'tighten, get oneself together to', and *doaimmahit* 'execute', cf. ex. (10-a)–(10-c).

(10) a. «Kákáos» prográmmaráidu **spoahkkaluvvui johtui** NRK1 kanálas
«Kákáos» program.series tap.PASS.PRT.3SG motion.ILL NRK1 channel
'The «Kákáos» series kicked off on the NRK1 channel'

b. Mii gal leat gearggus **čavget** **johtui** birra
we certainly are ready pull.oneself.together.INF motion.ILL around
jándora, lohká son.
day, said s/he
'We certainly are ready to pull ourselves together to get going all day, s/he said.'

c. . . . **doaimmahit johtui** čorgenbargguid.
. . . execute motion.ILL cleaning.work.ACC.PL
'. . . start cleaning.'

Including multi-word verbs in valency descriptions is important in grammar checking, disambiguation and machine translation. In grammar checking one wants to identify real word errors as in the case of *váldit vara* 'take blood' and the multi-word verb *váldit vára* 'take care'. While both can have a locative argument, in the case of *váldit vara* 'take blood', it needs to be human/animate (*sus* 's/he (Loc.)'). The argument of the multi-word verb, on the other hand, can be inanimate, e.g. *kulturárbbis* 'cultural heritage (Loc.)' in ex. (11-b). Since *kulturárbbis* 'cultural heritage (Loc.)' is inanimate and locative, *vara* 'blood' can can be identified as a real word error by means of its governor's valency. Valency errors can also be identified. In ex. (11-c), *váldit vára* appears with an argument in accusative case, *dáiddavuorkká* 'art archive (Acc.)', which should be locative case. While accusative case arguments can be matched with *váldit* 'take', they cannot be matched with *váldit vára* 'take care'. In the ungrammatical ex. (11-d), the passive *lea váldon* 'has been taken care of', on the other hand, appears with a BENEFICIARY in nominative case, *davvisámegiella* 'North Sámi (language)', even though the active verb does not have an accusative argument that can alternate with a nominative in a passive construction. Informant *H* prefers a BENEFICIARY in locative case, *davvisámegielas* 'North Sámi (Loc.)'. In ex. (11-e), on the other hand, the BENEFICIARY *kulturárbbiin* 'cultural heritage (Loc. Pl.)' is realized in locative case in the passive construction. In order to successfully detect the case error in ex. (11-d), the multi-word verb needs to be distinguished from the simple verb as *váldit* 'take' alone can have a nominative subject when passivized.

(11) a. olmmái lei gárremin ja su dolvo Guovdageidnui gos
man was drunk and brought him Guovdageaidnu.ILL where
doavttir válddii **sus** vara.
doctor took he.LOC blood.ACC
'the man was drunk and they brought him to Guovdageaidnu where the doctor took a blood sample from him.'

b. . . . de váldit seammás *vara iežamet **kulturárbbis**.
. . . then take at.the.same.time blood.ACC own culture.heritage.LOC
'. . . then we take care of our own cultural heritage at the same time.'

c. . . . váldit vára Savio ***dáiddavuorkká**, čájáhusaide,
. . . take care.of Savio's art.archive.ACC exhibition.ILL.PL,
seminára ide . . .
seminar.ILL.PL . . .
'. . . take care of Savio's art archive for exhibitions, seminars . . . '

d. ***Davvisámegiella** lea dál buoremusat vára váldon buot
North.Sámi.NOM has now best care taken all

sámegielain
Sámi.language.LOC.PL
'Of all Sámi languages, North Sámi has been best taken care of'

e. Lea dehálaš ahte dáin **kulturárbbiin** váldo vára.
is important that these.LOC cultural.heritage.LOC.PL take.PASS.3SG care
'It is important that the cultural heritage is taken care of.'

3.2.1.4 Linguistic considerations: Governing verb vs. auxiliary

In principle, only governing verbs receive a valency annotation in *valency.cg3*, i.e. a valency tag specifying an argument constellation with regard to semantic roles, morpho-syntax and selection restrictions. Auxiliaries, on the other hand, are annotated only with regard to their morpho-syntactic potential, i.e. their potential to appear with a non-finite form (e.g. infinitive, perfect participle, etc.). This is based on the assumption that only governing verbs have semantic arguments, while auxiliaries need to appear in periphrastic constructions with a governing verb in a non-finite form, where the full lexical verb is the governor, cf. Magga (1986, p.7).[8] However, auxiliaries and governing verbs cannot necessarily be clearly distinguished, cf. Magga (1986, p.8). The presence of a non-finite form can be a sign of the auxiliary status of the finite verb. However, it can also be an obligatory argument of the finite verb and receive a semantic role from it, or it can be free modification expressing, for example, purpose or cause.

There are prototypical auxiliaries like *leat* 'be', *veadjit* '(to) possibly', and *dáidit* '(to) probably'. In the case of other verbs, e.g. *galgat* 'shall', *áigut* 'want', *šaddat* 'become', *fertet* 'must', *sáhttit* 'can' and *viššat* 'care to do', views are diverging with regard to their status, and a number of verbs are considered to be both auxiliaries and governing verbs depending on their context (Magga, 1986, p.14). Magga (1986, p.56) also mentions diachronic change in the status of a verb, i.e. in the case of *áigut* 'want'. The verb had been previously used as an auxiliary only, but is used as both auxiliary and governing verb in more recent language use. Magga (1986, p.18) uses both morphological (e.g. incomplete inflectional paradigms), semantic and syntactic criteria to distinguish between auxiliaries and governing verbs. Here, I only take into account valency-related criteria. Magga (1986, p.22) notes that the subject, and all other parts of the sentence, in a periphrastic construction are related to the governing verb, not the auxiliary. It follows that even the subject receives its semantic role from the governing verb. The auxiliary itself is interchangeable, cf. Magga (1986, p.34). Furthermore, other arguments and free modifications describe the governing verb, not the auxiliary (Magga, 1986, pp.36–38), cf. ex. (12), where *johtilit* 'quickly' can modify either *oahpai* 'learned' or *čállit* 'write' as

[8] "Styreverb som f.eks. DÁIDIT kan ikke danne setning alene med et subjekt slik f.eks. VUOLGIT kan."

both are governing verbs.

(12) Skuvllas oahpai son **johtilit** čállit.
school.LOC learn.PRT.3SG s/he quickly write.INF
'1. At school s/he learned to write quickly.'
'2. At school s/he quickly learned to write.' (Magga, 1986, p.37)

Magga (1986, p.35) also notes that only governing verbs passivize in such a way that the object role of the active version is the same as the subject role of the passive sentence.

In *valency.cg3*, verbs that can appear with an infinitival phrase, but cannot govern a nominal phrase (e.g. *dán* 'this'), are considered auxiliaries and do not receive a valency tag. Auxiliaries are marked for their ability to have a non-finite complement with the tag *<Inf>*. The tag does not include a semantic role specification. The set *INF-V* below taken from *valency.cg3*, and originally from *disambiguation.cg3* and *dependency.cg3*, specifies verbs that are annotated with the *<Inf>*-tag. Verbs that can appear with an infinitival phrase or govern a nominal phrase are either considered to be governing verbs or polysemous, i.e. both governing verbs and auxiliaries. In the first case, the infinitive receives a semantic role with regard to its governor and the governing verb receives a valency tag. In case of polysemy, the governing verb/auxiliary receives both a valency tag specifying the semantic role of the non-finite form (e.g. *<TH-Inf>*) and the *<Inf>*-tag.

```
LIST INF-V = "arvat" "astat" "áigut" "álgit" "beassat" "berret" "boahtit" "dáhttut"
"dáidit" "dárbbašit" "diktit" "duostat" "fertet" "figgat" "galgat" "geahččalit"
"gillet" "gártat" "iskat" "háliidit" "lávet" "máhttit" "nagodit" "nagadit"
"nuhkket"  "ribahit" "seahtit" "sihtat" "soaitit" "suovvat" "sáhttit" "stađđat"
"veadjit" "viggat" ;
```

The verb *dárbbašit* 'need' is both an auxiliary and a governing verb. In ex. (13-a), it appears with the infinitive *diehtit* 'know' and is considered an auxiliary. In ex. (13-b), it appears with a direct object that is a THEME, and is considered to be a governing verb. The auxiliary *galgat* 'shall' in ex. (13-c) can appear with a DESTINATION-argument, cf. also Ylikoski (2016, p.219). Sentences like ex. (13-c) are considered ellipses of ex. (13-d)–type sentences, where the verb appears with a motion verb infinitive and a DESTINATION-argument, cf. Magga (1986, p.42),[9] and Helbig and Schenkel (1973, p.57).[10] However, in *valency.cg3*, auxiliaries that frequently appear with a DESTINATION-argument without a governing verb, e.g. *galgat* 'shall', are annotated with the valency tag *<DE-Ill-Plc>*

[9]"En vanskelighet representerer tilfelle der et bevegelses-verb er "underforstått": (116) Don dáiddat Márkanii (mannat/manname)? 'Skal du (dra) til krirkestedet?' [...] Her er det adverbialet som antyder hva som er underforstått."

[10]"Deshalb werden auch die Infinitive bei Hilfsverben nicht als besondere Mitspieler gewertet, sondern zusammen mit dem Hilfsverb als strukturelles Zentrum betrachtet. Wenn die modalen Hilfsverben im Satz allein (ohne Vollverb) erscheinen, handelt es sich um eine elliptische Reduzierung um das Vollverb, die an der Bedeutung des Satzes nichts ändert [...]"

as well. The advantage of this explicitness is that arguments in these constructions can easily be matched with the governor. The tag format is discussed further in Section 3.2.2.

(13) a. Lávdegottit *dárbbašit* **diehtit** buot áššiid birra mat gusket ...
committees need know.INF all things about that touch ...
'Committees need to know about all things that touch ...'

b. ... *dárbbašit* **ovttasbarggu** lagamus ránnjáiguin.
... need cooperation.ACC nearest neighbor.COM.PL
'... we need cooperation with the nearest neighbors.'

c. Dál vuolggán Anárii ja boahtte vahkku *galggan* **Girkonjárgii**.
now leave Inari.ILL and next week will Kirkenes.ILL
'Now I leave for Inari and next week I shall go to Kirkenes.'

d. Go gerge boradeamis, de galge **skuvlii** *vuolgit*.
when are.done eating, then will school.ILL go.INF
'When we are done eating, we will go to school.'

Governing verbs with an infinitive typically distribute the same semantic role to the infinitive as to a nominal argument. The verb *liikot* 'like' in ex. (14-a) appears with the infinitive *valáštallat* 'work out', which is a THEME. The verb *vuolgit* 'leave' in ex. (14-b), on the other hand, appears with a PURPOSE-infinitive, *borjjastit* 'sail'.

(14) a. Lean álo liikon **valáštallat**, ovdal čuigen olu ...
have always liked work.out.INF, before skied a.lot ...
'I have always liked to work out, I used to ski a lot ...'

b. Bárdni vulggii **borjjastit**.
boy went sail.INF
'The boy went sailing.'

3.2.2 Valency tags

In *valency.cg3*, valencies are mapped onto potential governors in the form of valency tags. Valency tags refer to semantic roles, morpho-syntactic specifications and selection restrictions or specific word forms in argument constellations. For a complete list of all valency tags cf. Appendix C.

Valency tags in *valency.cg3* are inspired by Bick's (2000) transitivity tags as shown in the excerpt of his tagset below (Bick, 2000, p.160). His tags refer to intransitivity, ditransitivity, and transitivity, but include references to particular cases (*ACC* (accusative), *DAT* (dative)), part of speech (*ADV*), obligatoriness, and a limited number of selection restrictions (human, inanimate, animal).

```
<vt>    monotransitive SUBJ V ACC
<vd>    monotransitive SUBJ V DAT
<vp>    monotransitive SUBJ V PIV
<va>    monotransitive SUBJ V ADV (TEMP, QUANT, LOC, DIR)
<vK>    copula SUBJ V SC
```

```
<vi>    intransitive inergative SUBJ V
<ve>    intransitive ergative V SUBJ
<vdt>   ditransitive SUBJ V ACC DAT
<vtp>   ditransitive SUBJ V ACC PIV
<vta>   ditransitive SUBJ V ACC ADV (LOC, DIR)
<vtK>   transitive prædicative SUBJ V ACC OC
<vU>    impersonal V
```

However, they do not explicitly refer to each argument in the frame and do not include semantic roles. Valency tags in *valency.cg3*, on the other hand, explicitly specify each argument with regard to its semantic role, morphological realization and lexical restrictions. Each argument constellation is specified by a separate valency tag, which is why a governor typically receives multiple valency tags to cover the full valency potential. For practical reasons (i.e. trying to keep valency tags as minimal as possible), subject roles are not included in the valency frames unless they make a significant difference in distinguishing between two meanings of a verb. This is the case in ex. (15-b), where the subject of *guoskat* 'concern' is obligatory and needs to be abstract, while the illative argument can either be a place (*<AG-Nom-Abs><TH-Ill-Plc>*) or abstract (*<AG-Nom-Abs><TH-Ill-Abs>*). In ex. (15-a), on the other hand, *guoskat* 'touch' has an animate subject (here implicit in the first person singular form of *guoskat*) and a concrete object (*<TH-Ill-Obj>*). By means of identifying the subject, the two meanings of *guoskat* 'touch' can be distinguished.

For testing purposes some argument descriptions referring to the subject are specified separately in *valency.cg3* as single tags disconnected from their frames, e.g. for a nominative AGENT (*<AG-Nom-Any>*) or a nominative EXPERIENCER (*<EX-Nom-Any>*). The subject can partly be dropped in the North Sámi language (in first/second person), cf. Svonni (2015, p.164). Covering all combinatory possibilities, including subjects, in valency tags doubles the amount of valency tags for each governor, which is why subjects are not included in the valency tags. However, they may be included in future versions of *valency.cg3*.

Facultative arguments are not marked with regard to their facultativity. Instead, one valency tag containing the argument and another one without the argument is specified. However, the ability of a governor to appear without any argument whatsoever is not codified in a valency tag, as both syntactic and pragmatic constraints allow almost any verb to appear without an argument in a specific context. For *gáđaštit* 'envy', three valency tags are listed as can be seen below. The first one includes both an animate THEME in locative case (*mus* 'me') and a REASON in accusative case (*dan* 'this'), cf. ex. (15-c) (Nielsen, 1926-1929, p.341). A second tag lists a REASON in accusative case only (*stáhtadoarjaga* 'subsidies'), cf. ex. (15-d). A third tag lists a THEME in accusative case (*bearašolbmuid* 'parents'), cf. ex. (15-e). Those three valency tags make up the valency potential of *gáđaštit* 'envy' in *valency.cg3*.

```
<TH-Loc-Ani><RS-Acc-*Ani>
<RS-Acc-*Ani>
<TH-Acc-Ani>
```

(15) a. . . . go dáhpedorpmis gusken buolli **gintalii**
. . . when unluckily touched burning candle.ILL
'. . . when unluckily I touched the burning candle'

b. **Dát** guoská maiddái **guođohanáigodahkii**.
this.NOM concerns also herding.season.ILL
'This also concerns the herding seasons.'

c. gáđaštii **mus dan**
envied I.LOC that.ACC
's/he envied me that' (Nielsen, 1932-1960*b*, p.12)

d. Ii gánnet gáđaštit Nuorttan[a]stte **stáhtadoarjaga**
not worth be.jealous.INF Nuorttanaste subsidy.ACC
'It is not worth it to be jealous of the Nourttanaste subsidies'

e. iige son gáđaš **bearašolbmuid**
not.either s/he be.jealous family.people.ACC.PL
'nor is s/he jealous of the parents'

A governor can receive several valency tags, each of which expresses a separate argument constellation with differences in either number of arguments, semantic roles, morphological tags or selection restrictions. This allows both for an unambiguous identification of a possible argument combination and word sense disambiguation. The valency tagset for each verb is meant to be exhaustive in the sense that all possible frames are given a tag and are matched to the verb in question. However, valency tags are added on the fly, and not all governors have been tagged according to their full valency potential yet.

Valency tags in *valency.cg3* consist of one or more arguments, cf. Table 3.7. *<TH-Acc-Any><RE-Loc-Ani>* is a tag that describes a frame with two arguments, the first of which is a THEME in accusative case of any semantic prototype and the latter of which is a RECIPIENT in locative case of an animate prototype. *<TH-ahte>* is a tag describing a frame with a THEME-argument realized by a subclause introduced by *ahte* 'that'. Selection restrictions are typically only relevant with regard to nouns, pronouns, adjectives and adverbs, and not to subclauses and non-finite verb specifications. *<Acc><TH-Inf>* is a valency tag describing accusative + infinitive constructions, where the semantic argument is an infinitive THEME. However, the infinitive has a subject in accusative case, which is also specified syntactically in the valency tag of the matrix verb as the governor requires an accusative + infinitive construction, not an infinitive argument only. *<TH-Acc-Any><árvvus>* is a valency tag that is added to a multi-word verb consisting of a verb and the locative singular form *árvvus* 'value' governing a THEME in accusative case. *<0>* is a valency tag that is added to an avalent verb, e.g. a weather verb or an impersonal passive verb.

Valency tag	***Governor* + example**
<TH-Acc-Any><RE-Loc-Ani>	e.g. *jearrat* sus dán 'ask her/him this'
<TH-ahte>	e.g. *jáhkkit* ahte 'believe that'
<Acc><TH-Inf>	e.g. *lohkat* su boahtit 'say that s/he comes'
<TH-Acc-Any><árvvus>	e.g. *atnit* sis árvvus 'value them'
<0>	e.g. *dánsojuvvui* 'there was dancing'

Table 3.7: Different types of valency tags in *valency.cg3*

3.2.2.1 Semantic role specifications in valency tags

Here, I present semantic role specifications within North Sámi valency tags and discuss the role of the following criteria in relation to semantic roles: universality, potential to account for syntactic alternations, uniqueness, minimalism (i.e. as few roles as possible), verb specificity, and independence of inherent semantics of arguments. Semantic role specifications in valency tags refer to a set of 24 semantic roles, cf. Table 3.8, which take Bick's (2007*c*) set of semantic roles (17 roles + 36 adjunct roles)[11] as a starting point, but also take into account Sammallahti's (2005) description (24 roles)[12] and Aldezabal's (2004) database for 100 Basque verbs (21 roles).[13] Their role sets differ not only in size, but also in their theoretical basis for constructing semantic roles, illustrated by the examples in Table 3.9. Both Bick's (2007*c*) and Aldezabal's (2004) role sets are used in machine readable grammars and corpus annotation. While Bick (2007*c*) uses the roles within a Danish, Spanish, and Portuguese Constraint grammar, Aldezabal (2004) has built a Basque database based on Levin's (1993) verb classes. Sammallahti's (2005) semantic roles, on the other hand, are part of a linguistic description of North Sámi. Unfortunately, Sammallahti's (2005) examples do not coincide with Bick's (2007*c*) and Aldezabal's (2004) examples, which is why the examples in Table 3.8 compare similar rather than equivalent constructions.

[11]'agent', 'causative agent', 'cognizer', 'speaker', 'patient', 'donor', 'recipient', 'beneficiary', 'experiencer', 'theme', topic domain, stimulus, result, 'message', 'state of affairs, fact', 'role', 'co-argument', 'static attribute', 'resulting attribute', 'source material', 'possessor', 'content', 'identity', 'location', 'origin, source', 'destination', 'path', 'social position', 'temporal location', 'temporal origin', 'temporal destination', 'temporal extension', 'frequency', 'extension, amount', 'cause', 'comparation', 'concession', 'condition', 'effect, consequence', 'purpose, intention', 'instrument', 'manner', 'accompanier', 'meta adverbial', 'dummy adverbial', 'reflexive', 'medial', 'vocative', 'focalizer', 'event, act, process', '(top) predicator', 'denomination', 'verb-incorporated'

[12]'agent', 'experiencer', 'patient', 'automaton', 'changer', 'mover', 'stative', 'theme', 'contents', 'consequence', 'result', 'owner', 'instrument', 'counterforce', 'benefactive','referent', 'place', 'source', 'path', 'goal', 'possessor', 'donor', 'conveyer', 'receiver' (Sammallahti, 2005, p.304)

[13]gaia 'theme', helburuko kokapena 'destination location', esperimentatzailea 'experiencer/agent', jarduera 'occupation', gai ukitua 'affected theme', helburuko egoera 'destination situation', kausa 'cause', neurria 'extent, amount', gai sortua 'created theme', abiapuntuko kokapena 'point of departure', iturria 'source', modua 'manner', egoera 'state/situation', bidea 'path', edukitzailea 'container', kokapena 'location', abiapuntua 'source (in exchange', edukia 'content, possessed item', denbora 'time', helburua 'goal', ezaugarria 'attribute'

Tag	Role	Example with a *governor* + argument bearing the respective role
TH	THEME	Maid don *liikot* **bargat** friddjabottuin? 'What do you *like* **to do** in your free time?'
AG	AGENT	Mánná gáskkáhalai **beatnagii**. 'The child *got bitten* by **the dog**.'
EX	EXPERIENCER	Eanaš bivdit ravgejit, eai ge *bállet* **guoli** albma láhkai darvánit. 'Most fishermen pull hard, they do not *give* **the fish** *time* to fasten properly.'
RE	RECIPIENT	Jeagil *addá* **bohccui** nu ollu álšša 'Lichen *gives* **the reindeer** so much energy'
DE	DESTINATION	dat *bistá* **juovllaide** 'it *lasts* until **Christmas**'
MA	MANNER	*mannat* **bures** '*go* **well**'
LO	LOCATION	Issát *bázii* **stohpui** smiehtadit. 'Issát *stayed* **in the house** to think.'
SO	SOURCE	mo *sirdit* máhtu **guovddážis** meahcásteaddjiide 'how to *move* knowledge **from the center** to the hunters'
PA	PATIENT	Muhto bohccot *borret* **muoraid** ja birgejit bures. 'But the reindeer *eat* **trees** and managed well.'
PR	PRODUCT	Tekstiilajoavku *lea gorron* **biktasiid** 'The textile group *has sewn* **clothes**'
PT	PATH	*vázzit* **dán geainnu** '*walk* **along this way**'
IN	INSTRUMENT	*hupmat* **telefuvnnas** '*talk* **on the phone**'
XT	EXTENT	*sirdit* dan **5 mehtera** '*move* it **5 metres**'
CO	CO-ARGUMENT	Sii *háleštit* **mánáiguin** das main sii beroštit 'They *speak* **with the children** about what they care about.'
PO	POSSESSOR	Stivralahtut *gullet* iešguđet **joavkkuide** 'The committee members *belong to* different **groups**'
PU	PURPOSE	*Manná* boazu várrái jassaguoraid **guohtut**. 'The reindeer *goes* to the mountain **to graze** next to the snow patches.'
RS	REASON/CAUSE	*jápmit* **nealgái** '*die* **of hunger**'
RO	ROLE	bargat **oahpaheaddjin** '*work* **as a teacher**'
BE	BENEFICIARY	*veahkehit* **su** '*help* **her/him**'
AT	ATTRIBUTE	Girji *lea* **logakeahttá**. 'The book *is* **unread**.'
RF	REFERENT	Spiinnit duođai *sulastahttiba* **du jiellahiid**! 'The pigs really *resemble* **your favorite children**!'
OR	ORIGIN	Dat gelbbolašvuohta maid oahppit *ožžot* **musihkkafágas** 'The competence that the students *get* **from the music subject**'
ID	IDENTITY	Oahpahusgiella *lea* **sámegiella** 'The teaching language *is* **Sámi**'
PV	PARTITIVE	Stuora *oassi* **eanandoalus** 'A big *part* **of agriculture**'

Table 3.8: The North Sámi semantic role set used in *valency.cg3*

Valency tags from *valency.cg3* are used in real word error detection, valency error detection, and lexical selection in machine translation. In general, all those tasks require a deeper syntactic analysis, which is why the semantic role set needs to be syntax-oriented, on the one hand, and semantically oriented, on the other hand. A semantic role set that is valid for any language, and based on semantic rather than syntactic criteria, is desirable for tasks such as machine translation. However, Helbig and Schenkel (1973, p.63) have already pointed out that obligatory arguments in one language can be free modifications or non-realizable in another language. While English verbs of displacement, e.g. *go*, can have an argument expressing EXTENT, the realization of EXTENT in Basque is ungrammatical, cf. ex. (16-a). In my initial valency experiments with lexical selection in Basque-North Sámi machine translation, cf. Wiechetek and Arriola (2011), I used valencies for lexical selection of verb translation equivalents. As verb sense disambiguation of the verb *sartu* '1.enter, 2.put' coincides with differences in their valency frames, they can be used to distinguish between translation equivalents. While the first sense has four matching arguments (AGENT, THEME, SOURCE, DESTINATION), the second sense has only three matching arguments (AGENT, SOURCE, DESTINATION). Syntactically, the argument constellations differ from each other in their respective languages. In ex. (16-b), *sartu* 'enter' coincides with the valency tag *<AG-Abs-Any><DE-Ine-Any>* and can be matched with *mannat (sisa)* 'go (in)' (*<AG-Nom-Any><DE-Ill-Any>*). In ex. (16-c), on the other hand, *sartu* 'put' coincides with the valency tag *<AG-Erg-Any><TH-Abs-Any><SO-Abl-Any><DE-Ala-Any>* and can be matched with *bidjat* 'put' (*<AG-Nom-Any><TH-Acc-Any><SO-Loc-Any><DE-Ill-Any>*). Taking into account valencies is also of clear advantage when translating cases in machine translation. While Basque inessive typically translates into North Sámi locative, the inessive DESTINATION-argument *tabernan* 'in the bar' should be translated by means of illative case in ex. (16-b). This information is encoded in the valency tag of *mannat (sisa)* 'go (in)'.

(16) a. ***lau metro** joan naiz (sukaldetik gelara)
four meters go AUX.1SG (kitchen.ABL room.ALL)
'I have gone four meters (from the kitchen to the bedroom)' (Estarrona et al., 2016, p.7)

b. **Mikel tabernan** sartu da.
Mikel.ABS bar.INE enter AUX.ABS3SG
'Mikel has entered the bar.' (p.k. Ainara Estarrona)

c. **Mikelek paperak poltsatik kaxoira** sartu ditu.
Mikel.ERG paper.ABS.PL bag.ABL box.ALL put AUX.ABS3PL.ERG3SG
'Mikel has taken the papers out of the bag and put them into the box (p.k. Ainara Estarrona)

Semantic roles are also used to account for different morpho-syntactic realization of the same argument type, especially in alternations that show a different perspective, but

contain the same arguments, cf. also Section 3.2.3.3. In Bick's (2007*c*) and Sammallahti's (2005) role systems (cf. Table 3.9), the subjects of intransitive motion verbs such as *walk, go*, or *come* are considered to have the same semantic role (AGENT) as the subjects of verbs such as *read*, *fetch*, etc. Vinka (2002, p.97) argues for the agentivity of the subject of these verbs (e.g. *viehkat* 'run') based on their availability to the causative alternation, i.e. the subject *Máhtte* in ex. (17-a), and hence also the object of the derived (causative) verb in ex. (17-b) is an AGENT. Vinka (2002, p.97) uses an independent agentivity test for volition, which consists of adding *mielastis* 'willingly, gladly' to the underived verb, as in ex. (17-a).

(17) a. **Máhtte** mielastis viegai
Máhtte willingly run.PRT.1SG
'Máhtte ran willingly' (Vinka, 2002, p.97)

b. Mon viegahin **Máhte**.
I run.CAUS.PRT.1SG Máhtte.ACC
'I caused Máhtte to run/ I chased Máhtte' (Vinka, 2002, p.97)

Sammallahti (2005)	**Aldezabal (2004)** (EADB)	**Bick (2007c)**	***valency.cg3***
		Movement	
$\mathbf{X}_{\text{AG}}$ vázzá $\mathbf{Y}_{\text{SO}}$ $\mathbf{Z}_{\text{DE}}$ 'X walks from X to Z'	$\mathbf{X}_{\text{TH}}$ joan $\mathbf{Y}_{\text{SO}}$ $\mathbf{Z}_{\text{DE}}$ da 'X went from Y to Z'	X_{AG} anda 'X walks'	*vázzit* 'walk' AG-SO-DE
biila$_{\text{MOVER}}$ vuolgá 'the car leaves'			
-	-	marchar **7km**$_{\text{XT}}$ 'walk 7km'	*vuolgit* 'leave' AG-XT
		Transitive movement	
Dulvi$_{\text{AUTOMAT}}$ doalvvui **stobu**$_{\text{MOVER}}$ 'The flood took the house'	$\mathbf{X}_{\text{CAUSE}}$ ekarri/eraman $\mathbf{Y}_{\text{TH}}$ $\mathbf{Z}_{\text{DE}}$ du 'X took Y to Z'	$\mathbf{X}_{\text{TH}}$ manda $\mathbf{Y}_{\text{SO}}$ **para** $\mathbf{Z}_{\text{DE}}$ 'X send Y to Z'	*doalvut* 'bring' AG-TH-SO-DE
Máret$_{\text{AG}}$ doalvvui **Máhte**$_{\text{PA}}$ **stohpui**$_{\text{DE}}$ 'Máret took Máhtte to the house.'			
		Subject roles	
$\mathbf{X}_{\text{EX}}$ massii $\mathbf{Y}_{\text{TH}}$ 'X lost Y'	$\mathbf{X}_{\text{EX/AG}}$ $\mathbf{Y}_{\text{TH}}$ ahaztu du 'X forgot Y'	$\mathbf{X}_{\text{AGCOG/EX}}$ esquece $\mathbf{Y}$ 'X forgets Y'	*vajálduhttit* 'forget', *massit* 'lose' AG-TH
$\mathbf{X}_{\text{AG}}$ geahčai $\mathbf{Y}_{\text{TH}}$ 'X looked at Y'	-	$\mathbf{X}_{\text{EX}}$ mira 'X looks'	*geahčča*t 'look' AG-TH
$\mathbf{X}_{\text{EX}}$ oinnii $\mathbf{Y}_{\text{TH}}$ 'X saw Y'	$\mathbf{X}_{\text{AG/EX}}$ ikusi $\mathbf{Y}_{\text{TH}}$ du 'X saw Y'	$\mathbf{X}_{\text{EX}}$ vê $\mathbf{Y}_{\text{TH}}$ 'X sees Y'	*oaidnit* 'see' EX-TH
		Causatives	
$\mathbf{X}_{\text{AG}}$ goaruhii $\mathbf{Y}_{\text{AG}}$ **gávtti**$_{\text{RESULT}}$ 'X made Y sew the costume'	-		*goaruhit* 'make sew, make eat' AG-AG-PR

$\mathbf{X}_{\text{AG}}$ viegahii $\mathbf{Y}_{\text{PA}}$ 'X persecuted Y/made Y run'	-	$\mathbf{X}_{\text{AGCAUS}}$ fez desaparecer $\mathbf{Y}_{\text{AG}}$ 'X made Y disappear'	*viegahit* 'make run, persecute' AG-AG
Symmetric verbs			
-	$\mathbf{X}_{\text{TH}}$ $\mathbf{Y}_{\text{TH}}$ aldatu du 'exchanged X with Y'	-	*molsut* 'exchange (with)' TH-TH
-	$\mathbf{X}_{\text{EX/AG}}$ $\mathbf{Y}_{\text{EX/AG}}$ besarkatu du 'X hugged Y'	$\mathbf{X}_{\text{AG}}$ abraça $\mathbf{Y}_{\text{PA}}$ 'X hugs Y'	*salastit* 'hug' AG-CO
Object roles			
$\mathbf{X}_{\text{AG}}$ lávllui **lávlaga**$_{\text{CONSEQUENCE}}$ 'X sang a song'	$\mathbf{X}_{\text{EX/AG}}$ abestu $\mathbf{Y}_{\text{TH}}$ du 'X sang Y'	$\mathbf{X}_{\text{AG}}$ canta $\mathbf{Y}_{\text{TH}}$ 'X sings Y'	*lávlut* 'sing' AG-TH
$\mathbf{X}_{\text{AG}}$ duddjui **guvssi**$_{\text{PR}}$ 'X made a cup'	-	$\mathbf{X}_{\text{AG}}$ produz $\mathbf{Y}_{\text{RS}}$ 'X produces Y'	*duddjot* 'make' AG-PR
Beneficiary, recipient, etc.			
-	$\mathbf{X}_{\text{SO}}$ $\mathbf{Y}_{\text{DE}}$ deitu du 'X called Y'	**Ela**$_{\text{AG}}$ **lhe**$_{\text{BE}}$ chamou por telefono. 'She calls him on the phone'	*riŋget* 'call' AG-RE
X attii $\mathbf{Y}_{\text{RE}}$ ruđa 'X gave Y money'	$\mathbf{X}_{\text{SO}}$ $\mathbf{Y}_{\text{TH}}$ $\mathbf{Z}_{\text{DE}}$ eman du 'X gave Y to Z'	$\mathbf{X}_{\text{AG}}$ dar $\mathbf{Y}_{\text{TH}}$ **a** $\mathbf{Z}_{\text{BE}}$ 'X give Y to Z'	*addit* 'give' AG-TH-RE
X muitalii $\mathbf{Y}_{\text{BE}}$ **Z** 'X told Y Z'	$\mathbf{X}_{\text{EX/AG}}$ $\mathbf{Y}_{\text{TH}}$ **Z** esan du 'X said Y to Z'	$\mathbf{X}_{\text{SPEAKER}}$ diz $\mathbf{Y}_{\text{MESSAGE}}$ **a** $\mathbf{Z}_{\text{RE}}$ 'X says Y to Z'	*muitalit* 'tell' AG-TH-RE
X rabai $\mathbf{Y}_{\text{BE}}$ uvssa 'X opened the door for Y'		ajuda **a** $\mathbf{Y}_{\text{BE}}$ 'help Y'	*rahpat* 'open' AG-TH-BE
***leat* 'be'**			
-	$\mathbf{X}_{\text{TH}}$ $\mathbf{Y}_{\text{AT}}$ da 'X is Y'	-	$\mathbf{X}_{\text{TH}}$ lea **čeahppi**$_{\text{AT}}$ 'X is smart'
sus$_{\text{PO}}$ lea **biila**$_{\text{CONTENT}}$'s/he has a car'	$\mathbf{X}_{\text{PO}}$ $\mathbf{Y}_{\text{TH}}$ du 'X has Y'	$\mathbf{X}_{\text{PO}}$ possui $\mathbf{Y}_{\text{TH}}$ 'X has Y'	*leat* 'have' PO-TH
sus$_{\text{EX}}$ lea čottadávda. 's/he has a sore throat'	-	$\mathbf{X}_{\text{TH}}$ está **doente**$_{\text{AT}}$ 'X is sick'	*leat* 'have' PO-TH
Others			
$\mathbf{X}_{\text{CONTENT}}$ sulastahttá $\mathbf{Y}_{\text{RF}}$ 'X resembles Y'	-	-	*sulastahttit* 'resemble' TH-RF
-	-	**X** trabalha **como guía**$_{\text{RO}}$ 'X works as a guide'	*bargat* 'work (as)' AG-RO

Table 3.9: Semantic role annotation in Sammallahti (2005), Aldezabal (2004) and Bick (2007c) and *valency.cg3*

However, in Aldezabal's (2004) role set for Basque, these verbs have a THEME-subject accounting for the ergative-absolutive alternations of these verbs. The subject of *sartu* ('1.enter, 2.put') is an AFFECTED THEME in the intransitive variant as is the object of

the transitive variant. The ergative, on the other hand, is a CAUSE, cf. Aldezabal (2004, p.188) and Estarrona et al. (2016, p.4). As this research prioritizes syntactic (monolingual) tasks (grammar checking, semantic role annotation) over semantic (bilingual) tasks such as machine translation, I will focus on syntactic regularities within North Sámi, rather than taking into account those types of alternations in other languages such as Basque. Subjects of motion verbs are therefore annotated as AGENTS in causative constructions in *valency.cg3*.

While valency annotation should account for syntactic alternations of the same semantic roles, it can also be important to distinguish between the semantic implications of morpho-syntactic differences. Sammallahti (2005, pp.60–71) distinguishes between role alternations in different types of passives. While Sammallahti (2005, p.65) classifies the object of the active verb *oaidnit* 'see' as a THEME in ex. (18-a), he considers it an EXPERIENCER in the "adversative passive" of ex. (18-b). In the "intentional passive" of ex. (38-d), he classifies it as a THEME, and in the "automative passive" of ex. (38-e) he considers it a CONTENT (as it does not correspond to the original active with an AGENT and therefore is not controlled). While some constructions are productive, others have a more or less lexicalized meaning. In *valency.cg3*, all three subjects of passive verbs and the object of the active construction are considered THEMES. Preserving semantic roles in diathesis alternations serves the purpose of accounting for missing arguments during grammar checking.

(18) a. Máhtte **oinnii** Máreha.
Máhtte see.PRT.3SG Máret.ACC
'Máhtte saw Máret.'

b. Máret **oainnáhalai** Máhttii.
Máret see.PASS.PRT.3SG Máhtte.ILL
'Máret was seen by Máhtte.' (Sammallahti, 2005, p.62)

c. Máhtte **oidnojuvvui**.
Máhtte see.PASS.PRT.3SG
'Máhtte was seen.' (Sammallahti, 2005, p.61)

d. Máhtte **oidnui**.
Máhtte see.PASS.PRT.3SG
'Máhtte was visible.' (Sammallahti, 2005, p.61)

The arguments' uniqueness is useful when accounting for missing arguments in grammatical error detection, which is why arguments should generally occur only once with respect to a single governor. There are cases where more than one occurrence of the same role is conceptually meaningful, i.e. in coordination, causative and symmetric constructions. Sammallahti (2005, p. 75, 78) annotates two AGENTS to certain causative constructions. Bick (2007*c*), however, distinguishes formally between an AGENT (*AG*) and a CAUSATIVE AGENT (*AGcaus*). In *valency.cg3*, I do not distinguish between AGENTS

and CAUSATIVE AGENTS, as CAUSATIVE AGENTS can be distinguished from non-causative AGENTS by means of a morphological tag and/or the morphological case of the AGENT. More than one role of the same kind can also appear with verbs expressing a certain symmetric relation of two arguments. However, views differ as to what a symmetric relation is. Aldezabal (2004, p.278,293) assigns two THEME-arguments to the verbs *aldatu* 'distinguish (sth. from sth.)' and *konparatu* 'compare (sth. with sth.)', formally distinguishing between THEME1 and THEME2 or THEME and CO-THEME. Sammallahti (2005, p.78), on the other hand, assigns different roles to *Máreha* (THEME) and *Ánnes* (REFERENT) in ex. (19-a). In *valency.cg3*, I distinguish between a THEME and a REFERENT in ex. (19-a). In the reciprocal construction in ex. (19-c), I distinguish between AGENT and CO-ARGUMENT of the verb *hállat* 'talk'. In certain passive constructions as in ex. (19-b), Sammallahti (2005) also sees two EXPERIENCERS. In *valency.cg3*, I distinguish between a THEME (*Máret*) and an EXPERIENCER (*Máhttii*).

(19) a. Máhtte ii earuhan *Máreha* **Ánnes**.
Máhtte not distinguish Máret.ACC Ánne.LOC
'Máhtte didn't distinguish Máret from Ánne.' (Sammallahti, 2005, p.99)

b. *Máret* oainnahalai **Máhttii**.
Máret see.PASS.PRT.3SG Máhtte.ILL
'Máret was seen by Máhtte.' (Sammallahti, 2005, p.62)

c. Erla čilge iežas hállat **háldiiguin**
Erla explains herself speak underground.beings.COM.PL
'Erla explains that she speaks with the underground beings'

To unambiguously identify the arguments, role distinctions should be made if certain types of arguments co-occur with the same governor. This holds, for example, for the distinction between a PATIENT (*áiggi* 'time (Gen./Acc.)') and a PRODUCT (*oassái* 'part (Ill.)') in ex. (20-a) or in ex. (20-b). At the same time, a semantic role accounts for mutually exclusive morpho-syntactic realizations of the same argument, e.g. when the same argument can be realized as a noun phrase, an adpositional phrase, a subclause, etc. Lastly, a minimal role set is useful to minimize semantic role annotation rules and semantic role specifications in error detection rules. This is why I do not distinguish between arguments that express e.g. permanent and non-permanent changes in *valency.cg3* as Nickel and Sammallahti (2011) do. Nickel and Sammallahti (2011) distinguish between the role for *lávlla* 'song' in ex. (20-c), i.e. CONSEQUENCE (a non-permanent product), and *gákti* 'costume' in ex. (20-d), i.e. (permanent) PRODUCT. Nor do I distinguish between sub-roles within different domains as Bick (2007*c*) does. He distinguishes between a THEME and a TOPIC, the latter of which is a THEME in the domain of a cognitive or communicative action or activity. He also distinguishes between a RESULT and a MESSAGE, the sub-RESULT role for communicative actions, i.e. the object of verbs of saying, confirming, and justifying. The corresponding AGENT subcategory is a SPEAKER.

(20) a. ...ferte juohkit **áiggi** soađi maŋŋel guovtti *oassái*.
...have.to split time.ACC war after two part.ILL
'...has to split the time after the war into two parts.'

b. ...riikka ovddasvástádus njulget **áššI** ovttaskasa *buorrin*.
...country's responsibility straighten.out thing individual good.ESS
'...the country's responsibility to straighten out the thing for the individual good.'

c. Máhtte lávlu **lávlaga**.
Máhtte sings song.ACC
'Máhtte sings a song.'

d. **Gákti** gorrojuvvui.
costume.NOM sew.PASS.PRT.3SG
'The costume was sewn.'

Role distinctions are also important when distinguishing between two verbs or even verb groups, e.g. between certain confusion pair members for real word errors such as different forms of *áddet* 'understand' and *addit* 'give'. A confusion pair consists of two (or more) real word forms that are likely to be confused in writing. The MANNER-argument *bures* 'well', as shown in ex. (21-a) distinguishes the verb *áddet* 'understand' from the verb *addit* 'give', which it is often confused with in spelling. Therefore a MANNER-argument is considered part of the valency of *áddet* 'understand', but not of the valency of *addit* 'give'. A TIME-adverbial like *guhká* 'long', which appears in ex. (21-b), on the other hand, is implied in the meaning of a verb like *ádjánit* 'last', and therefore considered part of its valency. However, it is not implied in the meaning of the verb *lohkat* 'read', and therefore not considered part of its valency. Valency specifications help to identify the specific verb in the respective argument constellation.

(21) a. addit *bures vs. áddet **bures**
give well vs. understand well
'give well vs. understand well'

b. lohkat guhká vs. ádjánit **guhká**
read long vs. take long
'read for a long time vs. take a long time'

Since selection restrictions are referred to separately within valency tags in *valency.cg3*, semantic roles should be independent of the semantic features of their arguments. Animacy/humanness is naturally an important and grammaticalized feature in many human languages. Both Bick (2007*c*) and Sammallahti (2005) have a default human/animate AGENT. But while Bick's (2007*c*) semantic role with respect to a governor does not depend on the actual animacy of the argument, Sammallahti (2005, p.41) categorically distinguishes between animate and inanimate nouns in specific positions. He mentions only one exception to the animate EXPERIENCER-subject of *jugahallat* 'be drinkable', i.e. the inanimate noun *viinnit* 'wine (Nom. Pl.)', shown in ex. (22).

(22) **Viinnit** dat gal jugahalle.
wine.NOM.PL that definitely drink.PASS.PRT.3PL
'The wines were definitely drinkable.' (Sammallahti, 2005, p.62)

Sammallahti (2005) and Nickel and Sammallahti (2011, p.368) distinguish between an AGENT and an AUTOMATON, the latter of which is used only in inanimate examples. In ex. (23-a), *Máret* is classified as an AGENT. In ex. (23-b), on the other hand, *dulvi* 'flood' is considered an AUTOMATON even though the action is physical in both cases and the meaning of the verb itself does not change. Bick (2007*c*), in contrast, takes into consideration the metaphoric use of verbs without changing the argument structure, so that, for example, a text can be an AGENT if used with a verb that typically occurs with a human subject. 'Control' is defined by both Bick (2007*c*) and Sammallahti (2005). But while Bick (2007*c*) sees a strong physical component in the definition of 'control', for Sammallahti (2005), only animates can control an action. Also, in cases of movement with a vehicle, as in ex. (23-d), the subject is not considered an AGENT, but a MOVER by Sammallahti (2005). The semantic role of an argument in Sammallahti's (2005) and Nickel and Sammallahti's (2011) systems depends not only on its inherent semantic features, but also on the inherent semantic features of the other arguments of the frame. In other words, the role of *stobu* 'house' (i.e. MOVER) in ex. (23-b) is distinguished from the role of *Máhte* (Acc.), i.e. PATIENT, in ex. (23-c). This is purely based on the semantic role of the subject, i.e. AUTOMAT in the first case and AGENT in the second case.

I do not make these types of distinctions in *valency.cg3* as it would mean doubling the amount of valency tags not only with respect to animate vs. non-animate subjects, but also with respect to different subject-object constellations. This is not beneficial for a minimal tag inventory. I subscribe to Bick's (2007*c*) view on metaphorical extensions of valency frames with e.g. prototypically animate subjects, and also consider inanimate arguments AGENTS in those frames.

(23) a. **Máret** vázzá
Máret walks
'Máret walks' (Nickel and Sammallahti, 2011, p.368)

b. **Dulvi** doalvvui *stobu*.
flood took house.ACC
'The flood took the house.' (Nickel and Sammallahti, 2011, p.368)

c. Máret doalvvui **Máhte** stohpui.
Máret took Máhtte.ACC house.ILL
'Máret took Máhtte to the house.'

d. **Máret** ollii Mázii/ **Máhtte** manná fatnasa fárus.
Máret reached Máze.ILL/ Máhtte goes boat.GEN by
'Máret reached Máze/ Máhtte goes by boat.'

While Bick (2007*c*) and Aldezabal (2004) distinguish between spatial and temporal roles

with regard to DESTINATIONS, SOURCES and LOCATIONS, in *valency.cg3* reference to time or place is made in the selection restrictions. However, the roles remain general, as can be seen in the valency tags for SOURCE-arguments *<SO-Loc-Time>* vs. *<SO-Loc-Plc>*, which differ only in their reference to selection restrictions.

3.2.2.2 Morpho-syntactic specifications in valency tags

While semantic role specifications make up the first part of the argument description of a valency tag, the second part of a valency tag typically refers to morpho-syntax, e.g. illative case. It can also refer to a set generalizing over several morphological tags, a lemma, a particular word form in the case of idiomatic constructions or to a clause, e.g. a finite or non-finite clause.

3.2.2.2.1 Morphological constraints

Table 3.10 gives an overview of valency tags that refer to morphological case, lemmata, non-finite verb-forms, parts of speech, and word forms. Morphological constraints typically refer to specific cases (nominative, illative, accusative, essive, etc.) or lemmata of adpositions. In addition, some tags refer to a specific verb form, i.e. infinitive in the case of single infinitive arguments, or to a part of speech in the case of an adverb (*Adv*), e.g. *<MA-Adv-Manner>*. This tag describes the valency realized in ex. (24-a), where a manner adverb, i.e. *bures* 'well', has the MANNER-role.

Postpositional phrases are often alternative to case realizations of the same semantic role. The verb *suhttat* 'get angry', for example, can be used with a THEME expressed by a postpositional phrase with *ala* 'at' as in ex. (24-b). Here *ala* 'at' cannot be replaced with other postpositions denoting direction (e.g. *vuollái* 'under', *lusa* 'to', *maŋŋái* 'after'), which is why the postposition is explicitly stated in the valency tag *TH-ala-*Plc*. In some cases, both a case (e.g. essive case) and a part of speech (e.g. adjective) as in *<MA-Ess-Adj>* are referred to in a valency tag, cf. ex. (24-c). Verbal arguments, e.g. infinitives or actio essive (i.e. progressive) forms, are specified by morphological constraints only (i.e. no selection restrictions) if no other elements are required.

(24) a. manná **bures**
go.PRS.3SG well
'it goes well'

b. Guovssahasat suhtte nuorat viellja **ala**
Northern.lights got.angry younger brother.GEN on
'The Northern lights got angry at the younger brother'

c. Viesu siste lea buot **čáhppadin** gožuduvv[a]n.
house.GEN inside is everything black.ESS cover.in.ash
'Inside the house, everything is covered in black ash.'

Valency tag	**Example with the governor in question**
	Morphological case
<DE-Ill-Time>	dat **bistá** juovllaide 'it **lasts** until Christmas'
<SO-Loc-Lang><DE-Ill-Lang>	go gártá **jorgalit** luondduálbmoga gielas omd. dárogillii. 'when one will **translate** from an indigenous language to e.g. Norwegian'
<TH-Ill-Any>	Boađát **áibbašit** Lucia-feasttaide. 'You are going **to long for** the Lucia parties.'
	Postpositions
<TH-gaskkas-Any>	**oktavuohta** sápmelaččaid gaskkas '**connection** between Sámi people'
<TH-ala-*Plc>	Mon **luohtán** du ala. 'I **trust** you.'
<TH-badjel-Ani>	geahčča**t** iežaset lunttaid **vuoitit** badjel joavku 'watch one's own boys **win** over the group'
<TH-birra-Any>	Mii fertet duostat **hállat** dán birra 'We have to dare to **talk** about this'
<TH-ovddas-Any>	Dássážii in leat gullan ovttage **ákkastallamin** dásseárvvu ovddas. 'Until now, I have not heard anyone **arguing** in favor of gender equality.'
<LO-maŋŋil-Time>	Dáhkidus **loahpahuvvo** maŋŋil 30 jagi 'The insurance **ends** after 30 years.'
	Non-finite verb forms
<BE-Acc-Ani><TH-Inf>	**rávvejedje** olbmuid jurddašit aivve buriid jurdagiid '**they advised** the people to think only good thoughts'
<PU-AktioEss>	Son **fitná** poasttas páhka viežžamin. 'S/he **takes a trip** to the post office to fetch the parcel.'
	Part of speech
<LO-Adv-Time>	Man guhká sáhttá **vuordit**? 'How long can s/he **wait**?'
<MA-Adv-Manner>	**Mana** dearvan! 'Goodbye! (lit. go healthy)'
<MA-Ess-Adj>	Viesu siste lhea buot čáhppadin **gožuduvv[a]n**. 'Inside the house, everything is covered in black ash.'
	Word forms
<PA-Acc-ieš><LO-Ill-Any>	oahppit **čiekŋudit** iežaset fáttáide 'the students **immerse** themselves into the subjects'

Table 3.10: Different morphological specifications in the valency tags of *valency.cg3*

Valency	Example
Subclause without subjunction	
<TH-FS-Qpron>	*čilget* **makkár** dillái muhtumat gártet '*explain* **what kind of** situation some people are going into'
<TH-FS-Qst>	Ii dárbbašan *ballat* **máhttágo** juoigat 'S/he does not need to *fear* that s/he **cannot** yoik'
<TH-FS>	*dadjat* guohtumis **lea** buorre kvalitehta. '*say* the pastures **have** good quality'
Subclause with subjunction	
<TH-ahte>	Muhto áhčči fas *gáibidii* **ahte** mus lea sámegiel namma. 'But my father *demanded* **that** I have a Sámi name.'
<TH-go>	Earát *liikojit* **go** leat eambbo guldaleaddjit. 'Others *like* **when** there are more listeners.'
<TH-jus>	váhnemat eai **liiko** *jus* vilgessáhpaniiguin fal bearehaga ovttastallat 'the parents don't **like** it *if* we socialize too much with the white mice'

Table 3.11: Valency tags for different types of finite subclauses in *valency.cg3*

3.2.2.2.2 Syntactic constraints

Apart from morphological constraints or part of speech specifications, syntactic constraints are also referred to in the valency tags, e.g. in the case of finite or non-finite clauses. Syntactic constraints refer to the clause's head, i.e. the finite or non-finite verb, a particular subjunction, e.g. *ahte* 'that', *go* 'when', and *jus* 'if', cf. Table 3.11. Alternatively, they refer to several obligatory parts of the clause, cf. Table 3.12. Finite subclause arguments are referred to by both a semantic role and a syntactic specification, but naturally lack a reference to selection restrictions. Different types of finite subclauses are distinguished by references to the characteristic subjunction or interrogative pronoun/adverb introducing the subclause, e.g. *ahte* 'that', *go* 'when', *gii* 'who', *mii* 'what', and *manin* 'why'. In ex. (25-a), the interrogative adverb *goas* 'when' introduces the finite subclause (*FS*) argument of *vuorddašan* 'I wait', which is why *vuorddašit* 'wait' receives the valency tag *<TH-FS-Qpron>*. If the subclause is introduced neither by a subjunction nor by a question pronoun/adverb, the form of the finite verb is either left unspecified (*<TH-FS>*) or it is specified in terms of its question particle (*<TH-FS-Qst>*). In ex. (25-b), the argument of *mearridit* 'decide' is a finite clause headed by a finite verb with a question particle *addet go* 'do you understand', which is why *mearridit* 'decide' receives the valency tag *<TH-FS-Qst>*.

(25) a. Dás de čohkkan, vuorddašan **goas** soitet oahput álggahuvvot
here then sit, waiting when might teaching.PL begin.INF
'Here I sit then, wondering when the class might begin'

b. ... galgá Norgga ráđđehus mearridit **addet go** stádadáhkádusa
... will Norwegian government decide give.PRS.3PL Q state.insurance

Romsa Olympiijagilvvuid lágideapmái
Tromsø Olympic.Game.GEN.PL committee.ILL
'... the Norwegian government will decide whether they will give state insurance to the Tromsø Olympic Games committee'

Valency tags for non-finite clauses specify a non-finite argument and an accusative argument, which is the subject of the non-finite governor and at the same time the object of the matrix verb, cf. Table 3.12. In contrast to infinitival arguments with the valency tags *<TH-Inf>*, *<PU-Inf>*, etc., these clauses typically require another argument in addition to the non-finite form to form a grammatical sentence. While the verb *liikot* 'like' can have a simple infinitival argument, it cannot be the governor of an accusative + infinitive construction like the verbs *doaivut* 'hope', *jáhkkit* 'believe', and *ballat* 'fear', cf. ex. (26-a). Both accusative and infinitive are therefore specified in the valency tag of the respective governor. Some verbs like *doaivut* 'hope' can appear in both types of constructions, i.e. with only an infinitive, e.g. *vásihit* 'experience' in ex. (26-b), but also with an accusative (*olbmuid* 'people') and an infinitive (*geavahit* 'use') in ex. (26-c).

(26) a. Mun doaivvun/jáhkán/balan/*liikon *su* **boahtit**.
I hope/think/fear/like s/he.ACC come.INF
'I hope/think/fear/like s/he will come.' [p.k. H]

b. Mun doaivvun **vásihit** seammá boahttevašvuođas maid ...
I hope experience.INF same future.LOC too ...
'I hope to experience the same in the future too ...'

c. Ja mii doaivut *olbmuid* **geavahit** vejolašvuođa deaivvadit singuin.
and we hope people.ACC use.INF chance meet them.COM.PL
'And we hope that the people use the chance to meet with them.'

Magga (1986, p.179) distinguishes between different types of accusative and infinitive constructions based on the role of the accusative argument, the infinitive and the subject of the matrix-verb's subject. "Object-clauses" such as the one in ex. (27-a) can be replaced with the accusative pronoun *dán* 'this' or a subclause sentence with *ahte* 'that'. Accusative and infinitive clauses, on the other hand, are clauses, in which the matrix verb governs the accusative semantically. Magga (1986) does not use an elaborate semantic role set, and uses the terms AGENT and PATIENT predominantly syntactically, more like *subject* and *object*. However, he notes that a more detailed semantic categorization is possible, cf. Magga (1986, p.190). Like Magga (1986), I distinguish between object-clauses and accusative + infinitive clauses, where both accusative and infinitive can be interpreted as two single arguments of the matrix verb, cf. Table 3.12. Magga (1986, p.218) addresses the difficulty in deciding on the status of the infinitive and accusative as either one or two arguments of the matrix verb. In *valency.cg3*, object-clauses such as those in ex.

(26-a) and ex. (27-a) are considered to have only one semantic role (THEME) for the infinitive. They are prototypically governed by *verba sentiendi/declarendi* such as *lohkat* 'claim'. According to Magga (1986, p.176), the event described in the matrix verb does not have any influence on the accusative argument in the construction. The accusative is governed semantically by the infinitive and therefore has only a semantic role with respect to its infinitival governor. Syntactically, however, the accusative is also governed by the matrix verb, which is why it receives the valency tag *<Acc><TH-Inf>* in *valency.cg3*.[14] Accusative-infinitive clauses, on the other hand, consist of two independent arguments, one of which is a THEME, i.e. the infinitive. The accusative has a semantic role with respect to the matrix verb, even though there is also a clear semantic relation between the infinitive and the accusative. In *valency.cg3*, I distinguish between accusatives that are RECIPIENTS, EXPERIENCERS, BENEFICIARIES, and PATIENTS. Communicative verbs with a RECIPIENT-role , e.g. *ávžžuhit* 'prompt', *čuorvut* 'call', *átnut* 'plead', or *gohčču t* 'order', receive the tag *<RE-Acc-Ani><TH-Inf>*. For those verbs, typically the construction in ex. (27-b) is synonymous to the one in ex. (27-c). The illative *sutnje* 's/he (Ill.)' in ex. (27-d), on the other hand, does not have the same role as the accusative *du* 'you (Acc.)' in ex. (27-e). While *sutnje* 's/he (Ill.)' is a RECIPIENT, *du* 'you (Acc.)' is an AGENT with respect to *vuolgit* 'leave'. Therefore, *lohpidit* 'promise' receives the same valency tag as *lohkat* 'say' in ex. (27-a). For verbs with the valency tag *<RE-Acc-Ani><TH-Inf>*, on the other hand, the construction in ex. (27-c) is synonymous to the one in ex. (27-b).

(27) a. Son lohká **Deanu** *leat* issoras guhkes čázádat ...
s/he claims Deatnu.ACC be.INF extremely long body.of.water ...
'S/he claims that Deatnu is an extremely long river ...'

b. son ávžžuhii/gohčui **su** *boahtit*
s/he prompted/called s/he.ACC come.INF
's/he prompted/called him/her to come'

c. son ávžžuhii/gohčui **sutnje**: *boađe*!
s/he prompted/ordered s/he.ILL: come
's/he prompted/ordered him/her: come!'

d. Ledje lohpidan **sutnje** *boahtit* ruoktot, ...
had promised s/he.ILL come.INF home, ...
'They had promised him/her to come home, ...'

e. Mun lohpidan **du** *vuolgit*.
I promise you.ACC leave.INF
'I promise that you can leave.'

Verbs like *veahkehit* 'help', *neavvut* 'advise', *oahpistit* 'advise', and *rávvet* 'advise' have accusative BENEFICIARY arguments which typically alternate between a frame with an argument in accusative case and a frame with an accusative and infinitive, cf. ex. (28-a).

[14] In Chapter 5, the tag is referred to as *<TH-Acc-Any><TH-Inf>*, which is kept in the old form so that the reader can recover the earlier version of the rule file and reproduce the results.

They receive the valency tag *<BE-Acc-Ani><TH-Inf>*. The verb *rávvet* 'advise' appears with a BENEFICIARY in accusative case and an illative THEME, cf. ex. (28-b), or with an illative BENEFICIARY and an accusative THEME, cf. ex. (28-c) (Nielsen, 1926-1929, p.263).

(28) a. ... veahkehit **guollebivdiid háhkat** áhpebivdui heivvolaš fatnasiid.
... help fishermen.ACC get.INF ocean.fishing.ILL suitable boats
'... help the fishermen to get suitable boats for ocean fishing.'

b. rávvii **min** *gulolašvuhtii*
advise we.ACC adherence.ILL
'advise us to adhere'

c. *maid* áiggut don dalle **munnje** rávvet?
what.ACC want you then me.ILL advise
'how do you want to advise me?'

Verbs like *balddihit* 'scare' receive the valency tag *<EX-Acc-Ani><TH-Inf>* because of parallel constructions with an accusative EXPERIENCER (*olbmuid* 'people (Acc. Pl.)' and a THEME realized as a subclause (*ahte deanoluossanálli lea uhkiduvvon* 'that the Deatnu-salmon is threatened') as in ex. (29-a). Verbs like *bidjat* 'put; get to do sth.', on the other hand, can have a causative meaning suggesting an additional AGENT. However, the role depends on the infinitive verb, which is why the accusative arguments are considered PATIENTS with respect to the matrix verb, e.g. *olbmuid* 'people' in ex. (29-b).

(29) a. ... balddihan *olbmuid* **ahte** deanoluossanálli lea uhkiduvvon
... scared people.ACC that Deatnu.salmon is threatened
'... s/he made the people scared that the Deatnu-salmon might be threatened'

b. bidjat *olbmuid* **jurddašit** das mii rasisma lea
get people.ACC think.INF it.LOC what racism is
'get people to think about what racism is'

Accusative and infinitive constructions with verbs like *njoarrat* 'pour', shown in ex. (30-a), are not annotated as frames with two arguments in *valency.cg3*. Verbs of this type typically alternate between constructions as the one shown in ex. (30-a) and the one in ex. (30-b).

(30) a. Mon njorren *gáfe* **čoaskut**
I poured coffee.ACC cool.down.INF
'I poured the coffee to cool it down' (Magga, 1986, p.202)

b. Mon njorren gáfe vai (gáffe) **čoasku**.
I poured coffee.ACC so (coffee.NOM) cool.down.PRS.3SG
'I poured the coffee to cool it down.' (Magga, 1986, p.202)

Valency	Governor examples
<Acc><TH-Inf>	*lohkat* 'claim', *diehtit* 'know', *jáhkkit* 'believe'
<RE-Acc-Ani><TH-Inf>	*ávžžuhit* 'prompt', *čuorvut* 'call', *gohčču̇t* 'order', *jearrat* 'ask', *sártnuhit* 'persuade'
<BE-Acc-Ani><TH-Inf>	*álggahit* 'help', *veahkehit* 'help', *neavvut* 'advise', *rávvet* 'advise', *oahpistit* 'guide'
<EX-Acc-Ani><TH-Inf>	*oalgguhit* 'encourage', *balddihit* 'scare', *árvvosmuhttit* 'encourage', *bođđet* 'incite'
<PA-Acc-Ani><TH-Inf>	*bágget* 'force', *addit* 'make do sth.', *dájuhit* 'get sb. to do sth. wrong'

Table 3.12: Valency tags for accusative + infinitive constructions in *valency.cg3*

3.2.2.3 Selection restrictions in valency tags

Valency tags in *valency.cg3* refer further to semantic selection restrictions. The selection restriction typically refers to a semantic prototype, positively or negatively, or to a lemma in the case of an idiomatic construction, cf. Table 3.13. Selection restrictions are only specified for those parts of speech that are annotated with regard to a semantic prototype in the respective *lexc* lexicon, i.e. nouns, pronouns, adjectives, adverbs, and adpositions. Verbs, finite subclauses and infinitival constructions, on the other hand, are not specified with regard to their selection restrictions. Selection restrictions in *valency.cg3* refer to measure, money, time, frequency, vehicle, language, body, animates, human, to name but a few.

Selection restrictions typically do not influence the grammaticality of a sentence in the same way as morpho-syntactic constraints, i.e. violations of selection restrictions can be made deliberately to create a specific meaning, and may depend on specific domains of texts (e.g. communication verbs with inanimate subjects can be acceptable in fiction, etc.).

Below I illustrate a number of cases where selection restrictions serve various practical purposes. In some cases they can be used for verb sense disambiguation of polysemous verbs where the arguments do not differ morpho-syntactically. The verb *addit* 'give' with a human accusative and an infinitive as in ex. (31-a) means 'get sb. to do sth.' (Magga, 1986, p.194). In ex. (31-b), *addit* 'give' appears with a non-human accusative and an infinitive, meaning 'give'. The infinitive is not part of the valency here. Verb sense disambiguation and also valency disambiguation can be achieved by identifying the semantic prototype of the accusative.

Selection restrictions are also used to distinguish between different semantic domains of the same roles, e.g. in the case of verbs that ask for a SOURCE and a DESTINATION. While in ex. (31-c) the valency of *jorgalit* 'translate' requires a SOURCE-argument of the language-prototype category and a DESTINATION-argument, in ex. (31-d), the valency of

Tag	Verb	Example with governor and argument
Animacy		
<CO-mielde-Ani>	vuolgit	Vuolgi **mu mielde**! 'Come **with me**!'
<OR-Loc-HumGroup>	leat	Bárdni lea riggámus **sogas**. 'The boy is from the richest **family**.'
Concrete		
<IN-Com-Veh>	vuodjit	vuodjit **skuteriin** muhtin joga badjel. 'drive **with the scooter** over some river.'
<PT-rastá-Plc>	mannat	manai dušše rastá **luotta** 'she just walked across **the path**'
<PA-Acc-Food>	jugistit	de jugistii **sávtta**. 'then s/he drank a little **juice**.'
Abstract		
<AG-Nom-Abs><TH-Ill-Abs>	guoskat	**Dat** guoská **ráhkisvuhtii**. 'It concerns **love**.'
<TH-Ess-Wthr>	birget	Gal golmmaiguin fáhcaiguin birget **buolašin**. 'With three pairs of mittens one manages **when it is cold**.'
<TH-Acc-Dance>	dánsut	Ollugat dánso **swinga** 'Many people danced **swing**'
<XT-Acc-Measure>	johtit	maŋimus **15 kilomehtera** johten johkafanassáhtuin. 'the last **15 kilometers** I travelled by riverboat.'
<XT-Acc-Time>	maŋŋonan	Barggut leat maŋŋonan badjel **guokte mánu** 'The work has been delayed over **two months**'
<XT-Ill-Money>	vuovdit	vuovdit **alimus haddái** 'sell **to the highest price**'
<IN-Ill-Lang>	čállit	čállet **sámegillii**. 'they write **in Sámi**.'
Negated/underspecified		
<TH-Com-*Ani>	veahkehit	váhnemat eai máhte veahkehit **leavssuiguin** 'parents cannot help **with the homework**'
<TH-Loc-Any>	ballat	Gánda guhte balai **gufihttariin** 'The boy who was afraid of the **underground beings**' / balan **čázis** 'I am afraid of **water**'

Table 3.13: Selection restrictions in valency tags in *valency.cg3*

the verb *bistit* 'last' has a DESTINATION-argument of the time prototype category.

(31) a. **adde** *su* *bargat* dal dan barggu
let him.ACC work.INF now the work
'they let him do the work now' (Magga, 1986, p.194)

b. Jus mahká dálveguohtun **addá** *vejolašvuođa* *ealihit* bohccuid
if alleged winter.pastures gives opportunity.ACC maintain.INF reindeer
'If we say that the winter pasture makes it possible to maintain the reindeer'

c. ...jorgalit luondduálbmoga **gielas** omd. **dárogillii**.
...translate nature.people's language.LOC e.g. Norwegian.ILL
'...translate from an indigenous language to for example Norwegian.'

d. Márjjábeaivvit álget bearjadaga, ja bistet **sotnabeaivái**.
Mary.days begin Friday, and last Sunday.ILL
'Marian feast days begin Friday and last until Sunday.'

Selection restrictions can further be used in semantic role annotation. They are used to identify accusative arguments of transitive verbs that predominantly appear intransitively. Potential objects of those verbs are usually semantically restricted, i.e. the object of *borgguhit* 'smoke' is typically a member of the substance-prototype category, and the object of *vázzit* 'walk' is typically a member of the education prototype cateogry as in ex. (32-a).

Selection restrictions can also help to match arguments with their governors in elliptical constructions. While the verb *suovvat* 'let, allow' has an animate accusative argument in its valency, i.e. *<AG-Acc-Ani><TH-Inf>*, *deaddilit* 'press, print' has an accusative of any type in its valency, i.e. *<TH-Acc-Any>*. In ex. (32-b), the inanimate accusative *namas* 'her/his name' can therefore unambiguously be matched with the governor *deaddilit* 'press, print'. The construction is elliptical, i.e. the accusative argument of *suovvat* 'let, allow' is missing, which makes it difficult to map the arguments in the first place.

(32) a. vázzit **skuvlla**
walk school.ACC
'go to school'

b. ...muhto son ii **suova** *namas* **deadd[i]lit** aviisii.
...but s/he not let name.ACC.PXSG3 print.INF newspaper.ILL
'...but s/he does not allow her/his name to be printed in the newspaper.'

In *valency.cg3*, selection restrictions generally refer to the prototypical use of the verb. Helander's (2001) description shows another approach. Helander (2001, p.69) uses inherent semantic features to show the full potential of semantic prototypes in specific arguments of the verb by specifying multiple alternative valency frames that differ only in their selection restrictions. In *valency.cg3*, I typically only show one selection restriction, either with a positive (e.g. *-Ani*) or negative restriction (*-*Ani*), unless differences in selection restrictions coincide with semantic role differences. If the non-prototypical use

is more frequent than the prototypical use, this will be reflected in the valency tag, cf. Table 3.13.

In grammar checking, selection restrictions are used to find erroneous morpho-syntactic realizations of a particular argument. This is done by associating arguments with their governors and annotating their roles. If the valency of a governor refers to a particular selection restriction of an argument, the argument can be distinguished from free modifications or arguments of other governors if it agrees with this particular selection restriction. Selection restrictions should therefore not be too restrictive. Preferably, they should only exclude impossible semantics. However, depending on the register, domain, etc., any semantic prototype may be possible and grammatical. Selection restrictions should therefore specify prototypical and frequent semantics or be left underspecified as *-Any*. Many prototypically physical verbs, such as *doallat* 'hold', have prototypical concrete accusative THEMES, but are used with a wide range of THEMES in *SIKOR* including many abstract THEMES, e.g. *dási* 'level (Acc.)' in ex. (33-a), *profiila* 'profile' (which should be in accusative case *profiilla* 'profile', not nominative) in ex. (33-b), and *sártni* 'speech' in ex. (33-c). They can be thought of as different senses of a polysemous verb. As word sense disambiguation is not the primary goal, the selection restriction to the accusative argument is left underspecified as *<TH-Acc-Any>*.

(33) a. . . . rusttet mii doallá alla internašu[vn]nalaš **dási**.
. . . equipment that holds high international level.ACC
'. . . equipment that meets international standards.'

b. Sámediggi berre dás doallat vuollegis ***profiila**.
Sámi.parliament should here keep low profile.NOM
'The Sámi parliament should keep a low profile regarding this.'

c. . . . doalai **sártni** gussiide.
. . . held speech.ACC guest.ILL.PL
'. . . s/he held a speech for the guests.'

3.2.3 Valency frames

A governor typically has multiple valency frames. The multiplicity of frames is due to different phenomena. Some are caused by rule-based diathesis alternations, while others are due to the facultativity of an argument, synonymous morpho-syntactic variants, and polysemy.

Verb	Synonymous arguments
dolkat 'get fed up'	THEME: locative, illative
liikot 'like'	THEME: illative, *locative, *accusative
sulastahttit 'resemble'	REFERENT: illative, accusative, *comitative
nohkkot 'run out of'	THEME: locative, illative
oahpásmuvvat 'get to know'	CO-ARGUMENT: comitative, illative
riidalit 'argue'	THEME: locative, *alde*, *badjel*, *geažil*

Table 3.14: North Sámi verbs with synonymous valencies

3.2.3.1 Synonymous valencies

In *SIKOR*, many synonymous morpho-syntactic realizations of the same argument types can be found, some of which are represented in Table 3.14. 'Synonymous' means here that the same semantic role is realized differently morpho-syntactically, leaving aside subclauses and non-finite constructions. These realizations can have slight differences in meaning.

The verb *sulastahttit* 'resemble' appears with both a REFERENT in illative case (*dieselmutuvrii* 'diesel motor'), cf. ex. (34-a), and one in accusative case (*lávlagiid* 'songs'), cf. ex. (34-b). *SIKOR* also includes examples with both a THEME in accusative case and a REFERENT in comitative case, cf. ex. (34-c). According go Informant *H* the sentence i ungrammatical, and *sulastahttit* 'resemble' should be replaced with *buohtastahttit* 'compare'. The verb *nohkkot* 'run out of' appears with both an illative and a locative THEME, cf. ex. (34-d)–(34-e).

(34) a. ...musihkka galgá sulastahttet ovtta boares **dieselmutuvrii**.
...music.NOM should resemble one old diesel.motor.ILL
'...the music should not resemble an old diesel motor.'

b. Luohti han sul[a]stahttá japánalaš boares **lávlagiid** ...
joik.NOM it resembles Japanese old song.ACC.PL ...
Joik resembles ancient Japanese songs ...'

c. **Dan** ii sáhte man ge láhkái *sulastahttit eará **gielaiguin**
that.ACC not can in any way compare other language.COM.PL
'That, one cannot compare in any way with other languages'

d. Mii nohkkuimet **mielkkis**.
we ran.out.of milk.LOC
'We ran out of milk.'

e. Son nohkkui **niestái**.
s/he ran.out.of food.ILL
'S/he ran out of food.'

Synonymous arguments are realized not only as different cases, but also as adpositional phrases. Ylikoski (2009, p.57) further mentions verbs that appear with arguments realized

by means of morphological cases and adpositional phrases in synonymous constructions. Many of these alternating constructions can be found in *SIKOR*. The THEME-argument of *riidalit* 'argue' can be realized as a nominal phrase in locative case (*mas* 'what') as in ex. (35-a), or as adpositional phrases with *geažil* 'because of' (cf. ex. (35-b)), *alde* (cf. ex. (35-c)), and *badjel* 'over' (cf. ex. (35-d)).

(35) a. Mii čuvget dás **mas** riidalit ja ...
we clarify here which.LOC argue and ...
'We clarify here what we dispute ...'

b. ... go lei riidalan muhtin áiddi **geažil** máŋga jagii[d].
... because had argued some fence.GEN.PL because.of many years
'... because s/he had argued over a fence for many years.'

c. riidalit luopmániid **alde**
argue cloudberry.GEN.PL on
'argue about cloudberries'

d. Son orruge dolkan riidalit bartta **badjel**.
s/he seems sick.of argue hut.GEN over
'S/he seems to be sick of arguing about the hut.'

3.2.3.2 Polysemy

Apart from synonymous realizations of certain arguments, the polysemy of a governor can justify multiple valency frames. According to Bick (2012), polysemy and the correspondence of a lexeme to many translation equivalents typically co-occurs with differences in the valency structure either syntactically or semantically. In *valency.cg3*, the verb *bidjat* 'put' is one of the verbs with the most valency frames (17 frames), some of which are related to its use in a multi-word expression, cf. Section 3.3, Table 3.22. While some valency tags belong to the same translation of a verb, e.g. *bidjat.1* 'put, place' in Table 3.15, most valency tags coincide with different translations.

(36) a. ... áigu bidjat dan **giehtagirjji** iežas *neahttasiidui*.
... intends put the handbook.ACC own website.ILL
'... s/he intends to put the handbook on his website.'

b. ... *johtui* bidjat dárbbašlaš heivehuvvon **oahpahusa**.
... motion.ILL put necessary adapted teaching.ACC
'... start necessary adapted teaching.'

c. Son oačču̇i maid oaggungilvvu *luhka*, **maid** galgá bidjat *ala*
s/he got also fishing.competition coat, which.ACC shall put on
'S/he also got a fishing competition coat, which she shall put on'

d. ... go biilla **mutuvra** biddjo *ala*.
... when car motor put.PASS.3SG on
'... when the car motor was turned on.'

e. Bija **uvssa** *gitta*!
put door.ACC closed

Meaning	**Translation**	**Valency tag**	**Example**
bidjat.1	put, place	∼ sth. somewhere (<TH-Acc-Any><DE-Ill-*Ani>), ∼ sth. on sth. (<TH-Acc-Any><LO-ala-Any>), ∼ sth. together (<TH-Acc-Any><oktii>)	ex. (36-a)
bidjat.2	start up, implement	<TH-Acc-Any><johtui>, <TH-Acc-Any><doibmii>, <TH-Acc-Any><fápmui>	ex. (36-b)
bidjat.3.a	dress, put on	<TH-Acc-Elect><ala>	ex. (36-c)
bidjat.3.b	turn on	<TH-Acc-Clth><ala>	ex. (36-d)
bidjat.4	close	∼ sth. <TH-Acc-Any><gitta>	ex. (36-e)
bidjat.5	present	∼ sth. (<TH-Acc-Any><ovdan>)	ex. (36-f)
bidjat.6	remove, put away	<TH-Acc-Any><eret>	ex. (36-g)
bidjat.7	name	<RE-Ill-Any><TH-Acc-Any><namman>	ex. (36-h)
bidjat.8	define	∼ sth. (<TH-Acc-Any>)	ex. (36-i)
bidjat.9	cause/get	∼ sb. to do sth. (<AG-Acc-Ani><TH-Inf>), ∼ sb. to go somewhere (<TH-Acc-Any><mátkái>)	ex. (36-j)
bidjat.10	start moving	∼ somewhere (<mátkái><DE-Ill-Plc>), (<TH-Inf>)	ex. (36-k)

Table 3.15: The valency variation of *bidjat* 'put' in *valency.cg3*

'Close the door!'

f. Ráđđehus bidjá *ovdan* **stuorradiggedieđáhusa** ...
government puts forward parliament.message.ACC ...
'The government presents the parliament message ...'

g. Jus biehttaledje, de sáhtii stivra bidjat *sin* **eret**
if refused, then could board put them.ACC.PL away
"vaikko goas".
"whenever"
'If they refused, then the board could fire them at anytime.'

h. ... **maidda** bijaime *namman* **Ginna**, **Galka**, **Borta**
... which.ILL.PL we.put name.ESS Ginna.ACC, Galka.ACC, Borta.ACC
'... which we named Ginna, Galka, Borta'

i. ... ja goas bidje **rájiid**?
... and when put.PRT.3PL border.ACC.PL
'... and when did they define the borders?'

j. It galgga **áhkát** bidjat *gođđit* ...
not should wife.ACC.PXSG2 put knit.INF ...
'You should not make your wife knit ...'

k. Sii geat ikte juo galge bidjat **mátkái** *Amerihkkái*.
they who yesterday already should put journey.ILL America.ILL
'The ones who already should have started the journey to America yesterday.'

The verb *bidjat* 'put' is not alone when it comes to polysemy. Verbs typically have more than one possible valency frame coinciding with polysemy and/or translation differences,

Verb	**Senses and valency tags**
earuhit	distinguish sth. from sth. <TH-Acc-Any><RF-Loc-*Plc>, dismiss sb. from their job <BE-Acc-Hum><LO-Loc-Pos>
riidalit	struggle with sth. <TH-Com-Any>, argue with sb. <CO-Com-Ani> / about sth <RS-nalde-Any>
bivdit	hunt sb. <RE-Loc-Ani>, ask sb. about sth. <TH-Acc-Any> <TH-Acc-*Ani><RE-Loc-Ani>
čuovgat	shine, light up, give light (to) <BE-Acc-Any>, receive <TH-Acc-Any>
bođđet	'incite sb. to do sth.' <TH-Acc-Any><bajás> <EX-Acc-Ani><TH-Inf>, 'distinguish sth. from sb.' <TH-Acc-Any><RF-Loc-*Plc> 'separate'
cealkit	'speak' <TH-birra-Any>, 'tell' <RE-Acc-Ani><TH-Inf>, 'fire' <TH-Acc-Hum><eret>
čuojahit	'play' <IN-Acc-Any>, 'call' <RE-Acc-Ani><TH-Inf> <RE-Acc-Ani>

Table 3.16: Some polysemous verbs and their valencies in *valency.cg3*

cf. Table 3.16. The verb *earuhit* is translated as 'dismiss' with a human accusative and a locative of the prototype position, as in ex. (37-a), and as 'distinguish' with an accusative argument and locative that is not of the place prototype category, as in ex. (37-b). Here, selection restrictions distinguish the two senses. The verb *riidalit* is translated as 'struggle' with a comitative THEME in ex. (37-c) and as 'argue' with a THEME realized by an adpositional phrase with *nalde* 'on' in ex. (37-d).

(37) a. . . . earuhii sámelogahaga *rektor* [. . .] **virggistis**.
. . . dismissed Sámi.college principal.ACC [. . .] position.LOC.PXSG3
'. . . dismissed the principal of the Sámi college [. . .] from his position.'

b. . . . muhto mii earuha *mu* seamma **ahkásačča**in Oslos?
. . . but what distinguishes I.GEN same age.LOC.PL Oslo.LOC
'. . . but what is it that distinguishes me from people my age in Oslo?'

c. . . . riidala **teoriijain**.
. . . struggle theory.COM
'. . . s/he struggles with the theory.'

d. . . . lea álbmot ja boazodoallu riidalan luossabivddu **nalde**.
. . . have people and reindeer.herding argued salmon.fishing.GEN on
'. . . the people and the reindeer herding industry have argued about salmon fishing.'

3.2.3.3 Diathesis alternations

Diathesis alternations are changes in the morpho-syntactic realizations of the same arguments of a governor, which can have slight differences in meaning. That means that constellations of the semantic roles change morpho-syntactically, either qualitatively or quantitatively. Practically, diathesis alternations can cause multiple valency tag assignments to a governor and constrain other valency tag assignments, which is why their behavior needs to be taken into account in valency annotation. I distinguish between alternations involving derivational affixes (i.e. passive, causative, reflexive, and reciprocal) and alternations where the change in the valency frame is not marked on the verb morphologically.

3.2.3.3.1 Alternations involving morphological derivations

In North Sámi, passive, causative, reflexive and reciprocal alternations all co-occur with morphological derivational processes. In *lexc*, some of those processes correspond to an underived form with derivational tags, while others are lexicalized, that is, they are listed under a new lemma, e.g. *rahpasit* 'open' as in 'the door opens' or *joatkašuvvat* 'be continued', cf. Table 3.17. However, the verb's semantic behavior depends not only on the type of derivational tag it has, but also on its combination with a lemma. As I do not deal with subject-roles systematically in *valency.cg3*, I primarily discuss the effects of the derivations on valency changes affecting object- and adverbial-arguments.

Passive derivations affect various arguments of a verb. Typically, the object of the active counterpart is moved into subject position unless it is deleted. The subject, on the other hand, disappears altogether from the valency or becomes a facultative argument. While I assume that those are different syntactic realizations of the same argument, Sammallahti (2005, p.61) distinguishes between different types of passives with different implications on his semantic roles. Below I will discuss three effects of passive derivations. Firstly, they can add a facultative argument in illative case to the original verb. Secondly, verbs with accusative objects in the active form lose their accusative object in their passive form. Thirdly, passive derivations of intransitive verbs or transitive verbs with restricted objects can become avalent verbs without a subject role.

Passives that can have an argument in illative case are categorized as "adversative passives" by Sammallahti (2005, p.62). According to Sammallahti (2005, p.62), the illative is an animate AGENT. However, inanimate illative arguments such as *vieruide* 'by the customs' in ex. (38-a) have the AUTOMATON-role. In *valency.cg3*, the illative is considered an AGENT irrespective of the animacy of the argument. The valency tag added to the derived verb is *<AG-Ill-Ani>*. As *báinnahallat* 'be influenced' is lexicalized in *valency.cg3*, the tag is directly added to the lemma. Its orthographical variant *báinnáhallat* is listed under the active form *báidnit* 'influence' and receives the valency tag in a combination

with the derivational tags *Der/h* and *Der/alla* or only *Der/halla*.

Valency changes with regard to the object role also require restrictions to accusative argument rules mapping, e.g. *<PA-Acc-Any>*. Those restrictions need to exclude passive forms, i.e. those forms receiving the tags *Der/h* and *Der/alla* or *Der/halla*. While the verb *báidnit* 'influence' is annotated with the tag *<PA-Acc-Any>*, passive forms need to be excluded from the annotation by a negative constraint, cf. also Section 3.2.4.

Impersonal passives such as the third person singular form *dánsojuvvui* 'there was dancing' in ex. (38-c) change the valency of the verb, making it avalant. The verb 'loses' the subject argument of the active form, and are marked with the valency tag *<0>* in *valency.cg3*.

(38) a. **Máhtte** lea báinnahallan dáčča **vieruide**.
Máhtte has influenced.PASS.PRFPRC Norwegian custom.ILL.PL
'Máhtte has been influenced by Norwegian customs.'

b. Dáčča **vierut** leat báidnán **Máhte**.
Norwegian custom.NOM.PL have influenced Máhtte.ACC
'Norwegian customs have influenced Máhtte.'

c. Dánsojuvvui.
dance.PASS.PRT.3SG
'There was dancing.'

d. Máhtte oidnojuvvui.
Máhtte see.PASS.PRT.3SG
'Máhtte was seen.'

e. Máhtte oidnui.
Máhtte see.PASS.PRT.3SG
'Máhtte was visible.'

As derivational tags are typically ambiguous, ambiguities need to be taken into account in restrictions to object-role mapping. While the derivational tags *Der/PassL*, cf. *oidnojuvvui* 's/he was seen' in ex. (38-d), and *Der/PassS*, cf. *oidnui* 's/he was visible' in ex. (38-e), are unambiguous with regard to their object-lessness, they can be referred to directly in negative conditions of the respective rules. However, the derivational tags or tag combinations *Der/h Der/adda*, *Der/h Der/alla*, *Der/halla* are also used for passive derivations, i.e. loss of accusative argument, and frequentative derivations maintaining the valency structure. In order to assign the correct valency tag, one must disambiguate between passives and frequentatives. While the tag combination alone is ambiguous, one must test the verb + morphological tag combinations with regard to their ambiguity. Adversative passives can have both *Der/PassL*, *Der/PassS*, *Der/h + Der/alla*, *Der/h + Der/adda* and *Der/halla* tag combinations, which in turn can be passive types where the illative remains unexpressed. Therefore, the morphological tags alone cannot be used to match lemmata with the *<AG-Ill-Ani>* tag. The lemma itself needs to be categorized with regard to its ability to form a certain diathesis alternation, and in case of ambiguity

the syntactic context needs to be specified. Nielsen (1932-1960*a*, p.227) lists two entries for *borahallat*, i.e. 'give food several times, give several (animals or children) food' (frequentative) and 'be bitten (by)' (passive). However, in *SIKOR*, all 68 occurrences are passives as in ex. (39-a), cf. also Table 3.18. Also *gáskkahallat* 'be bitten', and *oainnahallat* 'unintentionally be seen, get caught' are unambigious passives in *SIKOR* except for one ambiguous case. The verbs *heivehallat* 'try to get to suit (sth.)', and *oahpahallat* 'try to learn, teach many times', on the other hand, do not have any passive occurrence in *SIKOR*. While *heivehallat* 'try to get to suit' has mostly frequentative causative readings, cf. ex. (39-b), *oahpahallat* is ambiguous with regard to frequentative causative ('teach many times'), cf. ex. (39-d), vs. conative uses ('try to learn'), cf. ex. (39-c). Although there is no ambiguous passive-frequentative example in the corpus among the verbs investigated, the frequentative (lexicalized) causative-conative ambiguity implies differences in quantitative valency. While the lexicalized causative can have an illative BENEFICIARY, the conative reading cannot. The verb *oahpahallat* is therefore annotated with both *<TH-Acc-*Ani><BE-Ill-Ani>* and *<TH-Acc-*Ani>* at the same time. The second valency tag can apply both for the conative and for the causative reading. For verb sense disambiguation, this distinction would therefore not be sufficient.

(39) a. . . . de borahallá **čuoikkaide**
. . . then eat.PASS.PRS.3SG mosquito.ILL.PL
'. . . then s/he gets bitten by the mosquitoes'

b. . . . movt šaddet heivehallat *iežaset* ođđa **servodahkii**.
. . . how will.PRS.3PL adapt.FREQ.INF themselves.ACC new society.ILL
'. . . how will they adapt themselves to the new society.'

c. mánát ieža oahpahallet **sámegiela** mánáidgárddis
children themselves learn Sámi.ACC kindergarten.LOC
'children themselves learn Sámi in kindergarten'

d. oahpahallat **sámegiela** daidda ráves *sámiide* . . .
teach Sámi the.ILL.PL grownup Sámi.ILL.PL . . .
'teach Sámi to the Sámi adults . . . '

The second relevant diathesis alternation is the causative derivation. Morphological causatives are formed with the derivational tags *Der/Caus* (*jorgalahttit* 'make translate'), and *Der/h* (*borahit* 'make eat, feed'). While *Der/Caus* is morphologically unambiguous, forms that are annotated with *Der/h* can also be frequentative if they appear in combination with other derivational tags, i.e. *Der/alla* as *heivehallat* 'adapt in many ways'. Additionally, morphological derivations do not necessarily coincide with semantic causatives. Prototypically, a causative AGENT is added to the non-causative verb's valency, as in ex. (40-a), where the accusative/illative AGENT *Márehii/Máreha* enhances the valency frame of the verb. Sammallahti (2005, pp.77–79) calls these constructions "curative causatives". Other morphological causatives such as "transportative causatives", e.g.

Derivational tag	**Examples and valencies**
Der/PassL	*oidnojuvvot* (intentional passive: <TH-Nom-Any>), *dájuhuvvot* (adversative passive: <TH-Nom-Any><AG-Ill-Any>), *dánsejuvvui* (impersonal passive <0>)
Der/PassS	*oidnot* (automative passive <TH-Nom-Any>), *borrot* (adversative passive <TH-Nom-Any><AG-Ill-Any>)
Der/h + Der/alla, Der/h + Der/adda	*oainnáhallat* (adversative passive <TH-Nom-Any> <AG-Ill-Any>), *heivehallat* (frequentative <TH-Nom-Any>/<AG-Ill-Any>)
Der/halla	(only for transitive verbs) *oainnáhallat* (adversative <TH-Nom-Any>/<AG-Ill-Any>)
Der/h	*borahit* (permissive PA-Nom), *borahit* (causative <AG-Nom-Any> <AG-Ill-Any><PA-Acc-Any>), *goaruhit* (causative)
Der/d	*dovddadit* (reciprocal <AG-Nom-Any>/<TH-Nom-Any>), *basadit* (reflexive <AG-Nom-Any>/<TH-Nom-Any>), (continuative <AG-Nom-Any><TH-Acc-Any>), *divodit* (frequentative)
lexicalized	*rahpasit* 'open' (automative passive <TH-Nom-Any>), *joatkašuvvat* (automative)

Table 3.17: Valency tags for derived verbs in *valency.cg3*

njiejahit 'decrease' in ex. (40-b), which is derived from the intransitive *njiedjat* 'descend' (cf. ex. (40-b)), have a lexicalized meaning, and their additional argument, according to Sammallahti (2005, p.75), can no longer be considered an AGENT, but is rather a THEME or a PATIENT. None of the 14 occurrences of *njiejahit* 'decrease' has (or can have according to Informant *H*) an animate object. In *valency.cg3* they are annotated as regular transitive verbs with the tags *<TH-Acc-Any><SO-Loc-Any>*, *<TH-Acc-Any><DE-Ill-Any>* and *<TH-Acc-Any>* just like verbs such as *doalvut* 'bring'. Other morphological causatives, e.g. *gulahit* 'announce' from *gullat* 'hear', and *dieđihit* 'inform' from *diehtit* 'know', have an idiomatic meaning, cf. Vinka (2002, p.150) and can be lexicalized further like *gulahahttit* 'make announce', shown in ex. (40-e), and *dieđihahttit* 'make inform'.

(40) a. Máhtte goaruhii **Márehii/Máreha** gávtti.
Máhtte sew.CAUS.PRT.3SG Máret.ILL/Máret.ACC costume.ACC
'Máhtte made Máret sew the costume.'

b. Máret njiejahii **muoraid** váris.
Máret bring.down.PRT.3SG tree.ACC.PL mountain.LOC
'Máret brought the trees down from the mountain.'

c. Máret njiejai váris **muoraiguin**.
Máret go.down.PRT.3SG mountain.LOC tree.COM.PL
'Máret went down from the mountain with the trees.'

d. Báhppa **gulahii** heajaid.
pastor.NOM hear.CAUS.PRT.3SG wedding.ACC.PL
'The pastor announced the wedding.'

e. Mon báhpa **gulahahtten** heajaid.
I pastor.ACC announce.CAUS.PRT.1SG wedding.ACC.PL
'I made the pastor announce the wedding.'

As Table 3.18 shows, causative AGENTS are not only facultative, but also very infrequent in *SIKOR*. Typically, not only causative AGENTS but also the object roles of the non-causative can be omitted under various circumstances. In ex. (41-a)–(41-b), the causative *goaruhit* 'cause to sew, get sewn' appears without a causative AGENT and without a PATIENT-object, i.e. *čalmmi* 'eye' and *gávtti* 'costume'. However, they can be inferred from the context. The verb *lávlluhit* 'make sing', on the other hand, appears more frequently with a causative AGENT only, as in ex. (41-d). Typically, in constructions with both accusative and illative, the illative is interpreted as the BENEFICIARY rather than the CAUSATIVE AGENT,[15] like *buot mánáide ja bargiide* 'to all children and workers' in ex. (41-c). For *borahit* 'make eat', most of the examples have only one argument besides the subject, i.e. either a PATIENT (10 occurrences) like the accusative *tableahtaid* 'pills' in ex. (41-f), or an AGENT (34 occurrences) like *su* 's/he (Acc.)' in ex. (41-e). In addition, there are 32 occurrences of constructions with a PATIENT-subject, cf. ex. (41-g), cf. Sammallahti (2005, pp.67–69) ("permissive passive").

(41) a. ...mánná bártidii, soabbi basttii čalmmi bajil ja son šattai
...child was.in.accident, rod cut eye.GEN above and s/he had
vuolgit doaktára lusa **goaruhit**.
go doctor to sew.INF
'...a child was in an accident, a rod cut into the flesh above his/her eye and s/he had to go to the doctor to get it sewn.'

b. Háliidan gávtti, Gáivuona gávtti[], muhto in leat vel šaddan
want costume, Gáivuotna costume, but not have still become
goaruhit
sew.CAUS.INF
'I want a costume, a Gáivuotna costume, but haven't gotten around to getting it sewn'

c. leat sii goaruhan luhkaid buot **mánáide** ja
have they sew.CAUS.PRFPRC coat.ACC.PL all child.ILL.PL and
bargiide.
worker.ILL.PL
'they have ordered coats to be sewn for all children and workers.'

d. Niilo Rasmus lávlluha Ohcejoga **skuvlamánáid**.
Niilo Rasmus sing.CAUS.PRS.3SG Ohcejohka school.child.ACC.PL
'Niilo Rasmus made the Ohcejohka schoolchildren sing.'

[15]Informants H and N agree on the BENEFICIARY interpretation

e. ja sierra gáris borahii **su**.
and different bowl eat.CAUS.PRT.3SG him/her.ACC
'and s/he made him eat from another bowl.'

f. Borahit **tableahtaid**
eat.CAUS.INF pill.ACC.PL
'Make eat pills'

g. **Biergu** gal ii borat.
meat.NOM definitely not eat.INF
'The meat is definitely not edible.'

If Sammallahti's (2005) and Svonni's (2015) examples are excluded from the corpus material, constructions with two accusative arguments are only found once, cf. ex. (42). The example includes an accusative AGENT, i.e. *nieiddaid* ja *gánddaid*, and a topicalized accusative THEME, i.e. *divtta* 'poem'.

(42) ?**Divtta** sáhttá koaralohkama bokte dahje osiid vurrolagaid
Poem.ACC can choir.reading through or part.ACC.PL by.turns
nieiddaid ja *gánddaid* logahit.
girl.ACC.PL and boy.ACC.PL read.CAUS.INF
'One can make the girls and boys read the poem by turns or simultaneously.'

Causatives with accusative and illative arguments are more frequent. However, they only occur with specific verbs, cf. *borahit* 'make eat' (14 occurrences) and *jáhkkihit* 'make belive' in ex. (43-a). The verb *jáhkihit* 'make believe' also appears with THEMES expressed as subclauses and non-finite forms, cf. ex. (43-b) in addition to an illative or accusative AGENT, cf. also Vinka (2002, pp.55–56), who considers constructions of that type ungrammatical.

(43) a. Dat lea goit čilgehus *maid* **ránnjääahkuide** ledjen
that is anyway explanation which.ACC old.lady.neighbor.ILL.PL have
jáhkihan.
believe.CAUS.PRFPRC
'Anyway, that is the explanation that I have made the neighbor ladies believe.'

b. de livčči dát filbmenvuohki jáhkihan **mu** ahte dan maid
then would this way.of.filming believe.CAUS.PRFPRC I.ACC that that what
oainnán lea duohta.
see is true
'then this way of filming would make me believe that what I see is true.'

Reflexive and reciprocal derivations typically reduce the valency, as the object role is fused with the subject-role. Reflexive alternations move the object role into subject position. The verb *čuohpadit* 'cut oneself' is derived from the transitive verb *čuohppat* 'cut' and becomes intransitive. According to Sammallahti (2005, p.71), the subject of *čuohpadit* 'cut oneself', i.e. *Máhtte* in ex. (44-a), is both an AGENT and a PATIENT. While

Passive - frequentative			
Verb	**Passive**	**Frequentative**	**Others**
heivehallat	0	111	11 (reflexive), 1 (reciprocal)
oahpahallat	0	12	500 (conative)
oainnahallat	49	0	-
gáskkahallat	20	1 (undecided)	-
borahallat	68	0	-
Causative			
	+ causative AGENT only	**- causative AGENT**	**two arguments**
goaruhit	1 (Ill.)	65 (PRODUCT), 18 (PRODUCT+ BENEFICIARY)	7 (Ill.+Acc.), 2 (Acc.+ Acc.)
lávlluhit	6 (Acc.)	1	2 (Ill.+Acc.)
jáhkihit	11 (Acc.), 2 (Ill.)	13	16 (Acc.+*ahte*), 4 (Acc.+Inf.), 1 (Acc.+*dihte*)/Ill./Loc./finite sub-clause, 11 (Ill.+Acc.), 11 (Ill.+*ahte*)
borahit	36 (Acc.)	7 (Acc.), 32 (Nom.)	11 (Ill.+Acc.), 1 (Acc.+Acc.), 1 (*Acc.+Ill.)
logahit	1 (Acc.)	59 (Acc.)	6 (Ill.+Acc.), 2 (Acc.+ Acc.)
Reflexive			
	reflexive	**continuative**	
basadit	115	3	-
čuohpadit	8	131	-
geassádit	1,242	-	16 (transitive)
Reciprocal			
	reciprocal	**+ Acc.**	**+ Com.**
dovddadit	34	2	16
vuoiddadit	9	17	6
oaidnalit	106	19	-
riidalit	138	39	-

Table 3.18: The distribution of valencies of passive, causative, reflexive, and reciprocal verbs in *SIKOR*

the derivational tag *Der/d* is ambiguous with regard to the diathesis alternation and its effect on the valency, the combination of lemma and derivational tag can be unambiguous. The form *basadit* (*bassat* Der/d) 'wash oneself' is purely reflexive, as are most of the instances of *geassádit* 'withdraw' (1,243). However, five of these are analytical reflexives of *geassádit* 'withdraw', cf. ex. (44-c) where a reflexive pronoun (*iežaska* 'themselves') is used with the reflexive verb. The verb is also used transitively in the meaning of 'retract', cf. ex. (44-d). In *SIKOR*, there are only 8 reflexives. Much more frequent are transitive uses (continuative), as in ex. (44-b) or in the meaning 'perform surgery' (78 occurrences).

(44) a. **Máhtte** čuohpadii.
Máhtte cut.REFL.PRT.3SG
'Máhtte cut himself.' (Nickel and Sammallahti, 2011, p.409)

b. oahppi čuohpada **muitogoarttaid**.
student cuts.out commemoration.card.ACC
'the student cuts out commemoration cards.'

c. geassádan ?**iežaska** válgalisttus.
withdrawn oneself.3DU election.list.LOC
'they have withdrawn themselves from the election list.'

d. de molsu mearrádusa ja geassáda **váidaga**.
then changes decision.ACC and retract complaint.ACC
'then s/he changes the decision and retracts the complaint.'

The reciprocal alternation typically presupposes a symmetric relation between subject and object. The object is moved to subject position multiplied by coordination (alternatively by a plural) and the object position is deleted from the surface syntactic structure or realized with a reflexive pronoun (Kettnerová and Lopatková, 2013, p.160). According to Sammallahti (2005, p.71), the subject (*Máhte guovttos Márehiin* 'Máhtte and Máret') of a reciprocal verb such as *dovddadit* (*dovdat* Der/d) 'know each other' in ex. (45-b) has both an EXPERIENCER- and THEME-role. The verb *dovdat* 'know', on the other hand, has an EXPERIENCER-subject, and a THEME-object, cf. ex. (45-a). In *valency.cg3*, the second argument is annotated as CO-ARGUMENT.

In *SIKOR*, *dovddadit* 'know each other' is not only used reciprocally (34 occurrences). 16 occurrences can be considered analytical reciprocal constructions with an explicit reciprocal pronoun such as *guhte guimmiideaset* 'each other' as in ex. (45-d). Other non-reciprocal uses include constructions with a second argument in comitative case, such as *geainna* 'who (Com.)' in ex. (45-e). In ex. (45-f) *dovddadit* 'know each other' appears with an accusative, which does not satisfy the valency restrictions of the verb. According to Informant *H* and Informant *N*, the argument should be in comitative case. Alternatively, the verb should be *dovddiidit* 'get to know' as it is used with illative and accusative respectively, which can be identified by means of its valency.

(45) a. Máhtte dovdá **Máreha**
Máhtte knows Máret.ACC
'Máhtte knows Máret' (Nickel and Sammallahti, 2011, p.409)

b. **Máhte guovttos Márehiin** dovddadeaba
Máhtte.GEN together Máret.COM know.RECIP.3DU
'Máhtte and Máret know each other' (Ibid.)

c. Mii beassat dovddadit árgabeaivválaš **dáhpáhusaide** ja **muittuide**.
we get.to know everyday event.ILL.PL and memory.ILL.PL
'We get to know everyday events and memories.'

d. ...go ollu čiekčit dovddadit **guhte guimmiideaset**.
...since many players get.to.know each other.ACC.PXPL3
'...since many players get to know each other.'

e. Risten lei áidna, **geainna** lei álo bures dovddadan.
Risten was only, who.COM had always well known
'Risten was the only one whom he s/had always know well.'

f. Mii beassat dovddadit ?**Ándde**
we get know Ánde.ACC
'We get to know Ánde'

3.2.3.3.2 Alternations without morphological derivations

There are other alternations that do not involve morphological derivations but still change the quantitative or qualitative valency structure of certain verbs, cf. Table 3.19. Those need to be listed in separate sets that annotate valency tags and cannot be codified by morphological tags. In ex. (46-a), the argument of *vuodjit* 'drive' expressing the vehicle alternates between comitative *biillain* 'with the car' and accusative *biilla* 'the car', the latter of which is not an acceptable construction to Informant *H*. While the comitative construction is preferred by Informant *N*, the accusative is still acceptable. In *SIKOR* there are 223 instances of the lemma *vuodjit* 'drive' with a form of *biila* 'car' to its right, of which there are 33 (15%) accusatives, and 171 (85%) comitative forms. The verb *vuodjit* 'drive' further alternates between an intransitive motion and a transitive motion, synonymous to *vuojihit* 'transport sb.' with an accusative argument, e.g. *du* 'you' in ex. (46-b).

There are further alternations affecting the transitivity of certain verbs. Some verbs are used predominantly without an accusative argument, but can have an accusative argument in certain constructions. The accusative objects of those predominately intransitive verbs are typically restricted to certain semantic prototypes, substance (*duhpáha* 'tobacco') in the case of ex. (46-c), and place (*bávttiid* 'rocks') in ex. (46-d). The verb *borgguhit* 'smoke' has 44 instances (11%) with an expressed object and 404 instances with an unexpressed object (89%). Only 3 of 60 instances (5%) of *gakcut* 'climb' are used with an accusative object. 46 instances (77%) are used with a DESTINATION realized as illative case or a postpositional phrase with e.g. *ala* 'on', *badjel* 'over', etc. Annotating these verbs'

valency is useful as *lexc* transitivity tags, i.e. *IV* for intransitive verbs, and *TV* for transitive verbs, can only make very general specifications. In *verbs.lexc*, these verbs are either classified as transitive (as *borgguhit* 'smoke') or intransitive (as *gakcut* 'climb'). If the verb is classified as transitive, disambiguation rules are likely to analyze genitive-accusatives in object-less constructions as objects, rather than adverbials or objects of other verbs. In the case of intransitive verbs, on the other hand, disambiguation rules may not be able to capture the cases where the verb does appear with an object.

(46) a. vuodjit **biilla** vs. vuodjit **biillain**
drive car.ACC vs. drive car.COM
'drive the car'

b. mun **vuojan** vs. **vuojihan** du ruoktot
I drive.PRS.1SG vs. drive.CAUS.PRS.1SG you.ACC home
'I drive you home'

c. ...nissonolbmuide geat borgguhit **duhpáha**, lea váddáset šaddat
...women who smoke tobacco.ACC, is difficult become
mánálahkai
pregnant
'...for women who smoke tobacco it is more difficult to become pregnant'

d. Gavcco **bávttiid**.
climb.IMPRT.2SG rocks.ACC.PL
'Climb rocks.'

The verbs *dollet* 'grab' and *duolbmalit* 'tramp' are typically used with an accusative object. However, as multi-word verbs with *johtui* 'to the motion' or with a DESTINATION, e.g. *birra máilmmi* 'around the world' they are used intransitively. While the transitivity tag only specifies the verb's potential to appear with an accusative argument, valency tags can distinguish between the transitive reading (*<TH-Acc-Any>*) and the intransitive multi-word verb reading (*<johtui><DE-Ill-Plc>*).

(47) a. ...dollejit Soltun álbmotallaskuvlaoahppit **johtui** birra
...start.PRS.3PL Soltun college.students motion.ILL around
máilmm[i]
world.GEN
'...Soltun college students start travelling around the world'

b. Iđđedis mii dollet **johtui** Rimii
morning.LOC we started motion.ILL Rimi.ILL
'In the morning we headed for Rimi'

c. ...nuorat duolbmaledje **johtui** lávvardaga gáhtavuodjimis
...youth tramp motion.ILL Saturday street.driving.LOC
'...young people started pedalling Saturday at the street competition'

Other alternations do not involve quantitative valency changes, but qualitative changes. Helander (2001, p.65) describes an alternation of the verb *boahtit* 'come' between a frame with an illative and an essive argument as in ex. (48-a), and an infinitive and an ac-

Verb example	**Alternation: valency tags**
dollet 'grab'	<TH-Acc-Any>, <johtui><DE-Ill-Plc>
duolbmalit 'tramp'	<TH-Acc-Any>, <johtui><DE-Ill-Plc>
vuodjit 'drive'	<DE-Ill-Any>, <TH-Acc-Ani><DE-Ill-*Ani>
borgguhit 'smoke'	<PA-Acc-Substnc>
gakcut 'climb'	<LO-Acc-Plc>
boahtit 'come'	<PU-Inf>, <BE-Ill-Ani><veahkkin>
oahpahit 'teach'	<TH-Acc-*Ani><BE-Ill-Ani>, <BE-Acc-Ani><TH-Loc-Any>

Table 3.19: North Sámi verbs that participate in alternations without morphological derivations

cusative argument as in ex. (48-b). This alternation applies to most intransitive motion verbs (like *mannat* 'go', *girdit* 'fly', *joavdat* 'reach', *vuolgit* 'leave', etc.). While PURPOSE is expressed by a noun in essive case in ex. (48-a) (*veahkkin* 'as a helper') and the BENEFICIARY as an illative (*munnje* 'to me'), PURPOSE is expressed as an infinitive in ex. (48-b). The BENEFICIARY is an argument of the infinitive, not the matrix verb, hence the BENEFICIARY is missing in the valency frame of *boahtit* 'come' (*<PU-Inf>*). Other verbs like *oahpahit* 'teach', *neavvut* 'advise' and *rávvet* 'advise' alternate between a construction with the BENEFICIARY in illative case (*munnje* 'to me') and the THEME in accusative case (*d[á]rogiela* 'Norwegian'), cf. ex. (48-c), and a construction with the BENEFICIARY in accusative case (*mánáid* 'children') and the THEME in locative case (*oskkoldagas* 'religion (Loc.)') as in ex. (48-d).

(48) a. Dat bođii **munnje** *veahkkin*
s/he came I.ILL helper.ESS
'S/he came to help me'

b. Dat bođii *veahkehit* **mu**
s/he came help I.ACC
'S/he came to help me'

c. Báhppa oahpahii **munnje** *d[á]rogiela*.
priest taught I.ILL Norwegian.ACC
'The priest taught me Norwegian.'

d. Vuoigatvuohta oahpahit **mánáid** *oskkoldagas* ...
right teach.INF children.ACC religion.LOC ...
'The right to teach children about religion ...'

3.2.4 Valency rules in *valency.cg3*

The *CG* valency annotation grammar *valency.cg3* includes sets of potential governors that share at least one valency frame and rules that annotate valency tags to the members of these sets. The annotation rules are simple *SUBSTITUTE*-rules, which replace a certain part of speech with the same part of speech and a valency tag; cf. Didriksen (2010, pp.24–25) for the rule format. In Constraint Grammar, each analyzed token forms a cohort as below, i.e. one line with the form that is analyzed, and as many lines with lemma and tag combinations as there are distinct analyses. For the form *diehtit* 'know', there are two possible morphological analyses. One is an infinitive analysis (*Inf*) of the lemma *"diehtit"*. The other is a first person plural indicative analysis (*Ind Prs Pl1*) of the same lemma (*"diehtit"*) as illustrated below. While morphological differences are realized as different readings, which are to be removed or picked by disambiguation rules of a grammar, valency tags are simply added to a specific lemma without producing a new line. Rather than producing ambiguity they only increase the length of a line. This makes sense, as valency or verb sense disambiguation is not the foremost goal of valency annotation. Instead of adding ambiguity, valency tags are mostly used to reduce morpho-syntactic ambiguity.

```
"<diehtit>"
    "diehtit" V <TH-birra-Any> <TH-FS-Qst> <TH-ahte> <TH-Acc-Any> TV Inf
    "diehtit" V <TH-birra-Any> <TH-FS-Qst> <TH-ahte> <TH-Acc-Any> TV Ind Prs Pl1
```

Different Constraint Grammar rule types are used for either adding to ambiguity and more cohort lines, i.e. *MAP*-rules, or simply adding to the length of each line, i.e. *SUBSTITUTE*-rules. *SUBSTITUTE*-rules replace a certain tag (combination), here *V*, with another one, i.e. *V <TH-Acc-Any>*, including a valency tag to a specific target, here the set *TH-ACC-V*. This set specifies lemmata that have a THEME in accusative case, e.g. *diehtit* 'know', *dadjat* 'say', etc. *SUBSTITUTE*-rules can further specify context conditions for the annotation of this valency tag. Context conditions are specified by means of numbers referring to the relative position with regard to the target. Many context conditions refer to the form itself, i.e. *0*. Here the target is specified negatively, i.e. the form cannot be a passive form *(NEGATE 0 Der/PassL OR Der/PassS)*.

```
SUBSTITUTE (V) (V <TH-Acc-Any>) TARGET TH-ACC-V IF (NEGATE 0 Der/PassL OR Der/PassS)
```

However, sentential context can also be specified and valencies can potentially be disambiguated depending on the syntactic context without adding to the ambiguity of the cohort. *Valency.cg3* is much more potent than a regular valency lexicon, and can be thought of as a hybrid between a lexicon and a grammar, cf. also the structure of

Rules		%
SUBSTITUTE	440	100%
verb rules	410	93.2%
noun rules	20	4.5%
adjective rules	8	1.8%
adverb rules	1	0.2%

Table 3.20: Rule distribution within *valency.cg3* version r146069

VALLEX, which includes valency alternations in a grammar part of the lexicon (Žabokrtský and Lopatková, 2007).

Most rules of *valency.cg3*[16] refer to verbal governors (93.2%), cf. Table 3.20. 4.5% of the rules annotate valencies to nominal governors, and 1.8% to adjectival governors. Only one rule adds valency tags to adverb governors.

As discussed earlier, valency-changing derivations and inflections need to be taken into account in the annotation process. When annotating a valency tag that refers to an accusative PATIENT, both passive and reflexive derivations of base verbs with accusative PATIENTS need to be excluded as these derivations reduce the valency of the base verb. Unambiguous derivational tags can be referred to directly, e.g. *Der/PassS* (short passive), *Der/PassL* (long passive), *Der/Caus* (causative) and *Der/ahtti* (causative). Lexicalized verbs are referred to via sets, e.g. *CAUS-PA-ACC-ANY* (lexicalized causative verbs that can have a PATIENT and an unexpressed AGENT). Lastly, combinations of derivational tags and verb sets are used to refer to ambiguous derivations that are (mostly) unambiguous in lemma + tag combinations. The set *REFL-DER/D-V* includes lemmata of verbs that are unambiguously reflexive in combination with *Der/d*, e.g. *bassat* 'wash'. In combination with other tags, there can be further ambiguity. *Der/h* and *Der/alla* can mark a passive form, e.g. *gáskkáhallat* 'be bitten', or a causative frequentative form, e.g. *heivehallat* 'adapt'. Systematically ambiguous derivations are referred to by their tag and with a syntactic constraint, as in the derivational tag *Der/halla* with an animate illative (*LINK *0 Ill + Sem/Animate*), which is thought to be passive. This context condition refers to the left or right context without specifying a distance, i.e. **0*, and therefore performs verb sense disambiguation. The valency annotator *valency.cg3* is clearly used as a grammar here, and not just a lexicon. In ex. (49), *gáskkáhalai* 'was bitten' has the tag combination *Der/h Der/alla*, but is not a frequentative causative, but clearly a passive because of its illative animate AGENT *beatnagiidda* 'by the dogs'.

(49) ...gáskkáhalai heargi beanavuoddji **beatnagiidda**
...bite.CAUS;FREQ.PRT.3SG reindeer dogsledder.GEN dog.ILL.PL
'...reindeer was bitten by the dogsledder's dogs'

[16]version r146069 (Accessed 2017-01-05)

```
SUBSTITUTE (V) (V <PA-Acc-Any>) TARGET PA-ACC-V - Der/PassL - Der/PassS
OR CAUS-PA-ACC-ANY IF (NEGATE 0 (Der/h Der/alla) OR Der/halla
LINK *0 Ill + Sem/Animate BARRIER NPNHA - Pcle)(NOT 0 Der/d + REFL-DER/D-V );
```

To annotate valency tags specifying animate CAUSATIVE AGENTS of certain causative verbs, the rule refers both to lexicalized causatives, i.e. the set *AG-ACC-ANI-V* with its members *barggahit* 'make work', *borahit* 'make eat', *logahit* 'make read', etc. and to derived causatives. Derived causatives are referred to by the unambiguous causative tag *Der/Caus* in the combination with base verbs with an animate AGENT, i.e. members of the set *AG-NOM-ANI-V*, e.g. *bargat* 'work', *borrat* 'eat', *lohkat* 'read', etc. That way it is ensured that the subject of a non-causative verb and the object of a derived causative of the same base verb have the same role.

```
SUBSTITUTE (V) (V <AG-Acc-Ani>) TARGET AG-ACC-ANI-V OR
AG-NOM-ANI-V + Der/Caus OR  AG-NOM-ANI-V + Der/h + CAUS-DER/H-V;
```

Adjective rules such as the one below annotate valency tags to predicative forms of unambiguously predicative adjectives. This is done by specifying a negative constraint regarding adjectives that have ambiguous attributive and predicative forms and are listed in a set, i.e. *NOT 0 PRED-ATTR-ADJ*.

```
SUBSTITUTE (A) (A <TH-ahte>) TARGET TH-AHTE-A
(NEGATE 0 Attr LINK NOT 0 PRED-ATTR-ADJ);
```

While *valency.cg3* does not specify verb classes explicitly, there are many verbs with similar valency frame constellations. Verb classes can be inferred automatically from verb sets. However, members of one set do not necessarily need to coincide in other sets. One of those sets is *TH-SO-DE*, including verbs with a THEME, SOURCE and DESTINATION.

```
LIST TH-SO-DE-V = "bajidit" "bálkestit" "bidjalit" "bivdet" "botkkuhit" "bovdet"
"coggalit" "coggat" "čuovvulit" "dájuhit""deavdit" "fárrehit" "fievrridit" "fillet"
"gevret" "girdit" "guoddit" "gurgalit" "geassit" "gevret" "goivet" "hohccalit"
"jávkadit" "jođihit" "jorgalit" "láidestit" "láidet""leiket" "loktet" "máhcahit"
"nahkehit" "nivkalit" "njorrestit" "oaččudit" "oahpistit" "oavnnjildit" "oažžut"
"ofelaštit" "rádjat" "roggat" "sáhtašit" "sáddet" "sirdit" "suhppet" "suohpput"
"váldit" "viežžat" "vikkahit" "vuodjit" "vuolggahit" "vuolidit" ;
```

Its members have several valency tags in common including a valency tag for a THEME and a DESTINATION, i.e. *<TH-Acc-Any><DE-Ill-*Ani>*, and a THEME, a SOURCE and a DESTINATION, i.e. *<TH-Acc-Any><SO-Loc-Any><DE-Ill-Any>*.

3.3 Evaluation

I evaluated *valency.cg3*[17] with regard to both lexicon coverage and corpus coverage (both type and token) on a fully annotated version of *SIKOR*. The results are presented in Table 3.21. 1,718 governors (verbs, nouns, adjectives, and adverbs) have at least one frame. To compare, 2,730 verbs in *Vallex 2.0* are annotated with respect to their valency. In *valency.cg3*, 52.5% of the annotated governors have only one valency tag, 17.35% have two valency tags, 9.37% have three tags, and 20.61% have four or more valency tags. Altogether, there are 4,154 lexicon senses, i.e. different governor + valency tag combinations, cf. 6,460 governors in *Vallex 2.0*, and 414 different valency tags altogether. Table 3.22 gives an overview of the most polysemous governors in *valency.cg3*, i.e. the governors with the most valency tags. The verb *leat* 'be, have', a copula, auxiliary and governing verb, is naturally one of the most polysemous verbs and receives 24 valency tags. Other verbs with many valency tags are mostly motion verbs (*mannat* 'go', *bidjat* 'take', *boahtit* 'come', *čiekčat* 'kick', *vuodjit* 'drive', *časkit* 'hit'), communication verbs (*lohkat* 'read, claim', *dadjat* 'say', *muitalit* 'tell') and transaction verbs (*váldit* 'take', *oažžut* 'get', *addit* 'give').

I also evaluated *valency.cg3* on *SIKOR*. As regards corpus coverage, for practical reasons (i.e. only verbal governors have been annotated systematically), I only analyzed the coverage of verbal governors. Token coverage is 73.18%, meaning that 73.18% of the verb cohorts in the whole corpus are annotated by at least one valency tag or an auxiliary tag (<*Inf*>). Type coverage, on the other hand, is much lower: only 6.61% of all the verb types in the corpus are annotated. This means that the analysis is efficient and the most frequent governors have been annotated.

3.4 Conclusion

The work of this chapter resulted in a valency lexicon and grammar for North Sámi, i.e. *valency.cg3*[18], starting out as a systematic valency annotation of 500 verbs, and covering 1,718 verbs with at least one tag, but 47.5% with more than one tag. 20.61% receive four or more tags. Highly polysemous verbs such as *leat* 'be' receive up to 24 frames. A corpus analysis shows that while only 6.61% of the verb types are covered by a valency analysis, the overall coverage (tokens) is significantly higher, 73.18%. The valency grammar consists of Constraint Grammar rules which map multiple valency tags to specific targets (governors) under certain conditions. The valency grammar is used within automatic morpho-syntactic analysis/disambiguation, semantic role annotation and grammar checking where the valency tags are being directly referred to. While verbs are system-

[17] version r146069 (Accessed 2017-01-05)
[18] (Accessed 2017-01-05)

	valency.cg3	**%**
Lexicon coverage		
governors with at least one tag	1,718	100%
governors with one tag	902	52.5%
governors with two tags	298	17.35%
governors with three tags	161	9.37%
governors with four and more tags	354	20.61%
valency tags	414	-
lexicon senses	4,154	-
Corpus coverage (token)		
cohorts with a verb analysis	6,330,884	100%
verb cohorts with a valency tag	4,632,828	73.18%
Corpus coverage (type)		
cohort types with a verb analysis	20,029	100%
verb cohorts with a valency tag (type)	1,324	6.61%

Table 3.21: Lexicon and corpus coverage of the valency tags in *valency.cg3*

Tags	**Verb**	**Translation**	**Tags**	**Verb**	**Translation**
24	*leat*	be	13	*vuodjit*	drive
21	*mannat*	go	12	*časkit*	hit
19	*bidjat*	put	12	*oažžut*	get
18	*váldit*	take	12	*muitalit*	tell
16	*boahtit*	come	11	*addit*	give
16	*ballat*	fear	11	*vázzit*	walk
15	*lohkat*	read, claim	11	*vuolgit*	leave
14	*šaddat*	become	11	*soahpat*	agree
14	*dadjat*	say	11	*oaidnit*	see
14	*beassat*	get	11	*jáhkkit*	believe
14	*atnit*	use	11	*evttohit*	suggest
13	*čiekčat*	kick	11	*cealkit*	express

Table 3.22: The most valency-rich verbs in *valency.cg3*

atically annotated in *valency.cg3*, nouns, adjectives and adverbs are also sporadically annotated. While governing verbs are annotated with valency tags that specify the arguments' semantic roles, auxiliaries are only annotated with respect to their syntactic potential to appear with infinitive governing verbs. The distinctions between governing verbs and auxiliaries are based on Magga's (1986) valency-related criteria. In the Constraint Grammar formalism, annotations are added in a token-based manner. However, multi-word verbs are also annotated. Here, the valency tag containing a specification of the other multi-word parts is annotated on the verb. The annotation of multi-word verbs proves to be relevant as their valencies can differ from homonymous verbs that are not part of a multi-word verb. The token-based manner of *Constraint Grammar* requires that morphological derivations be taken into account in the valency rules, as derivations and inflectional tags can influence/change the valency potential of the verb.

Valency tags are directly integrated into a series of other Constraint Grammars, which is why their form is use-oriented. The system is flexible, as *valency.cg3* is a separate module and new verbs can be added on the fly. Valency tags refer to three important domains of linguistic analysis (semantic roles, morpho-syntax and selection restrictions), each of which is relevant when it comes to resolving error detection, ambiguities and machine translation issues. However, it is not necessary for all three to be specified in the valency tag. They can be left unspecified or be replaced with a concrete word form in the case of idiomatic constructions. Semantic roles are the identifiers of arguments in their alternative morpho-syntactic variations. The valency tags make use of a set of 24 semantic roles for North Sámi, which is based mostly on Bick's (2007*c*) semantic role set for Constraint Grammar. Semantic roles are typically unique with regard to their governor except for causative AGENTS/EXPERIENCERS, which can be distinguished from the non-causative AGENTS/EXPERIENCERS by their morphological case. Semantic roles are further distinguished from semantic prototype specifications, which are made within the selection restrictions. As opposed to Sammallahti's (2005) roles, semantic role distinctions in *valency.cg3* are not based on the animacy of the argument in question or other arguments of the same governor.

As regards morpho-syntax, valency tags refer to either morphological tags or syntactic labels for finite subclauses or non-finite clauses, i.e. accusative and infinitive constructions. I distinguish between verbs that only govern the infinitive semantically and syntactically require an accusative argument (i.e. "object clauses"), and governors that govern both the accusative and infinitive arguments semantically. In the latter case, the accusative can have various different semantic roles, while the infinitive is a THEME.

Lastly, selection restrictions are specified in the form of affirmed or negated semantic prototypes. With general verbs, they can be left underspecified. Otherwise, they specify prototypical and frequent use, or distinguish between two senses of a verb with an otherwise identical syntactical valency, which can coincide with differences in semantic

role.

In the case of synonymy, polysemy and diathesis alternations (both morphological and non-morphological), verbs receive several valency tags, in the latter case preserving the semantic roles of the arguments. Passives, causatives, reflexives, and reciprocals are described with regard to their *lexc* analysis (i.e. lexicalized lemmata, lemmata and tag combinations, and unambiguous derivational tags), their homonymy with other non-valency-changing derivations, and the realization of their arguments in *SIKOR*. There are interesting divergences between linguistic descriptions and corpus use, especially with regard to derived verbs. These include causatives in various constellations, i.e. with a CAUSATIVE AGENT only, without a causative AGENT and an illative-argument that is interpreted as a BENEFICIARY, etc. In addition, *SIKOR* shows that derived reflexives are used with reflexive pronouns and derived reciprocal verbs are used with reciprocal pronouns. Certain derivations are ambiguous with different argument constellations depending on the derivational variant.

Basic valency rules refer to governors, which can be constrained morphologically, i.e. with regard to their derivation or inflection. Rules specifying accusative arguments are typically restricted to non-passive and non-reflexive verbs by reference to the derivational tags. Adjective rules, on the other hand, are restricted to predicative forms of the adjective. However, valency rules can also be constrained syntactically and perform word sense disambiguation. When distinguishing a passive from a frequentative, a constraint to the rule can search for an animate noun in illative case. The valency annotation grammar is therefore not only a lexical database, but also a powerful grammar. As morphological derivations and syntactic/semantic valency changes do not necessarily coincide (i.e. *Der/halla*-tags are used for both passives and frequentatives), valency-wise coherent verbs are stored in sets, which are then used as targets for valency rules. These sets of governors naturally form verb classes, showing syntactic and semantic similarities and making it easy to uncover incoherences. However, they do not form syntactically-semantically coherent classes where one class-membership can directly be deduced from another class-membership.

The valency lexicon + grammar *valency.cg3* is a potent tool, that can be adapted to several applications in the future. The grammar can be extended by means of syntactic rules performing verbs sense disambiguation, which is relevant for machine translation as different translation equivalents typically coincide with valency differences. This can also require further specifications of selection restrictions and subject roles in the valency tags. As the coverage of verb types is only 6.61%, more verbs should be investigated systematically, thereby enabling better governor-argument matching, semantic role annotation, and improving grammar checking. As normative issues are discussed in further detail, e.g. with regard to grammar checking, cf. Chapter 5, *valency.cg3* can be extended to distinguish between normative and non-normative valencies.

In Chapter 4, I deal with semantic prototypes, which in addition to valencies are necessary to identify governor-argument relations in semantic role annotation, grammar checking and machine translation.

Chapter 4

Semantic prototype annotation

> Del rigor en la ciencia
>
> En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el Mapa de una sola Provincia ocupaba toda una Ciudad, y el Mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisfacieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el Tamaño del Imperio y coincidía puntualmente con él. Menos Adictas al Estudio de la Cartografía, las Generaciones Siguientes entendieron que ese dilatado Mapa era Inútil y no sin Impiedad lo entregaron a las Inclemencias del Sol y los Inviernos. En los Desiertos del Oeste perduran despedazadas Ruinas del Mapa, habitadas por Animales y por Mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas. (Borges, 1960)[1]

As the previous short story illustrates, the usefulness of a map lies in its generalization rather than its exact representation of every single detail. The same applies to a map of the semantic 'landscape' of a language. The semantic analysis of a language requires a careful choice of semantic categories, their granularity and their distinctions based on the objectives and their use.

This chapter deals with the semantic annotation of the North Sámi noun lexicon *nouns.lexc*, which, in addition to the valency annotation of potential governors (cf. Chapter 3), is one of the prerequisites for fully exploiting the information given in valency tags. Apart from semantic roles and morpho-syntactic specifications, valency tags specify semantic selection restrictions. Semantic roles can then be found automatically by means of morpho-syntactic specifications and selection restrictions. While grammar checking in *GoDivvun* builds on an existing morphological analysis in the North Sámi infrastructure

[1]"On Exactitude in Science . . . In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography." (Borges, 1999)

for syntactic analysis (*Giella-sme*), a semantic annotation is not available apart from sporadic semantic tags for female, male, surname, organization and place names in the proper noun lexicon *propernouns.lexc*. In addition, semantic sets in the constraint grammars for syntactic analysis and morpho-syntactic disambiguation group some (but not all) nouns semantically. As part of this dissertation, I annotate the North Sámi lexicon systematically by means of semantic prototype tags. The semantic prototype tags are related to each other in a hierarchy, which aspires to draw a complete semantic map of the world. Semantic annotation is needed for a deep syntactic analysis and higher level natural language processing, i.e. semantic role annotation, grammar checking and machine translation. The first section of this chapter is dedicated to the theoretical background of semantic annotation, focusing on distinction vs. definition in semantic analysis, the use of semantic annotation and its role in natural language processing. The second section describes the semantic prototype annotation of the North Sámi lexicon. It deals with the syntactic relevance of semantic categories in North Sámi, the development of semantic primitives, their hierarchical organization and semantic prototype categories with their central and peripheral members. A number of syntactic tests for testing prototype membership of central members are presented. Lastly, lexicon-related issues such as category membership of compounds and multiple membership in the case of polysemy and homonymy are discussed. The third section provides a quantitative evaluation of the lexicon and corpus coverage and a qualitative evaluation of the distribution of semantic prototype categories related to morpho-syntax in four test cases.

4.1 Background

In this section, I will discuss different approaches to the semantic categorization of predominantly nouns, both those that focus on a definition of the concepts and those that do not. I will then present a prototype approach that will be used for categorizing the North Sámi lexicon. Lastly, I will give a short overview over the use of semantic categorization in natural language processing relevant to this work.

4.1.1 Theoretical background

A semantic analysis typically assumes that words can be decomposed into meaning components or that they semantically refer to certain non-decomposable units or groupings of meanings. According to Wierzbicka (1996, p.148), there are two main branches of lexical semantics, the classical (Aristotelean) approach and the prototype approach. Wierzbicka (1996, p.237) applies the classical approach, i.e. her semantic analysis consists of a definition of the meaning of a word by means of culture-independent semantic primes:

> By "defining" a word, then, I mean, essentially, what Locke meant: "show-

> ing" the meaning of a definable (i.e. semantically complex) word in terms of indefinable (i.e. semantically simple) ones.

"Semantic primes" are simple, intuitive, and non-decomposable and describe complex and decomposable concepts in a universal and non-circular manner. According to Wierzbicka (1996, p.237), definitions are necessary "as a tool for understanding other cultures (and for making ourselves understood)".

Wierzbicka (1996, p.73) uses 55 irreducible "semantic primes" across all parts of speech, e.g.: SOME, MORE, SEE, HEAR, MOVE, THERE IS, (BE) ALIVE, FAR, NEAR, SIDE, INSIDE, HERE, A LONG TIME, A SHORT TIME, NOW, IF [...] WOULD, CAN, MAYBE, WANT, WORD. Consequently, 'mother' is defined as:

> X is Y's mother. =
> (a) at one time, before now, [Y] was very small
> (b) at that time, Y was inside X
> (c) at that time, Y was like a part of X
> (d) because of this, people can think something like this about X:
> "X wants to do good things for Y
> X doesn't want bad things to happen to Y"
> (Wierzbicka, 1996, p.155)

While Wierzbicka's (1996) semantic primes are defining, they are not necessarily syntactically relevant or valency-relevant, e.g. "alive" is syntactically relevant, but "can, maybe, want" are not. They can certainly be used in a non-circular definition of, for example, a culturally specific concept like the Japanese noun *amae*, which "is the noun form of *amaeru*" (Wierzbicka, 1996, p.238). However, Wierzbicka's (1996) definition of "mother" is not useful for a valency analysis, where humanness and possibly family-relationship are the most relevant features for identifying, for example, "mother" as the subject of specific verbs that require human subjects.

Lexical semantic descriptions in natural language processing tools are not necessarily meant as definitions. Lexical semantic descriptions in the Swedish machine-readable lexical resource *SALDO*, for example, "are not intended as definitions, but as loose – but hopefully accurate and useful – semantic characterizations of lexical entries" (Borin and Forsberg, 2009). Furthermore, the rule-based system *PALAVRAS* for Portuguese sentence analysis "does not claim to understand text, but only to structure or translate it, the final semantics lies (only) in the eye of the beholder" (Bick, 2000, p.365). Semantic categories in *PALAVRAS* are prototype categories that "draw distinction lines across the semantic landscape [...] by prototype similarity" and ask "is it more like A or more like B? rather than [...] asking Is it an A? or Is it a B?" (Bick, 2000, p.365).

In a syntactic analysis, syntactically relevant distinctions rather than full definitions are relevant in a semantic annotation. Therefore, semantic annotation within *Giella-sme*

focuses on semantic distinctions like those made in *PALAVRAS* rather than full definitions like those used by Wierzbicka (1996). I will therefore introduce semantic prototype categories for *Giella-sme*, similar to Bick's (2000) prototype categories for *PALAVRAS*.

The term "prototype" is taken from cognitive science, originally from a number of studies conducted by Rosch (1973), based on the assumption that humans distinguish between "things" in the world and classify them in a principled manner. Instead of assuming necessary and sufficient conditions for category membership like in the classical Aristotelean approach, prototype categories are composed of central members with a number of distinctive rather than defining features, some of which are shared by less central or peripheral members, and others that are not. Prototype category membership is graded with central and peripheral members and possible overlaps. Lakoff and Johnson (1980) assume that our conceptual system is to a great extent metaphorical and introduce prototypes also for abstract concepts. Lakoff (1987) further states that semantic categorization is based on the way humans conceptualize things depending on their cultural, social and political background, i.e. on cognitive knowledge as opposed to world knowledge. Bick (2000, p.365), on the other hand, considers world knowledge-based "'real' (i.e. not primarily syntactic) semantic classes" necessary in machine translation and artificial intelligence. However, he uses syntactically motivated structuring principles. As world knowledge and linguistic knowledge are hard to separate, in the discussion of semantic categorization that follows I will not distinguish between them.

4.1.2 Semantic categories in natural language processing

Higher-level natural language processing tools for deep syntactic and semantic parsing presuppose good lexical resources.

> Since human language is intertwined with human intelligence and human knowledge, full semantic analysis will not work without a certain degree of artificial intelligence and a huge bank of 'knowledge about the world'.
> Bick (2000, p.363)

A modified version of this bank of knowledge can either be accessed as a separate module or be directly included in the syntactic tool. Full-fledged lexical semantic databases are often developed separately of the syntactic tool and can later be accessed by it. Alternatively, semantic categories can be directly included in syntactic grammars either in the form of semantic sets or in lexica specifically developed for them. Below, I will describe the use of semantic categories in two machine-readable lexical resources and two syntactic tools relevant to this work, cf. Table 4.1.

While lexical semantic resources like *SALDO*[2] (Borin et al., 2008) (Swedish) and *Princeton WordNet* (English) (Miller, 1995) are large-scale ontologies in their own right,

[2] `http://spraakbanken.gu.se/eng/resource/saldo` (Accessed 2017-02-06)

	Entries	Semantic primitives	General categories	Parts of speech
Lexical resources				
SALDO (Swedish)	137,130	43 semantic primitives	39,384 primaries	all covered (35 categories)
WordNet 3.0 (English)	155,287	11 unique beginners	117,659 synsets	nouns, verbs, adjectives, adverbs
Syntactic tools				
PALAVRAS (Portuguese) (Bick, 2006*b*)	36,771	16 atomic features	∼160 prototype categories	nouns and proper nouns
XUXENg (Basque)	?	5 semantic features	5	nouns

Table 4.1: A comparison of semantic categories in *SALDO*, *WordNet 3.0*, *PALAVRAS* and *XUXENg*

syntactic tools do not necessarily need an elaborate system of categories and relations, but can perform well with a smaller set of semantic categories that serve the purpose of the specific task. Typically, syntactic tools and lexical resources do not only differ in the amount of semantic categories they use, but also in the lexicon coverage of their annotation. Independent lexical resources such as *WordNet* and *SALDO* categorize across parts of speech. While *WordNet* annotates nouns, verbs, adjectives and adverbs (but not function words), *SALDO* annotates as many as 35 different parts of speech, including function words. Syntactic tools like *PALAVRAS* and *XUXENg*, on the other hand, only use semantic categories for nouns.

While *SALDO* uses 43 semantic primitives to semantically classify its words (Borin et al., 2013, p.1197),[3] *WordNet* has 26 "unique beginners" for nouns, 15 for verbs, three for adjectives and one for adverbs.[4] These result in 117,659 general semantic categories, "synsets". *SALDO*, on the other hand, does not add labels to its lexical semantic generalizations, but, similarly to a dependency approach, produces 39,384 "primaries". *WordNet 3.0* has 155,287 entries,[5] comparable in size to *SALDO*, which has 137,130 entries.[6] While semantics from *WordNet* is used in word sense disambiguation, cf. Izquierdo Beviá et al. (2007), to my knowledge, there is no documented use of *WordNet* in automatic syntactic

[3]The primitives are: *all* 'all', *annan* 'other', *bara* 'only', *bra* 'good', *fort* 'quickly', *framme* 'in front', *färg* 'color', *för* 'for', *förbi* 'past', *före* 'before', *göra* 'do', *ha* 'have', *hur* 'how', *hända* 'happen', *i* 'in', *ja* 'yes', *just* 'exactly', *ljud* 'sound', *ljus* 'light', *med* 'with', *men* 'but', *mycken* 'a lot', *måste* 'must', *namn* 'name', *natur* 'nature', *när* 'when', *om* 'if', *om* 'about', *på* 'on', *rak* 'straight', *röra* 'move', *säga* 'say', *till* 'to', *tänka* 'think', *vad* 'what', *var* 'where', *vara* 'be', *varm* 'warm', *vem* 'who', *veta* 'know', *vid* 'by', *vilja* 'want', *öppen* 'open'

[4]`https://wordnet.princeton.edu/wordnet/frequently-asked-questions/database/noun.Tops` (Accessed 2017-04-24)

[5]`http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html` (Accessed 2017-04-24)

[6]`https://spraakbanken.gu.se/eng/research/saldo/statistics` (Accessed 2017-04-09)

analysis. *SALDO*, on the other hand, is used in *Sparv*,[7] a pipeline involving syntactic corpus analysis (Borin et al., 2016). In addition to a morphological analysis and dependency analysis, *Sparv* provides a lexical semantic analysis, which is used in word sense disambiguation, cf. Borin et al. (2013, p.1208). Syntactic analysis, on the other hand, is performed by the *MaltParser* (Nivre et al., 2007) and does not apply lexical semantic constraints.

Syntactic tools like *PALAVRAS* and the Basque grammar and style checker *XUXENg* do not use as many semantic categories as the previously described lexical resources. However, the categories that are used are functional with respect to the task that is performed by the tool. Oronoz (2009, p.146–147) mentions five semantic categories that *XUXENg* uses for postposition error detection apart from morpho-syntactic cues. These are *bizidun/bizigabe* 'animate/inanimate', *hizkuntza* 'language', *denbora* 'time', *gaia* 'material' and *gailua* 'device'. Bick (2009*b*) enhances the Portuguese parser *PALAVRAS* with 150–200 semantic prototypes and 16 atomic semantic features for syntactic analysis. The noun lexicon has 36,771 nouns entries with 43,514 senses, cf. Bick (2006*b*). His prototypes are typically used in "unification rules for selection restrictions [e.g. subject-verb unification and passive agent selection restriction], valency instantiation rules and head--dependent association rules" (Bick, 2006*b*). These are also the main uses of semantic categories in *Giella-sme*. Semantic prototype categories in *Giella-sme* are necessary in syntactic analysis and disambiguation, dependency analysis, grammar checking, and machine translation. They are therefore designed in a similar way to those in *PALAVRAS*, i.e. as a medium-sized set of general categories.

4.2 Annotation of the North Sámi lexicon

When designing a system of semantic categories, one needs to decide on the size of the set of semantic categories, the way membership is assigned (e.g. by means of syntactic or semantic tests) and the organization principles of the categories. As semantic categories in *Giella-sme* serve the purpose of improving syntactic analysis, especially in governor-argument matching, there is a particular focus on syntactic (valency) relevance.

[7] `https://spraakbanken.gu.se/sparv/#input=%23editor&lang=en&language=sv` (Accessed 2017-04-24)

4.2.1 Syntactic relevance of semantic categories

This section discusses the syntactic relevance of semantic categories in North Sámi, with examples from previous research, and natural language processing within *Giella-sme*.

In previous Sámi research, semantic categories have been found to be syntactically relevant for the selection preferences of adpositions, in possessive constructions, and for the distinction of senses of polysemous words. In addition, Wiechetek et al. (2010) use semantic prototypes for the lexical selection of translation equivalents in North Sámi to Lule Sámi machine translation in a particular context. While North Sámi *boaris* 'old' translates into Lule Sámi *vuoras* 'old' when modifying nouns that belong to the human or animal prototype category, it translates into *boares* 'old' elsewhere. The syntactic context, i.e. the attributive position of the form, is the same in both cases. However, the contexts differ semantically and can be referred to by means of semantic tags. The Constraint Grammar machine translation rule below selects (Lule Sámi) *boares* for attributive (North Sámi) *boaris* whenever there is a noun to its right that is not a member of the human or animal prototype category. Between *boaris* and the noun there should not be anything other than another adjective in attributive form.

```
SUBSTITUTE (A S0)(A S1) ("boaris"ri A Attr)(*1 N BARRIER NOT-Attr
LINK NOT 0 HUMAN OR ANIMAL);
```

Wiechetek (2012), on the other hand, uses semantic categories in case error detection within adpositional phrases. Many adpositions, such as *sisa* 'inside', are ambiguous. They have both an adposition and an adverb reading. Typically, a genitive noun to their left or right is taken as a reliable constraint to select the adposition reading. However, when checking for case errors, the otherwise reliable cue is misleading. Therefore, error detection rules make use of semantic cues in addition to morpho-syntactic context. The adverbial reading of *sisa* 'inside' is chosen when a noun of the building prototype in illative case appears to the right of *sisa*, to name an example (Wiechetek, 2012, p.38).

Antonsen et al. (2012) thoroughly analyzed the adpositions *miehtá* 'along, more than', *čađa* 'through', *rastá* 'through' and *maŋŋel* 'after', in pre- vs. postpositional use and their preferences with regard to appearing in expressions of extension, time, and movement. *Miehtá* 'along, more than', for example, tends to be used as a postposition mostly with nouns that are time expressions, and as a preposition mostly with nouns that are place expressions.

In their empirical study of synthetic and analytical adnominal possessive constructions – the latter including the reflexive pronoun *ieš* 'one's (own)' – Antonsen and Janda (2015) find that their distribution coincides with certain semantic preferences, both with regard to the possessor and the possessed. While the possessor belongs to the human prototype in 96% of the synthetic possessive constructions, it belongs to the human prototype in 97%

of the analytical constructions. The possessed, on the other hand, are 66% (synthetic) and 37% (analytical) nouns of the semantic categories 'relative', 'body', and 'owned'.

In natural language processing, semantic categories are also relevant in syntactic analysis and disambiguation, and semantic role labeling. In automatic semantic role annotation, the locative nouns *stális* 'out of steel' and *beavddis* 'on the table' in the text book examples (1-a) and (1-c) can be labeled with different semantic roles based on their semantic prototype membership. The noun *stális* 'out of steel' in ex. (1-a) is labeled with the ATTRIBUTE-role based on its membership of the material prototype category. Typically, these sentences contain verbs like *ráhkadit* 'make' or *snihkket* 'craft', which can be missing in elliptical constructions as in the corpus example (1-b). The locative *beavddis* 'at/on the table' in ex. (1-c) and in the corpus example (1-d), on the other hand, is labeled as a LOCATION based on its membership of the furniture prototype.

(1) a. Niibi lea **stális**.
knife is steel.LOC
'The knife is made of steel.' (Nickel and Sammallahti, 2011, p.239)

b. ... ráhkada niibbii[d], guvssiid ja náhpiid, dahje buot mii lea
... makes knifes, cups and milk.cups, or everything which is
muoras ja **čoarvvis**.
wood.LOC and horn.LOC
'... makes knifes, cups and milk cups, or everything that is made of wood and horn.'

c. Niibi lea **beavddis**.
knife is table.LOC
'The knife is on the table.' [L.W., p.k.]

d. Na miihan álgga[h]at go áhččige lea **beavddis** ...
well, we.certainly begin when father.also is table.LOC ...
'Well, we certainly begin when Dad is at the table ...'

The Constraint Grammar rules below map the ATTRIBUTE-role onto a noun of the material prototype, as in *stális* 'steel (Loc.)' in ex. (1-a) and *muoras* 'wood (Loc.)' and *čoarvvis* 'horn (Loc.)' in ex. (1-b). In a first step, the function *SETCHILD* associates the verb with the valency tag *<AT-Loc-Mat>* with its argument in locative case and the material prototype to its right. In a second step, the daughter of a verb with the valency *<AT-Loc-Mat>* receives the semantic role label *§AT* (ATTRIBUTE) if it is in locative case and a member of the material prototype category (*Sem/Mat*).

```
SETCHILD (V <AT-Loc-Mat>) TO (1 Sem/Mat + (Sg Loc)) ;

SUBSTITUTE N (§AT N) TARGET N IF (p (V <AT-Loc-Mat>)) (0 Sem/Mat + (Sg Loc));
```

The simplified Constraint Grammar rule for syntactic mapping below maps the subject predicative label *@<SPRED* onto a noun in locative case after a finite copula, *COPULAS*

+ *FMAINV*, if the locative is of the material prototype category, *Sem/Mat*.

```
MAP (@<SPRED) TARGET (N Loc) + Sem/Mat IF (*-1 COPULAS + FMAINV BARRIER NOT-ADV-PCLE ;
```

The rule is used in regular parsing, *functions.cg3*,[8] and disambiguation that is used in grammar checking, i.e. in *disambiguator.cg3*.[9] Especially when dealing with syntactically unreliable input, as in grammar checking, analyzing syntax only on the basis of syntactic constraints can be misleading. As there can be syntactic errors in the sentences, lexical semantic annotation can help to identify the intended syntax despite its morpho-syntactic error.

4.2.2 Semantic primitives as structuring principles

Semantic categories in *Giella-sme* are not randomly chosen, but are related to each other by means of certain structuring principles. Structuring principles not only ensure the completeness of a system, i.e. provide a semantic place for every lexeme, but often produce the semantic categories themselves. From a rule-developing perspective, it can be very tempting to create a new tag on the fly as soon as a rule can be applied to more than one word. From the standpoint of generalization, however, it is not desirable to have a collection of random tags that can only be used for one (marginal) rule. Therefore, semantic prototypes in *Giella-sme* are kept at a fairly general level, whereas *WordNet* has many singleton synsets as pointed out by Borin and Forsberg (2009, p.1193).

The lexical database *WordNet* uses classical lexical semantic structures, i.e. synonymy, antonymy, hyponymy, hyperonymy, meronymy, holonymy, troponymy, and entailment. For nouns, *WordNet* uses 25 "unique beginners", cf. Figure 4.1, that form separate hierarchies by means of hyperonym relations. According to Miller (1990, p.16), "the features that characterize a unique beginner are inherited by all of its hyponyms". However, these features are not explicitly stated in *WordNet*.

In the lexical semantic resource *SALDO*, a lexeme is described by one or several descriptors. The lexeme's main descriptor is "a semantically closely related entry which is more central, i.e., semantically and/or morphologically less complex, probably more frequent, stylistically more unmarked and acquired earlier in first and second language acquisition, etc." (Borin and Forsberg, 2009). Central members of *SALDO* are frequent words, stylistically simple (i.e. described by few semantic primitives) and also morphologically simple. Unlike in *WordNet*, the descriptors forming the hierarchies are organized by associations, which can be, but do not need to be, hyperonyms (Borin et al., 2013, p.1192). Instead, they can also be antonyms, or synonyms.

[8] version r151846 (Accessed 2017-04-27)
[9] version r151804 (Accessed 2017-04-26)

{*act, action, activity*}
{*animal, fauna*}
{*artifact*}
{*attribute, property*}
{*body, corpus*}
{*cognition, knowledge*}
{*communication*}
{*event, happening*}
{*feeling, emotion*}
{*food*}
{*group, collection*}
{*location, place*}
{*motive*}
{*natural object*}
{*natural phenomenon*}
{*person, human being*}
{*plant, flora*}
{*possession*}
{*process*}
{*quantity, amount*}
{*relation*}
{*shape*}
{*state, condition*}
{*substance*}
{*time*}

Figure 4.1: The 25 unique beginners for *WordNet* nouns (Miller, 1990, p.16)

In the syntactic parser *PALAVRAS*, Bick (2000, p.372) organizes semantic prototypes into one single hierarchy based on the hyperonym relation and 14 syntactically relevant semantic primitives, which are binary "atomic features", cf. Figure 4.2. Prototypes are fuzzy categories without necessary and sufficient conditions. Peripheral members of a category therefore do not need to inherit all semantic features of their mother. Bick's (2000) 22 general semantic prototype classes further split into 150–200 prototype tags without being structured by explicit atomic features.[10]

The structure of the hierarchy for North Sámi nominal prototype categories in *Giella-sme*, shown in Figure 4.3, is mainly based on Bick's (2000) hierarchy. It uses a number of the syntactically relevant atomic features of Bick's (2000) hierarchy, discarding some of them and adding a few new ones based on the syntactic peculiarities of North Sámi. Prototypes, on the other hand, are specific to North Sámi and differ from Bick's (2000) set. The hierarchy is also used to construct more general sets including several prototype categories in the syntactic file itself, e.g. *ANIMATE* or *CONCRETE*.

As in Bick's (2000) hierarchy, nouns are primarily split into concrete and abstract nouns. While the left, i.e. concrete, branch of Bick's (2000) hierarchy is directly applied to nouns in *Giella-sme*, the structure of the non-concrete side is adapted to syntactically relevant features in North Sámi. In particular systematic case homonymies like comitative singular/locative plural, genitive/accusative, cf. Trosterud and Wiechetek (2007), and case homonymies where possessive suffixes are involved are taken into account when choosing semantic structuring principles. In order to resolve these case ambiguities, one needs to distinguish between potential objects (accusative vs. genitive), potential posses-

[10] `https://gramtrans.com/deepdict/semantic_prototypes_overview.pdf` (Accessed 2017-02-06)

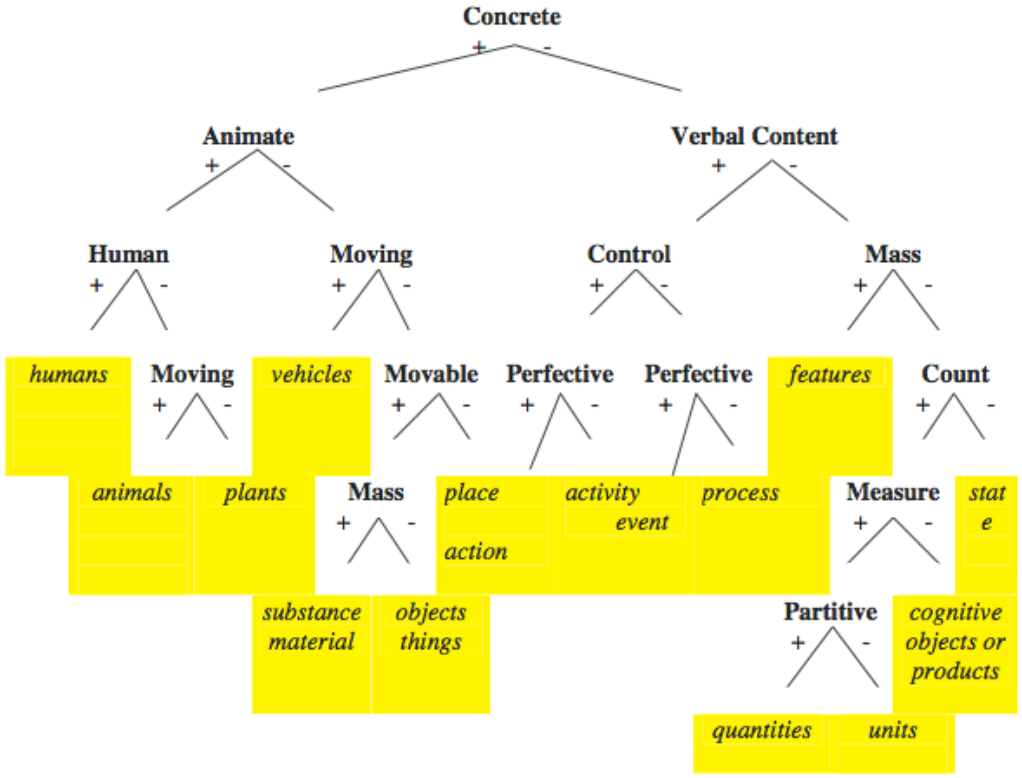

Figure 4.2: Hierarchy of semantic prototypes within *PALAVRAS* (Bick, 2000, p.372)

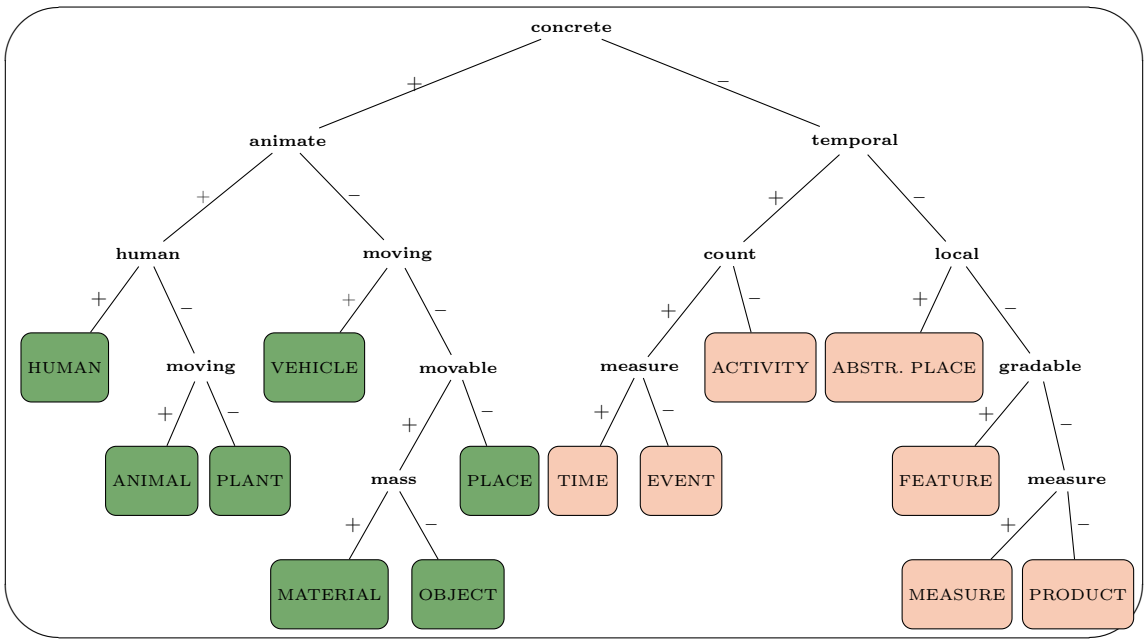

Figure 4.3: The hierarchy of semantic prototypes in *Giella-sme*

sors (genitive vs. accusative), whole-part relations (genitive vs. accusative and locative vs. comitative), potential instruments (comitative vs. locative) and the potential local, temporal or social space of an event (locative vs. comitative).

The leftmost branch of the hierarchy represents the most central category in human language, the human prototype, which is distinguished from animals, plants, other moving items (vehicles) and places (least similar to humans from all concrete categories). The categorization on the non-concrete branch of the hierarchy pays attention to case distinctions in North Sámi. The conceptualization of abstract categories often relates to concrete categories by means of metaphors. Some nouns are conceptualized as locations of actions similar to concrete places, while others are conceptualized as objects, instruments or actions. Abstract concepts are primarily split into concepts with and without a temporal dimension, cf. Bick's (2000) 'verbal content' feature. In North Sámi, these nouns can typically be used in temporal expressions, e.g. with temporal adpositions. Neither "control" nor "perfective" are particularly relevant in *Giella-sme*. Typically, nouns derived from verbs are categorized as 'actions/activities'. Those are distinguished from nouns of the time prototype category, which denote a certain time period and can function as units. Events also denote a certain time period, but are not used as units.

Non-temporal concepts are further split into local and non-local concepts. Local concepts, e.g. members of the domain prototype, are often conceptualized as places. Non-local concepts are split into further categories by the features gradable and measure into feature, measure and product prototypes. Product-prototypes are conceptualized as objects, e.g. visual products of actions or semantic concepts.

4.2.3 Semantic prototypes

The hierarchy in Figure 4.3 represents a general structure and contains annotation guidelines for 14 general semantic prototypes in *Giella-sme*, which are further divided into many more specific prototypes. The prototype tagset aspires to cover the full lexicon of North Sámi. Unlike lexical semantic resources such as *WordNet* and *SALDO*, *Giella-sme* is a predominantly syntactic resource, and semantic prototypes are used mostly to facilitate syntactic analysis rather than being part of an independent full-fledged lexical resource. Prototypes are tagged directly in the lexica for nouns (*nouns.lexc*), proper nouns (*propernouns.lexc*), and adjectives (*adjectives.lexc*).

As opposed to lexical resources like *WordNet* and *SALDO*, the basic unit of the lexica in *Giella-sme* is not the word meaning, but the lemma. Semantic prototypes tags are predominantly added to nouns. However, adjectives that can assume the syntactic functions as (human) nouns are also tagged. Adverbs that can assume the same function as specific nouns in argument positions of a verb are tagged as place or time prototype category members.

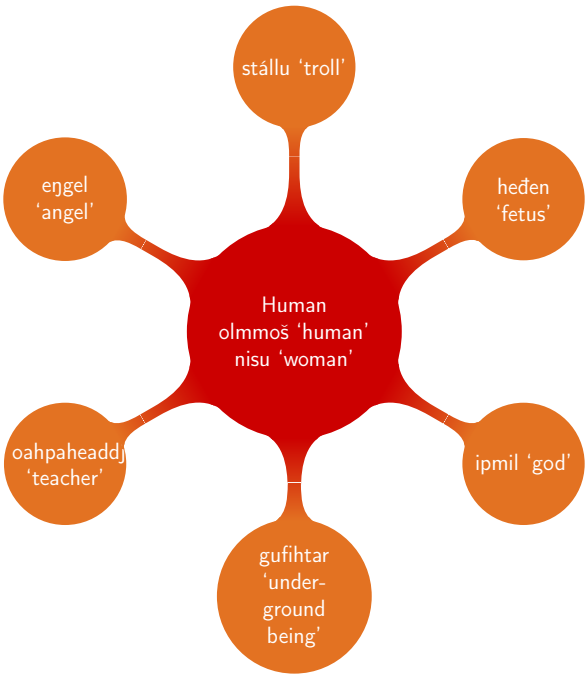

Figure 4.4: Central and peripheral members of the human prototype category

Below, I will analyze the central and peripheral members of two prototype categories, i.e. human and vehicle, by means of syntactic tests that are characteristic of each category's defining atomic feature. Figure 4.4 shows the human prototype category with central members *olmmoš* 'human' and *nisu* 'woman' and a number of less central members such as *oahpaheaddji* 'teacher', *gufihtar* 'underground being', *ipmil* 'god', *heđen* 'fetus', *stállu* 'troll', and *eŋgel* 'angel'.

While *gufihtar* '(mythical) underground being', and *stállu* 'troll' are human-like beings appearing in Sámi narratives, *ipmil* 'god' and *eŋgel* 'angel' are religious or spiritual beings thought of as superior to humans. They are not humans, but human-like. *Stállu* 'troll' for example can have *beallestállu* 'half-troll' offspring with humans, cf. ex. (2-a). Ex. (2-b) shows that some human and human-like concepts are not far from each other, and a man (*olmmái*) can (metaphorically) be a giant (*jiehtanas*). Also ownership of e.g. a reindeer herd is not only a typical attribute of a human, but can also be used for a *gufihtar* 'underground being', cf. ex. (2-c).

(2) a. **Beallestál[l]u** - *gean* namma lei Mikkel, ...
half-troll - who.GEN name was Mikkel, ...
'Half-troll - whose name was Mikkel, ...'

b. Boares olmmái lei **jiehtanas**, *gii* lávii olbmuid borrat.
old man was giant, who used.to humans eat
'The old man was a giant, who used to eat humans.'

c. Biera diđii ahte die lei **gufihttara** *eallu*.
Biera knew that that was underground.being.GEN herd
'Biera knew that that was the underground being's herd.'

Concepts like *eŋgel* 'angel' have typical human attributes, i.e. intellect, speech, emotions, and morals, expressed by the verbs *diehtit* 'know' and *cealkit* 'say' in ex. (3-a)–(3-b). The noun *heđen* 'fetus', on the other hand, denotes a human that is technically not yet developed enough to live on its own. Therefore, it is generally not conceptualized as a human, comparable to *eŋgel* 'angel'. Whereas it does not appear as the subject of verbal communication or emotion, it can assume the habitive function with subjects that denote human illnesses, cf. ex. (3-c).

(3) a. *Dihtet* go **eŋgelat** ahte moai bohte, áhčči?
know Q angel.PL.NOM that we come, Dad
'Do the angels know that we come, Dad?'

b. Muhto **eŋgel** *celkkii* sidjiide: Allet bala!
but angel said they.ILL: Don't fear
'But the angel said to them: do not be afraid!'

c. ...**heđemis** lea *Downs syndroma*.
...fetus.LOC has Down syndrome.ACC
'...the fetus has Down syndrome.'

One indication for humanness is the antecedent position of the relative pronoun *gii* 'who' based on the assumption that *gii* 'who' is a strong grammatical marker for the humanness of the antecedent. I investigated the use of the nouns in question as antecedents of *gii* 'who' as opposed to *mii* 'that', which can be used with both non-human and human antecedents (Nickel, 1994, p.123). Concrete persons that are referred to by their names should, according to Nickel and Sammallahti (2011, p.125), preferably be modified by relative clauses with *gii* 'who'. Table 4.2 shows the distribution of *mii* 'which, that' and *gii* 'who' with antecedents of the human prototype category and two other nouns of the animal prototype, i.e. *beana* 'dog' and *boazu* 'reindeer', as a reference group. While central members such as *olmmoš* 'human', *nisu* 'woman', and also *oahpaheaddji* 'teacher' are frequent in *SIKOR*, i.e. between 10,000 and 80,000 occurrences, less central and peripheral members have less than 1,000 occurrences. The low frequency of some of the peripheral members, i.e. less than 10 occurrences, makes it hard to study their syntactic properties, as the context of the syntactic tests may not even be represented in any example.

Central members of the human prototype are highly represented by relative clauses with *gii* 'who' (91–96%). *Eŋgel* 'angel', although not a human, ranges in the same percentage as prototypical humans. Surprisingly, *ipmil* 'god' does not. Some of the infrequent peripheral nouns, like *beallestállu* 'half-troll', *Mummistállu* 'Moomin', and *muohtastállu* 'snowman' are found with *gii* 'who' in 100% of of the relative clauses examined. However, these nouns are all represented by one or two examples only.

While *Mummistállu* 'Moomin' is a fictional character acting like a human and resembling a hippopotamus, *muohtastállu* 'snowman' is a sculpture rather than an acting

Noun	**Total**	***mii***	***gii***	% (***gii***)
Central members				
olmmoš ‘human’	77,202	359	7,756	95.58
nisu ‘woman’	8,024	31	404	92.87
oahpaheaddji ‘teacher’	14,854	67	717	91.45
Peripheral members				
ipmil ‘god’	1,269	16	8	33.33
eŋgel ‘angel’	450	2	23	92.00
jiehtanas ‘giant’	402	5	3	37.50
ulda ‘underground being, guardian spirit’	214	9	4	30.77
háldi ‘underground being, guardian spirit’	144	2	0	0.00
gufihtar ‘underground being’	120	3	2	40.00
heđen ‘fetus’	6	-	-	-
stállu ‘troll’	786	7	1	12.50
juovlastállu ‘Santa Claus’	373	6	5	45.45
Mummistállu ‘Moomin’	10	-	2	100.00
muohtastállu ‘snowman’	9	-	1	100.00
beallestállu ‘half-troll’	6	-	1	100.00
Non-members				
boazu ‘reindeer’	17,019	767	-	0.00
beana ‘dog’	3,948	174	2	1.14

Table 4.2: The distribution of members of the human prototype category as antecedents of relative subclauses

being, which is why the preference of *gii* 'who' is unexpected. Between 12 and 45% of the other peripheral human prototype category members (except for *ulda* 'underground being, guardian spirit') appear with relative clauses introduced by *gii* 'who'. Peripheral members of the human prototype category, like *gufihtar* 'underground being', are modified by relative clauses with both *gii* 'who', cf. ex. (4-a), and a form of *mii* 'that', cf. ex. (4-b). The results are not quite representative with only one example on each side.

(4) a. Mun de[dd]ohalan gabba bohccuin ja
I have.nightmares white reindeer.LOC.PL and
gufihttariin **geat** isket mu rábadit ...
underground.being.LOC.PL who.NOM.PL try snatch me ...
'I have nightmares of white reindeer and underground beings who try to snatch me away ...'

b. ... muittuhussan gufihttariin, **mat** ásse das ...
... reminder.ESS underground.being.LOC.PL, that.NOM.PL lived here ...
'... as a reminder of the underground beings, that lived here ...'

Unfortunately, there are only six occurrences of *heđen* 'fetus' in *SIKOR*, none of which is followed by a relative pronoun. Central members of the animal prototype, like *boazu* 'reindeer' and *beana* 'dog', on the other hand, are only antecedents of *gii* 'who' in less than 2% of the occurrences and strongly prefer relative clauses introduced by *mii* 'that'.

Figure 4.5 shows the vehicle prototype category with the central members like *biila* 'car', *fanas* 'boat' and *sihkkel/syhkkel/sykkel* 'bike'. All of these vehicles have a human driver and are moved by either a motor or human power. *Gielká* 'sled', on the other hand, is ambiguous between a motor-driven vehicle, cf. ex. (5-a), and a sled that moves by itself on an inclined surface/slope, cf. ex. (5-b). *Heargi* 'draft reindeer' is a peripheral member of the vehicle prototype category because it is an animal. However, as can be seen in ex. (5-c), *herggiin* 'draft reindeer (Com.)' is used parallel to *skohteriin* 'scooter (Com.)' as a vehicle by means of which the journey is performed.

(5) a. ... **gielkkás** lei fárta jobe badjel 120 kilomehtera diimmus.
... scooter.LOC had speed even over 120 km hour.LOC
'... the scooter reached speeds of over 120 km per hour.'

b. ollu muohta boahtá ja beasan čierastit **gielkkáin** fas
much snow comes and get.to sled sled.COM again
'a lot of snow is coming and I get to go sledding again'

c. Mátki mainna **herggiin** ádjánedje guokte beaivvi ovdal
journey which draft.reindeer.COM take.PRT.3PL two days back.then
'The journey that took them two days with a draft reindeer back then'

Vuoján 'vehicle' can denote a vehicle in general, whether motorized or non-motorized, cf. ex. (6-a). In ex. (6-b)–(6-c), both *heargi* 'draft reindeer' and *fanas* 'boat' are explicitly

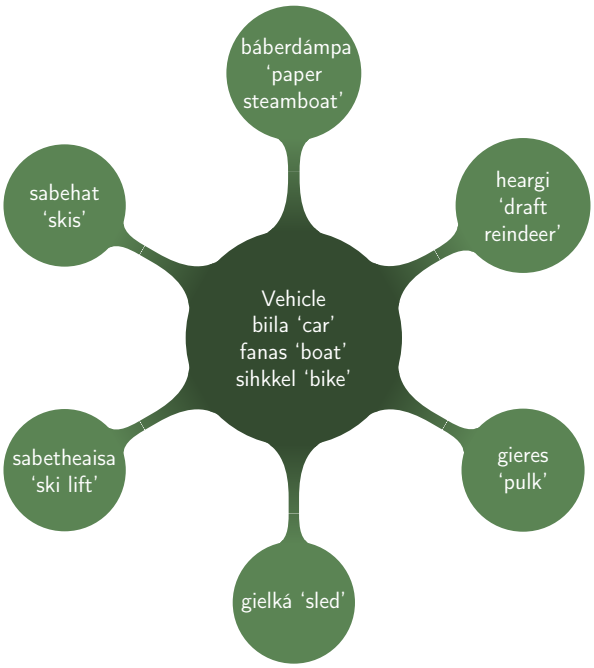

Figure 4.5: Central and peripheral members of the vehicle prototype category

classified as vehicles (*vuoján*). *Vuoján* 'vehicle' can also be used as a synonym for *skohter* 'scooter', cf. ex. (6-d), or *heargi* 'draft reindeer' as in ex. (6-e).

(6) a. ...johtalusa losit **vuojániiguin** nugo biilla ja tráktora
...traffic heavy vehicle.COM.PL like car and tractor
'...traffic with heavy vehicles like cars and tractors'

b. Eira jáhkká ahte dat *heargi* livčči šaddan buorre **vuoján**.
Eira believes that the draft.reindeer could become good vehicle
'Eira believes that the draft reindeer could become a good vehicle.'

c. Geainnuhis gilážis lei **fanas** áidna *vuoján* jus áiggui gosage vuolgit.
streetless village was boat only vehicle if wanted somewhere leave
'In a village without roads, a boat was the only vehicle if one wanted to go somewhere.'

d. Son bártidii maiddái iežas Lynx **vuojániin**.
s/he was.in.an.accident also own Lynx vehicle.COM
'S/he was also in an accident with her/his own Lynx scooter.'

e. Nuppi sadjái bođii son Anne Risten Sara **vuojániin** Gistein.
second place came s/he Anne Risten Sara reindeer.COM Gistein
'S/he came in second with Anne Risten Sara's reindeer Gistein.'

Sabetheaisa 'ski lift' is a peripheral member of the vehicle prototype. Although it has elements that move by electrical power, the construction itself is fixed. However, in ex. (7-a), it is the AGENT of the motion verb *vuolgit* 'leave'. *Sabehat* 'skis', on the other hand, is a peripheral member as skis do not seat a human and as they are a hybrid between clothes, i.e. something you can put on, cf. ex. (7-b), and a vehicle, i.e. something you

can move by.

(7) a. . . . dákko gokko **sabetheaissat** *vulget* vuolimuččas luohkás.
. . . here where ski.lifts leave lowest hill.LOC
'. . . here where the ski lifts leave from the lowest hill.'

b. N[a]hket **sabehiid** juolgái ja čuoigá Gaskkačorrui.
put.IMPRT.2PL ski.ACC.PL feet.ILL and ski.IMPRT.2SG Gaskkačorru.ILL
'Put the skis on your feet and ski to Gaskkačorru.'

The vehicle prototype is characterized by the atomic features -animate and +moving. It typically appears in comitative case with directed motion verbs, prototypically with the verb *vuodjit* 'drive', which has a subject AGENT, a SOURCE-argument, and a DESTINATION-argument, cf. ex. (6-a), (6-d), and (6-e). Members of the vehicle prototype category also appear in the subject position of a volitional directed motion verb, cf. *biila* 'car' in ex. (8-a). I consider the following (amongst others) directed motion verbs: *fievrridit/-fievrredit* 'transport', *suvdit* 'ship', *doalvut* 'bring', *johtit* 'travel', *boahtit* 'come', *vuodjit* 'drive', *vuolgit* 'leave', *ollet* 'reach', *finadit* 'pay a short visit', *leat jođus* 'be on the move', *ruohttit* 'run', *girdit* 'fly', *geasehit* 'transport', *viežžat* 'fetch', and *šuvgat* 'fly'. Both the inanimate *sihkkel* 'bike' in ex. (8-b) and the animate *heargi* 'draft reindeer' in ex. (8-c), appear as a subject of the verb *ruohttit* 'run' and the derived verb *ruohtastit* 'run away', which are typically used with animate subjects. In addition, there are deverbal nouns derived from directed motion verbs such as *johtolat* 'traffic', *fievrredeapmi* 'transport', *mátki* 'journey, trip', *sáhttu* 'ride', and *vuodjin* 'driving'. These are also considered to be indicators of members of the vehicle prototype category, cf. ex. (5-c) and (6-a).

(8) a. . . . go **biila** *bođii* meattá olles leahtuin.
. . . when car came past full speed.COM
'. . . when the car passed at full speed.'

b. Fáhkka *ruohtastii* **sihkkel** nuppeguvlui, eret luottas.
suddenly ran bike other.way, away path.LOC
'Suddenly the bike went the other way, away from the path.'

c. Dat lea vuosttaš **heargi** mii lea *ruohttan* vuollel 15 sekundda . . .
this is first draft.reindeer which has run below 15 seconds . . .
'This is the first draft reindeer that has run under 15 seconds . . . '

Table 4.3 shows the distribution of different members of the vehicle prototype category with regard to two different syntactic tests. Between 10 and 19% of the occurrences of *gielká* 'sled', *biila* 'car', *heargi* 'draft reindeer', *sihkkel* 'bike', *vuoján* 'vehicle' in *SIKOR* appear in comitative case together with a motion verb. For *fanas* 'boat' the percentage is significantly lower (6.5%), suggesting that it is less central. Only between 1 and 4% of the occurrences of *gieres* 'pulk' and *sabehat* 'skis' appear in comitative case together with a motion verb. *Sabetheaisa* 'ski lift' has no occurrence whatsoever in comitative case after a motion verb. At the same time it is not very frequent, i.e. 5 occurrences in total.

Noun	*SIKOR*	Test 1 X.com	%	Test 2 X.nom	%
Central members					
biila 'car'	5,610	790	14.08	141	2.51
fanas 'boat'	4,957	328	6.62	161	3.25
sihkkel 'bike'	678	84	12.39	8	1.18
gielká 'sled'	136	26	19.12	3	2.21
Ambiguous					
heargi 'draft reindeer'	3,122	522	16.72	39	1.25
vuoján 'vehicle'	585	62	10.6	7	1.2
Less central members					
sabehat 'skis'	493	29	5.88	8	1.62
gieres 'pulk'	82	3	3.66	-	-
sabetheaisa 'ski lift'	5	-	-	2	40

Table 4.3: The distribution of members of the vehicle prototype category in sentences with motion verbs

Most of the members can also appear in the subject position together with a motion verb, typically between 1 and 4%. Here, *fanas* 'boat' and *biila* 'car' are most prototypical apart from the infrequent *sabetheaisa* 'ski lift' with 40%.

4.2.4 Syntactic tests for category membership

The noun lexicon *nouns.lexc*[11] includes 78 tags for semantic prototype categories, cf. Table B.1 in Appendix B. These are based on the semantic hierarchy in Figure 4.3 and syntactically relevant categories applied in the constraint grammars within *Giella-sme*. Each prototype category is associated with one or several syntactic tests that are passed by central members of the category, but not necessarily by peripheral members. Syntactic tests mostly test the noun's ability to appear as an argument of certain verbs or in particular adpositional phrases.

4.2.4.1 Testing concrete categories

Concepts are primarily split into concrete and abstract concepts. Typically concrete concepts can be touched, and can be objects of the verb *guoskat* 'touch, concern'. They can also be approached and appear in a postpositional phrase with *lusa* 'to' after motion verbs such as *mannat* 'go'. However, abstract concepts can also be conceptualized as concrete concepts through metaphor, and the verbs themselves can be used metaphorically. Lakoff and Johnson (1980, p.25) elaborate on ontological metaphors:

[11] version r146045 (Accessed 2017-01-05)

> When things are not clearly discrete or bounded, we still categorize them as such, e.g. mountains, street corners, hedges, etc. [...] our experiences with physical objects (especially our own bodies) provide the basis for an extraordinarily wide variety of ontological metaphors.

None of the 3,742 occurrences of *lusa* 'to' in *SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection* (2015-03-01) occurs with a member of the non-concrete branch of the feature hierarchy in Figure 4.3. Of the 15,049 occurrences of the verb *guoskat* 'touch, concern', both concrete and abstract nouns appear in object position. Abstract nouns include members of the language prototype (*sámi giela* 'Sámi language' in ex. (9-a)), text prototype (*ohcamušaide* 'applications' in ex. (9-b)), and activity prototype categories (*nuoraidpolitihkii* 'youth politics' in ex. (9-c)). The verb *guoskat* itself has a concrete ('touch') and an abstract meaning ('concern'), which makes the reliability of the test problematic.

(9) a. Sámegieloahpahus ... guoská **sámi giela** **ja** **kultuvrra**
Sámi.language.teaching ... concerns Sámi language.ACC and culture.ACC
'Sámi language teaching ... concerns Sámi language and culture'

b. Seamma guoská **ohcamušaide** mánáidjoavkkuin.
same concerns application.ILL.PL children.group.LOC.PL
'The same goes for the **applications** from children groups.'

c. Mii guoská **nuoraidpolitihkii**, ...
which concerns youth.politics.ILL, ...
'Concerning youth politics, ...'

Some nouns can be members of prototype categories on both the concrete and abstract branch of the hierarchy. Nouns like *biebmu* 'food' are members of the food prototype category as something that can be eaten (cf. ex. (10-a)) and members of the event prototype category denoting the point in time when food is eaten (cf. ex. (10-b)).

(10) a. Mus lea **iđitbiebmu** fárus.
I.LOC is breakfast with
'I have the breakfast with me.'

b. Liikká bulii njálbmi go máisten dan **biepmu** maŋŋil
nevertheless burned mouth when tasted it meal.GEN after
'Nevertheless, my mouth burned as I tasted it after the meal'

Figure 4.4 shows different types of animate prototype categories. Animate prototypes include humans, animals and plants. Typically, they live, grow and act (plants, though to a lesser extent). However, the verb *šaddat* 'grow, become' is not only used for plants, cf. ex. (11-a). Also abstract nouns, e.g. events (*festivála* 'festival' in ex. (11-b)), and amounts (*lohku* 'number' in ex. (11-c)) can be subjects of these verbs. Typically, plants can blossom (*liedđut*) and be cultivated (*gilvit*). Members of the plant prototype category, such as *duottarrássi* 'tundra flower' shown in ex. (11-e) are not the only nouns

Prototype category	**Description**	**TEST (positive and negative)**
+concrete +animate +human		
Sem/Hum	human	X.NOM čállá reivve 'writes a letter', X, gii 'X, who'
Sem/Fem	female name	X.NOM lea nisson 'X is a woman'
Sem/Mal	male name	X.NOM lea olmmái 'X is a man'
Sem/Sur	surname	X.NOM lea goargu 'X is a surname'
Sem/Org	organization	X.NOM čállá reivve 'X writes a letter'
+concrete +animate -human		
Sem/Ani	animal	X.NOM ruohttá/eallá/suhttá 'X runs, lives, gets angry'
Sem/Plant	plant	X.NOM šaddá/stuorru/lieđđu 'X grows/blossoms', gilvit X.ACC 'plant X'
Sem/Group	group	X.NOM leat.3PL 'X are'

Table 4.4: +concrete +animate semantic tags for North Sámi

that can appear in the subject position of the verb *lieđđut* 'blossom'. Members of abstract prototype categories, such as *vuoigatvuohta* 'right', can also be found there, as shown in ex. (11-d). But again *gilvit* 'cultivate, seed' (1,140 results) can have members of abstract categories as objects, e.g. cognitive products, as in ex. (11-f).

(11) a. . . . ja šattut dárbbašit fotosyntese go galget **šadd[a]t** . . .
. . . and plants need photosynthesis if should grow . . .
'. . . and plants need photosynthesis to grow . . . '

b. **Festivála** lea šaddan hui viiddis . . .
festival has become very wide . . .
'The festival has become huge . . . '

c. . . . prošeaktaohcamiid **lohku** stuorru . . .
. . . project.application.GEN number grows . . .
'. . . the number of project applications grows . . . '

d. **Vuoigatvuođat** lieđđugohtet . . .
rights flourish.INCH.PRS.3PL . . .
'The rights begin to flourish . . . '

e. **Duottarrásit** liđđot alla duoddariin.
tundra.flowers blossom.PRS.3PL high tundra.LOC.PL
'The tundra flowers blossom in the high tundras.'

f. Ulbmil ođasreivviin lea sihke gilvit **máhtu** sámi diliiid birra
goal newsletter.COM is both spread knowledge Sámi issues about
'The goal of the newsletter is both to spread knowledge about Sámi issues'

Members of the human, female, male and surname prototype categories are all end nodes of the +human branch in Figure 4.3. While the human prototype is used in *nouns.lexc* and *adjectives.lexc*, female, male, and surname are only used in *propernouns.lexc*.

They were introduced to the proper noun lexicon[12] in 2005 partly because of homonymies between place-names and surnames in the Norwegian language (e.g. *Trosterud* is both a place and a surname) and their syntactic behavior (e.g. surnames appear after first names and are analyzed as syntactic heads). The general thought was that "they would come in handy later". In North Sámi, few nouns denote female (e.g. *nisu* 'woman') or male humans (e.g. *áhčči* 'father') only – most nouns are gender neutral – whereas, for example, Romance languages make gender distinction in profession expressions. Therefore, female, male, and surname prototypes are only used in the proper noun lexicon.

The organization prototype category belongs to the +human branch as well, as its prototypical members such as e.g. *girku* 'church' can be conceptualized as humans, cf. ex. (12-a). At the same time, its prototypical members represent a building, which can be a LOCATION, cf. ex. (12-b). The group prototype category typically refers to groups of humans or animals. Members of the group and human prototypes, such as *joavku* 'team', can typically be used with a finite verb in plural form even if the noun is a singular form, cf. ex. (12-c). Groups of animals, e.g. *eallu* 'herd' or *spierru* 'fish swarm', only appear with singular verbs, cf. ex. (12-d).

(12) a. **Girku** bargá sámegiela nannemiin ...
church works Sámi strengthening.COM ...
'The church works to promote the Sámi language ...'

b. Mii leat fitnan **girkus**.
we have visited church.LOC
'We have visited the church.'

c. Deanu kulturskuvlla **nuoraidteáhterjoavku** dánsot ...
Tana art.school youth.theater.group.NOM dance.PRS.3PL ...
'The youth theater group from the Tana Art School is dancing ...'

d. ... **eallu** mannagoahtá várrái.
... herd.NOM go.PRS.3SG mountain.ILL
'... the herd starts going to the mountain.'

Concrete inanimate non-moving movable prototype categories are represented in Table 4.5. The concrete inanimate moving vehicle prototype category was discussed in the previous section.

Non-moving but movable concepts are split in into +/-mass. Movable concepts belong to the object prototype category. The object prototype category is a larger category containing many subtypes that take into account use, as in the case of tools or clothes, and certain qualities, e.g. shape. Tools typically appear in comitative case together with certain verbs. Specific 'tools' are used in music, measuring, and hunting (weapons). Clothes, on the other hand, can be accusative objects of verbs of (un)dressing. Members of the food prototype category, on the other hand, can be accusative objects of verbs of

[12] version r4607 of *propernoun-sme-lex.txt* (Accessed 2005-10-05)

Prototype category	**Description**	**TEST (positive and negative)**
-animate, -moving, +movable, -mass		
Sem/Obj	object	sirdit X.ACC 'move X', guoskat X.ACC 'touch X'
Sem/Food	food	borrat/ráhkadit X.ACC 'eat/prepare X'
Sem/Drink	drinkable	juhkat X.ACC 'drink X'
Sem/Clth	clothes	coggat/bidjat X.ACC ala 'put on X'
Sem/Txt	written document	X.LOC čuožžu ahte 'in X, it is written that'
Sem/Tool	for manual work	divvut juoidá X.COM 'repair something with X'
Sem/Wpn	weapon	goddit X.COM 'kill/hunt with X'
Sem/Body	body part	mus lea bávčča X.LOC 'my X hurts'
Sem/Ctain	container	bidjat juoidá X.GEN sisa 'put something inside X'
Sem/Furn	furniture	čohkkedit X.GEN ala 'sit down on X'
-moving, +movable, +mass		
Sem/Mat	disintegrates when penetrated	ráhkadit juoidá X.LOC 'make sth. out of X'
Sem/Substnc	does not disintegrate	mannat X.GEN rastá 'go through X'

Table 4.5: +concrete -animate -moving +movable semantic prototype tags for North Sámi

eating. Furniture, body parts and containers can be LOCATIONS or DESTINATIONS for certain actions. In the case of two possible prototype category memberships, syntactic similarity is used as a criterium. Alternatively, the noun is considered to be a member of both possible prototype categories. The noun *oađđenseahkka* 'sleeping bag' is categorized as a cloth-object based on its similarity to *loavdda* 'blanket, cover' or *govčča* 'blanket' as both can appear in comitative case after the verb *gokčat* 'cover', cf. ex. (13-a)–(13-b). It also resembles *seaŋga* 'bed', which is a piece of furniture. But unlike a bed, it can be packed easily into a backpack and transported, cf. ex. (13-c) where *oađđinseahka* 'sleeping bag' is used with the verb *dohppet (fárrui)* 'grab'.

(13) a. ...heŋgojit gálvvut ja dasto gokčojit **loavdagiiguin**.
...hang goods and then cover tarp.COM.PL
'...the things are hung up and covered with tarps.'

b. ...*gokčat* **gokčasiin** ruskkaid go vuoj[i]hit tilheŋŋáriin ...
...cover tarp.COM garbage.ACC.PL if transport trailer.COM ...
'...cover it with a tarp if you transport it with a trailer ...'

c. ...*dohppe* sábehiid, soppiid ja **oađđinseahka** fárrui ...
...grab ski.ACC.PL, stick.ACC.PL and sleeping.bag.ACC with ...
'...take the skis, sticks and sleeping bag with you'

Movable non-mass concepts, on the other hand, belong to the material and substance

Prototype category	**Description**	**TEST (positive and negative)**
-moving -movable		
Sem/Plc	default place	leat X.LOC 'be in/at sth.'
Sem/Route	elongated	čuovvut X.ACC 'follow X', boahtit dán X.GEN 'come this X'
Sem/Plc-water	water place	vuodjat X.GEN rástá 'swim through X'
Sem/Plc-elevate	elevation	mannat X.GEN badjel 'go over X'
Sem/Build	built	mannat X.GEN sisa 'go into X', hukset X.ACC deike 'build X here'

Table 4.6: +concrete -animate -moving semantic prototype tags for North Sámi

prototype categories. Both are typically found in locative case in constructions with the verb *ráhkadit* 'make' as in ex. (14-a)–(14-b), characterizing what something else is made of, i.e. *silkkis* 'of silk' or *čázis* 'of water'.

(14) a. ...čeabet duojit leat *ráhkaduvvon* **silkkis**.
...neck handicrafts are made.PASS.PRFPRC silk.LOC
'...the necklaces are made out of silk.'

b. ...oastit mášiinna mii jándoris *ráhkada* viinna **čázis**.
...buy machine that day.LOC makes wine water.LOC
'...buy a machine that makes wine out of water in a day.'

Non-movable concepts, on the other hand, are typically places and are often LOCATION-arguments of verbal governors, cf. Table 4.6. They can be used with particular adpositions like *sisa* 'inside' (i.e. a building). They can also be followed (i.e. a route), or appear in genitive or accusative in place adverbials.

4.2.4.2 Testing abstract categories

The abstract branch of the feature hierarchy also pays attention to disambiguation and governor-argument matching. Abstract categories are split into those that have a time dimension (cf. Table 4.7) and those that do not (cf. Table 4.8).

Temporal concepts are often realized as time adverbials and can typically appear with temporal postpositions such as *maŋŋel* 'after' and *ovdal* 'before'. Members of the event prototype category such as *čájálmas* 'show' in ex. (15-a), members of the time and date prototype categories (e.g. *njukčamánu 1. b. 2005* '1 March 2005' in ex. (15-b)), and members of the activity prototype category (e.g. *searvan* 'participation' in ex. (15-c)) can all appear in temporal prepositional clauses with *ovdal* 'before'.

Prototype category	**Description**	**TEST (positive and negative)**
-concrete +temporal +count +measr		
Sem/Time	time expression	X.GEN maŋŋel 'after X', ádjánit X.ACC 'take X (amount of time)'
-concrete +temporal +count -measr		
Sem/Event	arranged or natural	X.GEN maŋŋel 'after X', mannat X.ILL 'go to X', lágidit X.ACC 'arrange X'
Sem/Edu	educational event	vázzit X.ACC 'walk', addit X.ACC 'hold X'
Sem/Wthr	weather event	Odne lea X.NOM 'Today there is X', birget X.ESS 'manage in X', X.GEN maŋŋel 'after X'
-concrete +temporal -count		
Sem/Act	activity/action	álggahit X.ACC 'start X', X.GEN maŋŋel 'after X'

Table 4.7: -concrete +temporal semantic prototype tags for North Sámi

(15) a. . . . gos oahppit ohppe sámivuođa birra *ovdal* **čájálmasa**.
. . . where students learned Sáminess about before show.GEN
'. . . where the students learned about being Sámi before the show.'

b. Ráđđehus lea mearridan ahte buot departementtat galget válbmet
government has decided that all departments should finnish
iežaset láhka- ja njuolggadusjorgalanplána *ovdal*
their law- and rule.translation.plan before
njukčamánu 1. b. 2005.
1 March 2005
The government has decided that all departments should hand in their plan for the translation of laws and regulations before 1 March 2005.'

c. *Ovdal* hanseáhtaid **searvama** Bergengávppašeapmái . . .
before Hanseat participation Bergen.trade.ILL . . .
'Before the participation of the Hanseats in the Bergen trade . . . '

Temporal expressions are further divided into countable and non-countable nouns (i.e. members of the activity prototype). Countable nouns are split into those that function as a measuring unit, i.e. the time prototype category, and those that do not, the event prototype category. Together with a numeral, events like *konseartta* 'concert (Gen.)' in ex. (16-a) can typically be counted, but they do not measure time if counted. Nouns of the time prototype category, e.g. *jahki* 'year', can further be used as adverbials in genitive or accusative case, cf. ex. (16-b). Nouns of the event prototype category, on the other hand, cannot be used as adverbials in genitive or accusative case. Events can be specified further, as educational or weather events, for example, depending on the context they appear in.

Prototype category	**Description**	**TEST (positive and negative)**
-concrete -temporal +local		
Sem/Plc-abstr	abstract place	mannat X.ILL 'go to X',
Sem/Dir	direction	mannat X.ILL 'go X way'
Sem/State	induced from the outside	leat X.LOC 'be in X', leat X.GEN dilis 'be in a state'
-concrete -temporal -local +gradable		
Sem/Feat	permanent and momentary characteristic	dovdomearka lea X 'the characteristic is', dus lea eanet X go mus 'you have more X than me'

Table 4.8: -concrete -temporal semantic tags for North Sámi

(16) a. Dan oktavuođas son doalai guokte **konseartta** Anáris ...
this context.LOC s/he held two concert.GEN Inari.LOC ...
'In this context s/he held two concerts in Inari ...'

b. ...allaskuvla lea máŋga **jagi** ožžon 300 000 kr ...
...highschool has many year.GEN received 300 000 crowns ...
'...the high school has received 300,000 crowns for many years ...'

Non-temporal concepts are divided into local and non-local concepts. Local concepts are often LOCATIONS and can be realized as postpositional phrases with local adpositions such as *siste* 'inside' in ex. (17-a), or in locative case such as *birrasiin* 'surrounding (Loc. Pl.)' in ex. (17-b). Non-local categories are split into gradable and non-gradable concepts. Gradable concepts are members of the feature prototype category, such as *dearvvašvuohta* 'health' and *guhkkodat* 'length'. They typically describe another concept (e.g. *sii* 'they' in ex. (17-c) and *áigodat* 'time period') and can be modified by a comparative such as *buoret* 'better' in ex. (17-c) or by *seamma* 'same' in ex. (17-d).

(17) a. **Kultursuorggi** *siste* lea ahtanuššama eaktun dat ahte ...
culture.branch.GEN inside is thriving.GEN requirement.ESS it that ...
'Inside the cultural branch the requirement for thriving is that ...'

b. Jávrrážiid **birrasiin** lea rássešaddu.
lake.GEN.PL surrounding.LOC.PL is grass.growth
'In the area around the lake, there is grass growth.'

c. ...ja *sis* lea *buoret* **dearvvašvuohta** go máŋgasis earáin.
...and they.LOC have better health than many.LOC other.LOC
'...and they have better health than many others.'

d. ...guokte vuosttaš áigodaga eai leat *seamma* **guhkkodagas** ...
...two first time.period not are same length.LOC ...
'...the first two time periods are not of the same length ...'

Non-gradable concepts, cf. Table 4.9, are divided into measurable and non-measurable

Prototype category	**Description**	**TEST (positive and negative)**
-concrete -temporal -local -gradable +measure		
Sem/Measr	measuring unit	lassánit 100 X.COM 'increase by X'
Sem/Curr	currency	máksit 100 X.GEN 'cost 100 X'
-concrete -temporal -local -gradable -measure		
Sem/Perc-emo	not countable	dovdat garra X.ACC. 'feel a strong X', Mun lean X.LOC 'I am in X'
Sem/Perc-phys	physical perception	oaidnit/dovdat X.ACC 'see/feel X'
Sem/Prod-vis	visual product	geahččat X.ACC 'watch X'
Sem/Prod-audio	audible product	guldalit X.ACC 'listen to X'
Sem/Prod-ling	linguistic product	čállit/dadjat X.ACC 'write/say X'
Sem/Lang	language	hállat X.ILL 'speak to X' jorgalit X.LOC Y.ILL 'translate from X to Y'
Sem/Rule	convention, rule	X.GEN mielde galgá bargat nie 'according to X one should do that way' / čuovvut X.ACC 'follow X'

Table 4.9: -concrete -temporal -local -gradable semantic tags for North Sámi

concepts. Measurable concepts include the measure and currency prototype categories, which can be used as units. Non-measurable concepts include the perception, product, tools, language, and rule prototype categories.

4.2.5 Semantic prototypes and the lexicon

Semantic prototype categories are implemented in the form of prototype tags in the North Sámi lexica, *nouns.lexc*, *propernouns.lexc*, *adverbs.lexc*, and *adjectives.lexc*. The lexicon contains primarily morphological information about inflection, derivation and compounding. The lexica are split into different modules for distinct parts of speech. However, the distinction between e.g. nouns and adjectives is not clear-cut, cf. Nielsen (1926-1929, pp.60–61). In ex. (18), the adjective *bealjeheapme* 'deaf' of the human prototype category is syntactically used like a noun, as an argument of the nominal governor *olámuddu* 'reach'.

(18) Nu šaddet dábálaš TV-sáddagat olámuddosis maiddái **bealjehemiide**
so become normal TV-broadcasts reach.ILL.PXSG3 also deaf.ILL.PL
'That way, normal TV broadcasts become available to deaf people'

Lexicon entries such as the one for *ealga* 'moose' in Figure 4.6 include the lemma, a specification of the compounding potential (i.e. should the first part of the compound be in nominative or genitive case or should it be a genitive plural form (*Cmp-*

N/SgN+CmpN/SgG+CmpN/PlG)), and a semantic prototype tag, i.e. *Sem/Ani*, separated by *+*-signs. Morpho-phonological information is given after the colon (i.e. the form used in the two-level transducer), followed by the reference to a continuation lexicon, i.e. *GOAHTI-A*, where the generation and analysis of the complete inflectional and derivational paradigm is specified.

```
ealga+CmpN/SgN+CmpN/SgG+CmpN/PlG+Sem/Ani:eal'ga GOAHTI-A ;
```

Figure 4.6: The lexicon entry for *ealga* 'moose' in *nouns.lexc*

Inflection and derivation that do not change the part of speech do not influence semantic prototype membership. However, in compounding, typically only the semantic tags of one of the compound parts should be preserved. North Sámi is a compounding language and according to the norm, the resulting compound of two or more lemmata is written as one word, cf. *Čállinrávagirji* (2003, p.60). For practical reasons, e.g. producing correct suggestions in spell-checking, *nouns.lexc* includes lexicalized compounds, which include semantic prototype tags referring to the compound rather than its parts. Typically, in dynamic compounding, nominal lemmata can form a compound with any other lemma in the noun lexicon. However, a few noun lemmata are not available as first parts of a compound. These are typically nouns appearing only in a specific morphological form in *nouns.lexc*, such as *allu* 'height' (which only appears in accusative case), and certain spellings of loan words such as *fax* 'fax'.

When assigning the semantic prototype tag of a dynamic compound, the morphological analyzer proceeds in the following manner: Whenever a lexicalized version of a compound is found in the lexicon, its analysis is preferred over the dynamic compound's analysis. If no lexicalized compound is listed, the prototype tags of all parts of the compound are preserved when processed by the North Sámi Xerox morphological finite state analyzer (*xfst*). For the hypothetical dynamic compound *ealgasadji* 'moose place' in Figure 4.7, l.1, both the animal prototype tag *Sem/Ani* for *ealga* 'moose' and the place prototype tag *Sem/Plc* for *sadji* 'place' are preserved. The *Giella-sme* system architecture includes a reformatter, *lookup2cg*, which then makes the format Constraint Grammar-compatible, cf. l.3–4 in Figure 4.7. There, only the prototype category of the last part of the compound, i.e. *Sem/Plc*, is preserved. However, this analysis requires that all compounds be head-final, and that the semantic prototype category of the head be transferred to the whole compound.

There are a number of compounds that, for various reasons, are not members of the same semantic prototype category as the last element of the compound, cf. Table 4.10, some of which are mentioned by Nickel and Sammallahti (2011, p.664). These

```
1    ealgasadji ealga+Sem/Ani+N+SgNomCmp+Cmp#sadji+Sem/Plc+N+Sg+Nom
2
3    "<ealgasadji>"
4            "ealga#sadji" Sem/Plc N Sg Nom
```

Figure 4.7: *Xfst* and *lookup2cg* analyses of the dynamic compound *ealgasadji* 'moose place'

Compound	**Meaning**	**Noun 1**	**Noun 2**
jahkebealle	'half a year'	*jahki* 'year' Sem/Measr_Time	
kilobealle	'half a kilo'	*kilo* 'kilo' Sem/Measr	
eadnebealle	'stepmother'	*eadni* 'mother' Sem/Hum	*bealle* 'half' Sem/Part
luossabealle	'half a salmon'	*luossa* 'salmon' Sem/Ani	
jumešbealle	'one of the twins'	*jumeš* 'twin' Sem/Hum	
luossalahkki	'half of a salmon'	*luossa* 'salmon' Sem/Ani	*lahkki* 'half' Sem/Part
mánnáriehpu	poor child	*mánná* 'child' Sem/Hum	*riehpu* 'poor creature' Sem/Hum
nástegállu	'star forehead' police	násti 'star' Sem/Obj	gállu 'forehead' Sem/-Body

Table 4.10: Semantically irregular compounds in North Sámi

can be compounds that are members of the prototype category of the left-most element. Alternatively, their prototype category membership can differ from the membership of any of the parts.

The compound *jahkebealle* 'half a year', cf. ex. (19), is a member of the time prototype category like the first part of the compound, cf. Figure 4.8, l.1. The prototype category of the second part of the compound (*Sem/Part*) is removed in the *lookup2cg*-analysis, cf. Figure 4.8, l.4–5. As the compound is lexicalized, cf. Figure 4.8, l.2, the lexicalized version with the correct semantic prototype tags can be selected.

(19) ...lei váttis bargodilli vuosttaš **jahkebeale**.
...was difficult work.situation first half.year.GEN
'...there was a difficult work situation during the first half of the year.'

Compounds with *-bealli/-bealle* 'half' are heterogeneous with regard to their prototype category. Not all compounds with *-bealli/-bealle* 'half' are left-headed compounds. While time and measure nouns do not lose their time-ness and measure-ness when split in half, half animate concepts may lose their animacy. While *luossa* 'salmon' is a member of both the animal and food prototype categories, *luossabealle* 'half a salmon' is only a member of the food prototype category, cf. ex. (20-a). *Eadnebealle* 'stepmother', on the other

```
jahkebealle        jahki+Sem/Measr+Sem/Time+N+SgNomCmp+Cmp#bealle+Sem/Part+N+Sg+Nom
jahkebealle        jahkebealli+Sem/Time+N+Sg+Nom

"<jahkebealle>"
        "jahkebealli" Sem/Time N Sg Nom
```

Figure 4.8: *Xfst* and *lookup2cg*analysis of the lexicalized compound *jahkebealle* 'half a year'

hand, is not a mother cut in half. The halfness is a metaphorical one, in the sense that she is not the biological mother. The semantic prototype category of the right-most noun *eadni* 'mother', i.e. human, is preserved in the compound. The same is true for *jumešbealli* 'twin', which is not a type of twin, but still a member of the human prototype, cf. ex. (20-b). When applied to words that have a plural connotation, *-bealli/-bealle* 'half' stresses the singularity, e.g. *jumešbealli* '(only one) twin'.

(20) a. Olles **luossabealit** bassojit dolas ...
whole salmon.half.NOM.PL fry.PASS.PRS.3PL fire.LOC ...
'Whole salmon halves are fried over the fire ...'

b. Mu oappás lea **jumešbealli**.
my sister.LOC has twin
'My sister has a twin.'

Also *-riehpu* 'poor creature' is underspecified as regards its prototype category. By itself, it is typically used to refer to animate concepts. As a compound it can be used to refer to humans, cf. ex. (21-b), animals, cf. ex. (21-a), and body parts, cf. ex. (21-c).

(21) a. **áldoriehpu** ii birge, ja ribaha miesi.
poor.reindeer.cow not manage, and loses calf
'the poor reindeer cow does not manage, and loses her calf.'

b. Vuoi **mánnariebut**!
oh child.poor.NOM.PL
'Oh poor children!'

c. **bierggasriehpu** leai galbmon skihččát, ja dat su moarsi gal
tool.poor had frozen stick.out.INF, and it his bride definitely
ii liikon go oinnii, ahte lea galbmon su irgi.
not liked when saw, that had frozen her groom
'his poor penis was frozen stiff and erect, and his bride did not like it when she saw that her bridegroom had frozen.'

Compounds with *-lahkki* are also heterogeneous. While *fanaslahkki* 'half boat' in ex. (22-a) can no longer be categorized as a vehicle, *juolgelahkki* 'half leg, broken leg' is still a body part, cf. ex. (22-b), and *beaivelahkki* is still a member of the time prototype category, cf. ex. (22-c).

(22) a. Boares **fanaslahkki** lea geavahuvvon goahtedáhkkin, ...
old boat.half is used hut.roof.ESS, ...
'The old boat half is used as a hut roof, ...'

b. Ollu bohccot mannet bealleheakkas ja **juolgelahkiiguin** mehcciide ...
many reindeer go half.dead and broken.leg.COM.PL forests ...
'Many reindeer go half dead and with broken legs to the forests ...'

c. Diimmut vásse, idja ja vel **beaivelahkki**.
hours went.by, night and also day.half
'The hours went by, the night and half a day as well.'

Other compounds do not assume the semantic prototype category of any of their parts. The compound *nástegállu* is a typical reindeer name, but is also used as a term for police as confirmed by informant *H* and mentioned by Svonni (2013, p.150) ("polis (äldre utryck)"). However, it cannot be found in *SIKOR*. The compound *nástegállu* is a member of the human prototype category, even though neither *násti* 'star', nor *gállu* 'forehead' is a member of the human prototype category. The compounds named in this section are lexicalized. However, listing all compounds with *-bealle* 'half', *-lahkki* 'half', *-riehpu* 'poor creature' is only possible if their compounding processes are not productive. According to Trosterud (2003, p.84f.), productivity of a morphological process is given when a calculation of all lexemes participating in the process is not possible. The global productivity of a word formation process according to Baayen (1993, p.181) is calculated by $P=n_1/N$, "where n_1 denotes the numbers of types with the required affix that occur only once (the so-called hapax legomena) and N the total number of tokens with this affix in some corpus". Baayen (1993) distinguishes *P*, the degree of productivity, from the number of types *V*, which is the extent of use.

Antonsen and Trosterud (2017) evaluate the productivity of a number of morphological processes for nouns, of which derived action nouns are the most productive, i.e. the productivity or the probability of encountering a new form is 7.47%. The productivity of compounding in general is 6.14%, which is slightly less than the productivity of the inchoative verbal derivation (6.73%). Table 4.11 shows that compounding with *-riehpu* 'poor' (30.5%) and textit-lahkki 'half' (29.5%) is far more productive than any of the processes described by Antonsen and Trosterud (2017). *N* in Table 4.11 includes all occurrences of this type of compound in *SIKOR*. However, the total number of occurrences (N) for these compounds is fairly small compared to the numbers reported by Antonsen and Trosterud (2017) (i.e. 70,759 for action nouns, 25,133 for inchoatives, and 1,602,886 for compounding in general), which is why the experiment should be repeated when a larger corpus is available. Nevertheless, the results for productivity show that semantic tagging for these types of compounds will need to be resolved in a rule-based manner in future analyzers.

Morphological formative	**Examples from *SIKOR***	**N**	**P=n_1/N**
-riehpu	áldoriehpu 'poor reindeer cow', mánnáriehpu 'poor child', bierggasriehpu 'poor device'	226	30.5%
-lahkki	beaivelahkki 'half a day', fanaslahkki 'a boat half', juolgelahkki 'broken leg'	61	29.5%
-ráidu	heargeráidu 'reindeer caravan', TV-ráidu 'TV series', várreráidu 'mountain chain'	2,797	7.9%
-bealle/-bealli	lihterbealle 'half liter', jumešbealli 'twin'	14,752	0.4%

Table 4.11: The productivity of left-headed/unpredictable compounds in North Sámi, where N = total occurrences of -x types of compounds, and P = the number of unique occurrences of specific compounds divided by N

4.2.6 Multiple categorization

Lemmata in *nouns.lexc* are members of one or several semantic prototype categories due to the choice of categories, polysemy and homonymy. The choice of categories has an impact on the (syntactic) generalizations that can be made. One lemma can belong to several categories that generalize over different aspects of the same lemma. The lemma *bearaš* 'family' is a member of both the human and the group prototype category. From its membership of the human prototype category (*Sem/Hum*) it follows that the noun can be used with a form of the relative pronoun *gii* 'who', and that it can be the subject of governors that require a human AGENT. Its membership of the group prototype category, on the other hand, indicates that *bearaš* 'family' denotes a group, usually implying that the singular noun can appear with a verb in third person plural.

Homonymy and polysemy can also be the reason for multiple semantic categorization. While homonyms are ethymologically and semantically unrelated, polysemous lemmata have two or more related meanings. Polysemy is a rather typical phenomenon in most languages. Borin et al. (2013, p.1199) mentions that in the Swedish lexical semantic resource *SALDO*, there is an "average [of] 1.1 senses per entry base form, and the most polysemous entry has 10 senses". Table 4.12 shows the frequency and semantic categorization of North Sámi nouns that have multiple semantic tags based on different senses or different semantic categorizations with syntactic implications.

The form *gássa* '1. gas, 2. box' is an example of homonymy. It is based on two identical lemmata of the same part of speech, each of which has a different inflection paradigm. The lemma *gássa* 'gas' is a member of the substance prototype category (*Sem/Substnc*), the lemma *gássa* (or alternatively *kássa*) 'box' is a member of the container prototype category (*Sem/Ctain*). While these are homonymous in nominative case, they differ in other cases. For example, the genitive case of *gássa Sem/Substnc* 'gas' is *gása*, but the genitive of *gássa Sem/Ctain* 'box' is *gássa*. The noun *gássa* is used in the sense 'gas' in ex. (23-a), where it

Noun	Semantic tags	***SIKOR***
Unrelated senses, different inflection paradigm (nominative, genitive)		
gássa	1. gássa, gása *Sem/Substnc* 'gas' 2. gássa, gássa *Sem/Ctain* 'box'	367
goddi	1. goddi, gotti *Sem/Ani* 'wild reindeer' 2. goddi, goddi *Sem/Hum* 'murderer'	328
doalli	1. doalli, doalli *Sem/Hum* 'host' 2. doalli, doali *Sem/Route* 'winter route'	298
beassi	1. beassi, beasi 'nest' *Sem/Build* 2. beassi, beassi 'bark' *Sem/Mat*	278
vádir	1. vádir, váhtara *Sem/Plant* 'maple' 2. vádir, vádira *Sem/Tool-measr* 'spirit level'	2
nelet	1. nelet, nelega *Sem/Part* 'fourth' 2. nelet, neleha *Sem/Food* 'clove'	0
Unrelated senses, same inflection paradigm		
luohkká	1. 'hill' *Sem/Plc* 2. 'class' *Sem/Group_Hum*, *Sem/Cat*	3,492
lávki	1. 'onion' *Sem/Fruit* 2. 'step' *Sem/Act*	1,261
sávdnji	1. 'seam' *Sem/Clth-part* 2. 'crack' 3. 'sauna' *Sem/Build*	87
fearga	1. 'ferry' *Sem/Veh* 2. 'color of the reindeer (fur)' *Sem/Feat-phys*	70
gáhttu	1. 'roof' *Sem/Build-part* 2. 'cat' *Sem/Ani*	29
linsa	1. 'lens' *Clth-jewl* 2. 'lentil' *Sem/Fruit*	9
biehkki	1. 'bit' *Sem/Obj* 2. 'reindeer mark' *Sem/Symbol*	5
Related meanings		
riekti	1. 'right' *Sem/Rule* 2. 'court' *Sem/Org*	5,341
mearri,-r-	1. 'amount' *Sem/Amount* 2. 'objective' *Sem/Semcon*	1,595
johtolat	1.'traffic' *Sem/Act* 2. 'street' *Sem/Route*	1,292
násti	1. 'star' i.e. object on the sky *Sem/Plc* 2. 'star' i.e. a famous person *Sem/Hum*	962
kruvdno	1. '(Norwegian) crown' *Sem/Curr* 2. 'crown' i.e. headdress *Sem/Obj*	634
vuoján	1. 'vehicle' *Sem/Veh* 2. 'reindeer' *Sem/Ani*	482
láse	1. 'glass' *Sem/Mat* 2. 'window' *Sem/Build-part* 3. 'drinking glass' *Sem/Ctain*	384
Other words with different semantic tags depending on the context		
bargu	1. 'work' *Sem/Act* 2. 'workplace' *Sem/Plc*	39,667
luondu	1. 'nature' *Sem/Plc* 2. 'human nature' *Sem/Feat-psych*	6,014
hip-hop	1. hip-hop' *Sem/Dance* 2. 'hip-hop' *Sem/Prod-audio*	61

Table 4.12: North Sámi nouns with multiple semantic tags

is coordinated with another noun of the semantic prototype category (*Sem/Substnc*), *olju* 'oil'. In ex. (23-b), on the other hand, it is used in its container (*Sem/Ctain*) sense 'box' and as the DESTINATION-argument of the verb *dovdnjet* 'hide'. Both coordination and government argument relations can serve as a clue in semantic prototype disambiguation, which also disambiguates the two senses and possible translations of *gássa*. The senses of the homonymous noun *gáhttu* '1. roof, 2. cat', on the other hand, only differ semantically, not morphologically. While *gáhttu* 'roof' is a member of the building-part prototype category, cf. ex. (23-c), *gáhttu* 'cat' is a member of the animal prototype category and is a Germanic loan word. The senses of the polysemous noun *láse* '1. glass (i.e. material and drinking glass) 2. window', on the other hand, are clearly related. They belong to the material prototype category, the container prototype category, cf. ex. (23-d), and the building-part prototype category.

(23) a. Dát ledje olj[]u ja **gássa** ...
these were oil and gas.NOM ...
'These were oil and gas ...'

b. Sii dovdn[j]ejedje guliid **gássaide** ...
they hid fish box.ILL.PL ...
'They hid the fish in the boxes ...'

c. Sin ovttaseallimis seamma **gáhtu** vuolde ...
their living.together.LOC same roof.GEN under ...
'During their time living together under the same roof ...'

d. ...gohčui son addit bárdnái ovtta **láse** viinni.
...called s/he give.INF son.ILL one glass.ACC wine
'...s/he asked for a glass of wine for her/his son.'

Different types of compounds with the same semantic head can reflect the polysemy of a noun, cf. Table 4.13.

The noun *ráidu* 'caravan, series' is used in various types of compounds. In ex. (24-a)–(24-b), both *heargeráidu* 'reindeer caravan' and *ráidu* 'caravan' are members of the animal prototype category. *TV-ráidu* 'TV series' and *ráidu* 'series' in ex. (24-c)–(24-d), on the other hand, are members of the visual product prototype category. *Várreráidu* 'mountain chain' in ex. (24-e) is a member of the elevated place prototype category. However, the simple noun *ráidu* does not have this connotation.

(24) a. Son, gii jođihii **ráiddu**, ii jietnadan maidege.
s/he, who led caravan.ACC, not said anything
'The one who led the caravan did not say anything.'

b. ...dolvo guhkes **heargeráiddut** rievssahiid Bossegohmárkaniidda.
...brought long reindeer.caravans ptarmigans Bossekop.market.ILL.PL
'...long reindeer caravans brought the ptarmigans to Bossekop Market.'

c. Makkár mánáid **TV-ráiddus** lei son mielde go lei mánná?
which children's TV.series.LOC was s/he with when was child

Polysemous noun	**Compounds with different semantic tags**
diibmu 'hour, class, clock'	*diibmu* Sem/Measr Sem/Obj 'hour' *lávlundiibmu* Sem/Event 'singing class'
ráidu 'caravan, series, chain'	*heargeráidu* Sem/Ani Sem/Group 'reindeer caravan' *girjegávperáidu* Sem/Org 'bookstore chain' *TV-ráidu* Sem/Prod-vis 'TV series'
linjá 'line, studies'	*elfápmolinjá* Sem/Route 'electricity line' *duojárlinjá* Sem/Edu 'art studies' *moallalinjá* Sem/Plc-line 'goal line'
bábir 'paper'	*biebmobábir* Sem/Mat 'waxed paper' *vearrobábir* Sem/Txt 'tax paper'
rápma 'frame'	*biilarápma* Sem/Obj 'car frame' *bálkárápma* Sem/Semcon 'wage frame'
luohkká 'class'	*internáhttaluohkká* Sem/Group Sem/Hum 'boarding school class' *vearroluohkká* Sem/Cat 'tax class'
foarbma 'form, shape'	*gáhkkofoarbma* Sem/Ctain 'cake mold' *čuojahanfoarbma* Sem/Feat-phys 'shape to play (music)'
kássa 'tub, fund'	*margariidnakássa* Sem/Ctain 'margarine tub' *loatnakássa* Sem/Org 'loan fund'
ivdni 'color, coloring'	*konditorivdni* Sem/Substnc 'pastry coloring' *čakčaivdni* Sem/Feat-phys 'autumn color'
riekkis 'circle, ring, tire'	*johtolatriekkis* Sem/Route 'roundabout' *biibalriekkis* Sem/Org 'bible circle' *bealleriekkis* Sem/Geom 'half circle' *giehtariekkis* Sem/Clth-jewl 'arm ring' *dálveriekkis* Sem/Obj 'snow tire'

Table 4.13: North Sámi polysemous nouns and their compounds

'In which children's TV series did s/he take part as a child?'

d. NRK lea ráhkadan **ráiddu** sámiid dili birra ...
NRK has made series.ACC Sámi situation about ...
'NRK has made a series about the situation of the Sámi ...'

e. **Várreráiddu** bákti lea guovtti oasis.
mountain.chain.GEN wall has two part.LOC
'The mountain chain's wall consists of two parts.'

The noun *forbma* 'form, shape', on the other hand, is a member of the container prototype category in ex. (25-a) and in the compound *gáhkkoforbmii* 'cake mold (Ill.)' in ex. (25-b). It is a member of the feature prototype category in ex. (25-a) and as a compound, *čuojahanfoarbma* 'playing shape', in ex. (25-d).

(25) a. Sii leat maid leiken dani **forpmaide**.
they had also poured tin mold.ILL.PL
'They had also poured tin into the molds.'

b. Biergosuohkadas leikejuvvo **gáhkkoforbmii**
minced.meat poured cake.mold.ILL
'The minced meat is poured into the cake mold'

c. sámi nieiddat ... leat maid buoret fysalaš **forpmas** go muđui
Sámi girls ... are also better physical form.LOC than otherwise
riika nissonat.
country's women
'Sámi girls ... are also in a better physical shape than other women in the country.'

d. Mis lea hui buorre **čuojahanfoarbma** dál
we.LOC have very good playing.shape now
'We are in a very good shape for playing now'

4.3 Evaluation

This section includes an evaluation of the lexicon and corpus coverage of the semantic prototype tags for North Sámi in *Giella-sme*. It further includes an evaluation of four examples that illustrate the syntactic relevance of semantic prototype tags in their distribution.

	nouns.lexc	**%**
Lexicon coverage		
Nouns (total)	91,825	
Nouns with a semantic tag	65,598	71.44%
Nouns with more than one semantic prototype	2,777	3.024%
Corpus coverage (token)		
Cohorts with a noun analysis	14,209,002	
Noun cohorts with a semantic tag	12,771,984	89.89%

Table 4.14: Lexicon and corpus coverage of North Sámi semantic prototype tags in *Giella-sme*

4.3.1 Lexicon and corpus coverage

I evaluated the morphologically and partly semantically tagged noun lexicon *nouns.lexc*[13] with regard to both lexicon coverage and corpus coverage (token) on a fully annotated version of *SIKOR*. The results are presented in Table 4.14.

The lexicon *nouns.lexc* has 50,403 entries and is comparable in size to the lexical semantic resources *WordNet* (95,600 entries) and *SALDO* (76,750 entries). Of the lemmata in the noun lexicon, 71% have at least one semantic tag and 3% have more than one. Corpus coverage is higher than lexicon coverage, which means that the items tagged in the lexicon have a high frequency. Of all noun analyses in the corpus almost 90% have at least one semantic tag. As expected, most nouns in the corpus belong to the human prototype category followed by the organization and place prototype categories.

4.3.2 Syntactic relevance of semantic prototypes

I used four test cases to evaluate the syntactic relevance of semantic categories. The evaluation is based on a corpus search of *SIKOR*.[14]

I tested the distribution of the nominal prototype categories for two postpositional constructions and two verb-object constructions. The postpositional phrases contain the postpositions *rastá* 'across' and *ala* 'on, onto', the latter of which is tested in the context of the verb *bidjat* 'put'. The verb-object constructions are governed by the multi-word verb *bidjat johtui* 'put into action, get started' and *máksit* 'pay, mean'.

The form *ala* 'on' preceeded by a noun in genitive/accusative case in combination with the verb *bidjat* 'put' occurs 410 times in *SIKOR*. The form is ambiguous as to a postposition- and adverb-reading. As an adverb, *ala* is part of the idiomatic construction meaning 'turn on', which requires an object. As an adposition, *ala* 'on, onto' requires a noun phrase in genitive case. In order to disambiguate between both the syntactic function of the noun and the part of speech of *ala* 'on', it is useful to know more about the semantic

[13]Version r146045 (Accessed 2017-01-04)
[14]containing 22,093,728 words (Accessed 2014-02-01)

distribution of the nouns in these syntactic contexts. An analysis of the occurrences of *ala* after *bidjat* shows that 96.29% (396 occurrences) are adpositional constructions and 3.71% (15 occurrences) are adverbial constructions.

67% of the nouns in the adverbial constructions are members of the el-object prototype or the audio-product prototype category (10 occurrences), 20% are clothes (3 occurrences), and the remaining 13% are members of the food category. The semantic range of objects in the adverbial construction is fairly restricted. However, some of the same categories also have single occurrences in adpositional constructions, i.e. 1.8% are members of the of el-object category (seven occurrences), 0.8% are members of the audio-product category (three occurrences), and 0.5% are members of the clothes prototype category (2 occurrences).

Members of the clothes prototype category typically appear with *ala* 'on' in an adverbial construction, cf. ex. (26-a). However, there are also single instances of nouns of the clothes prototype category in the postpositional construction, cf. ex. (26-b). The postpositional construction can be distinguished from the adverbial construction by identifying the accusative object of *bidjalit* '(quickly) put' (which is a derivation of *bidjat* 'put'), i.e. *dan* 'it'. In addition to a semantic analysis, a deep syntactic analysis associating governors with their arguments is necessary to fully disambiguate the adverb reading from the adpositional reading.

(26) a. Vuos bidjaleaba **biktasa** ala ja leage vuosttaš geardde ...
first put.PRS.3DU clothes.ACC on and is.also first time ...
'First they quickly put the clothes on and it is also the first time ...'

b. Son válddii eret bearralčiŋa ja bidjalii *dan* **biktasiid** ala
s/he took away pearl.jewelry and put it.ACC clothes.ACC.PL on
'S/he took away the pearls and quickly put them on top of the clothes'

The distribution of semantic prototype categories in the adpositional constructions is represented in Figure 4.9. The distribution of semantic prototype categories was much more spread out than for the adverbial construction. However, there are clear semantic tendencies as well. Animate categories like human, organization and animal are the most common ones. Together with nouns of the body prototype these make up 45%. Members of the furniture and place prototypes make up 22%. It is also interesting to note that 91% are concrete concepts.

While the verb *bidjat ala* 'put on' generally has a concrete meaning of placing something, cf. ex. (27-a), together with inanimate concrete categories it can have an abstract meaning as well. This is the case in ex. (27-c), where it is used with a member of the abstract noun *vuođđu* 'basis'. Even with members of animate categories like *studeanttaid* 'student (Acc. Pl.)' in ex. (27-b) *bidjat ala* 'put on' can have an abstract meaning.

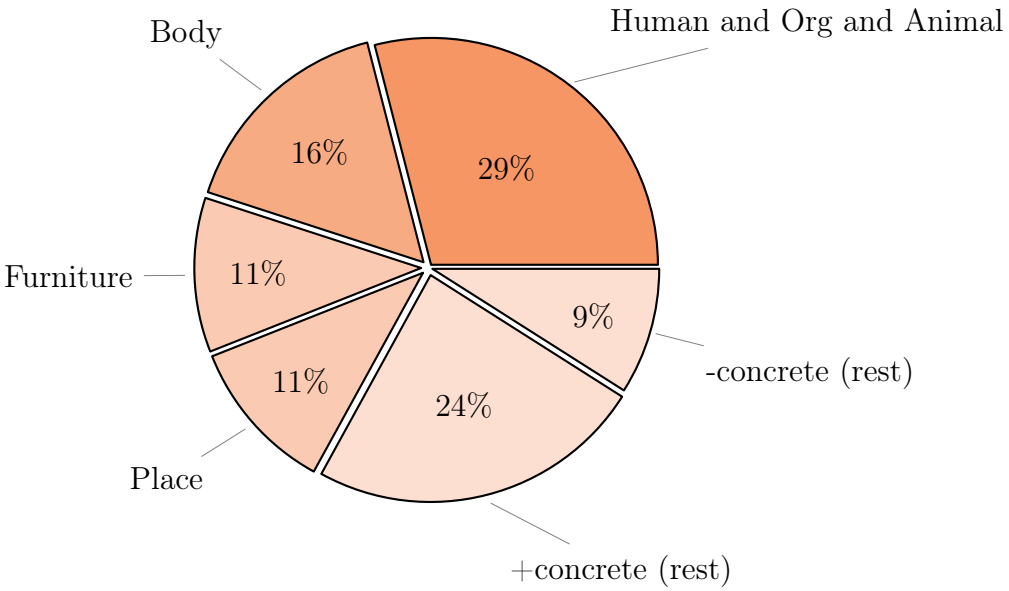

Figure 4.9: The distribution of semantic categories in dependents of *ala* 'on' co-occurring with *bidjat* 'put'

(27) a. Soalddát bijai bissuid **beavddi** ala.
soldier put rifles table.GEN on
'The soldier put the rifles on the table.'

b. ...ii go son leat bidjan **studeanttaid** ala beare stuora noađi.
...not Q s/he have put student.GEN.PL on too big load.ACC
'...hasn't s/he put too big a load on the students.'

c. ...bidjá našuvnnalaš oahppoplánaid sámi sisdoallu **vuođu**
...put national teaching.plan.ACC.PL Sámi content basis.GEN
man ala hukse
which.GEN on builds
'...the content of the national teaching plans lays a foundation on which one builds'

Typical sentences with *rastá* 'through' include adpositional phrases such as for example *joga rastá* 'across the river' in ex. (28-a) and *ráji rastá* 'across the border' in ex. (28-b).

(28) a. Sutnje dagai ovtta, sugaigo **joga** rastá ihkku
s/he.ILL did one, row river.GEN across night
'For her/him it didn't matter, if s/he rowed across the river during the night'

b. Leigo áidna vejolašvuohta mannat **ráji** rastá?
was only possibility go border.GEN across
'Was it the only possibility of crossing the border?'

Passing through any type of substance, like water, for example, to reach another place, typically prefers *rastá* 'through' as an adposition, rather than *badjel* 'over'. Verbs of motion that do not express passing through water, on the other hand, prefer the adposition

badjel ‘over’, cf. ex. (29-a). In order to detect lexical errors related to adposition use, it is useful to investigate the semantic preferences of each adposition. The use of *rastá* ‘through’ with the noun *šaldi* ‘bridge (Gen.)’ in ex. (29-b) should therefore be marked as a lexical error based on its semantic prototype category.

(29) a. . . . vuodjit skuteriin muhtin **joga** badjel.
. . . drive scooter some river.GEN over
‘. . . drive over some river with the scooter.’

b. *ii oktage sis sáhte vázzit rastá ovtta **šaldi** mas leat nálut
not one they.LOC can walk across one bridge.GEN which has nails
‘none of them can walk across a bridge that has nails’

Figure 4.10 shows the semantic distribution of genitive complements of *rastá* ‘across’ in *SIKOR*. Adpositional phrases with *rastá* ‘across’ appear predominantly (88%) with members of the place categories (i.e. place, route, organization). Of these, 77% are linear places, e.g. *rádji* ‘border’, members of the route prototype category, e.g. *luodda* ‘path’, and linear water-places, e.g. *johka* ‘river’. Marginally, there are a few abstract categories, such as *kultuvra* ‘culture’, a member of the feature prototype category, in ex. (30-a) and *fágá* ‘subject’, a member of the text prototype category, in ex. (30-b). Error detection rules can build on semantic preferences in order to find lexical errors.

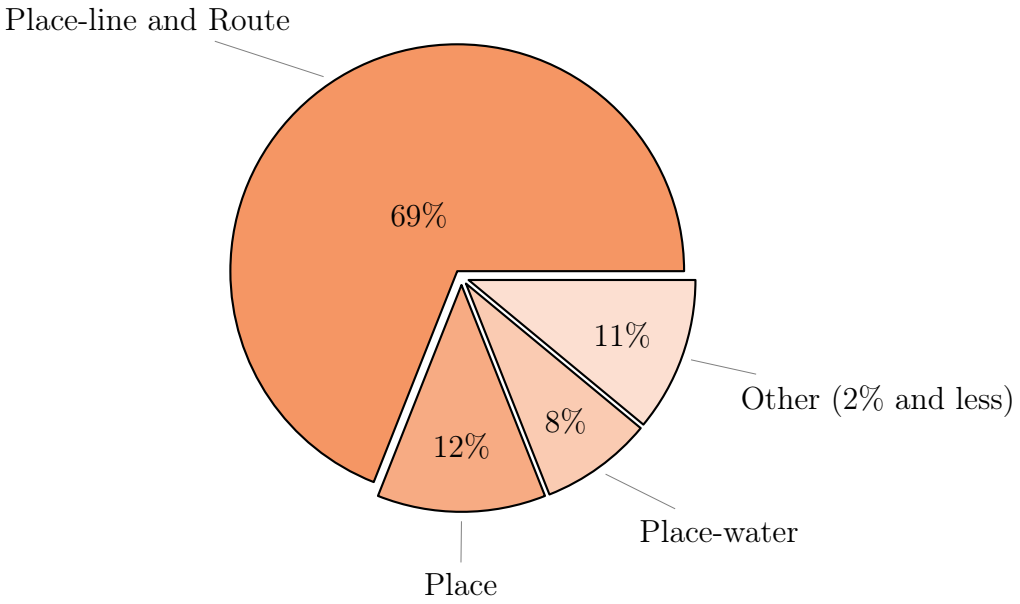

Figure 4.10: The distribution of semantic categories in dependents of *rastá* ‘through’

(30) a. Ovttasbargu **kultuvrraid** rastá
cooperation culture.GEN.PL across
‘Cooperation across cultures’

b. . . . reflekteret oahpaheami fágaid siskkobealde ja
. . . reflect teaching subject.GEN.PL within and

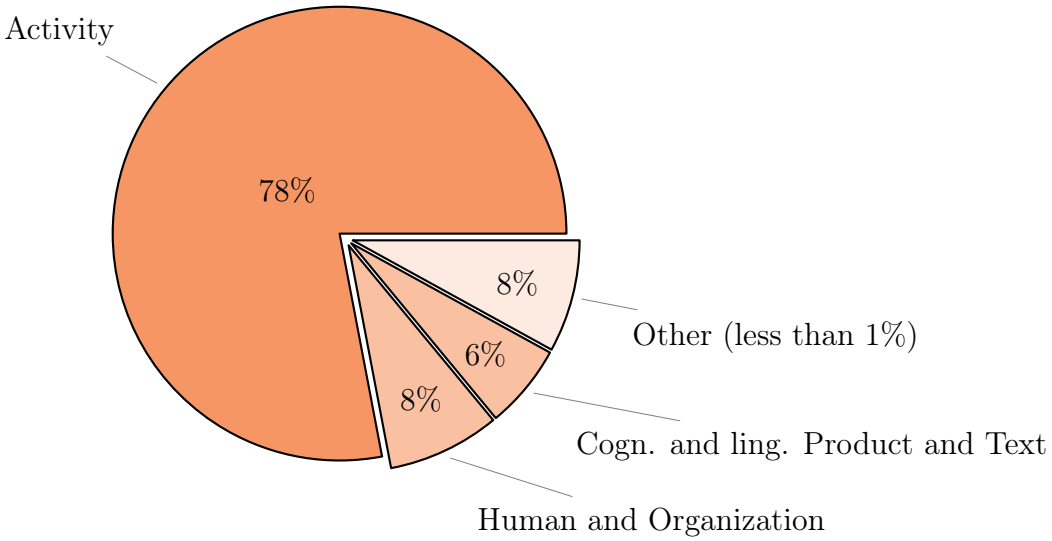

Figure 4.11: The distribution of semantic categories in objects of *bidjat johtui* 'put into action'

fágaid rastá ...
subject.GEN.PL across ...
'...reflect the teaching within and across the subjects ...'

Figure 4.11 represents the semantic distribution regarding the object position of the governor *bidjat johtui* 'put into action'. Recognizing accusative objects and distinguishing them from genitive modifiers is one of the most challenging tasks in morpho-syntactic disambiguation. Since morpho-syntactic contexts are often ambiguous, lexical semantic information regarding what kind of objects are probable facilitates disambiguation.

The distribution of the prototype categories of the objects associated with *bidjat johtui* 'put into action' shows clear semantic preferences. As much as 78% of the nouns belong to the activity prototype category. These are often fixed expressions such as *bidjat johtui doaimmaid* 'initiate activities' as in ex. (31-a). Other typical nouns are *prošeakta* 'project' as in ex. (31-b) or any compound of *bargu* 'work'. Eight percent belong to the human (e.g. *joavku* 'group' in ex. (31-c)) and organization (e.g. *skuvla* 'school' in ex. (31-d)) prototype categories. Six percent belong to the linguistic and cognitive product prototype categories and to the text prototype category. Membership of the cognitive product and activity prototype categories can be ambiguous: for example, *muitalus* 'story' is a product of and related to the activity of telling.

(31) a. ...hásttuha ráđđehusa bidjat johtui konkrehta **doaimmaid**
...challenges government put motion.ILL concrete activity.ACC.PL
'...s/he challenges the government to initiate concrete activities'

b. Davvi-Romssa Musea lea bidjan johtui máŋga **prošeavtta**
North-Troms Museum has put motion.ILL many project.ACC
'The North Troms Museum has initiated many projects'

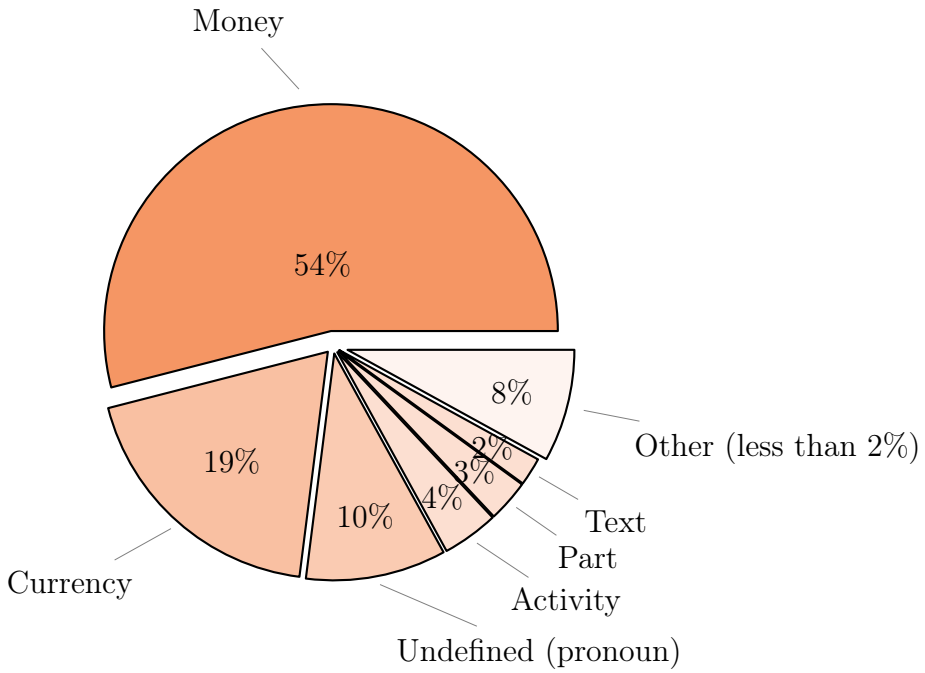

Figure 4.12: The distribution of semantic categories in objects of *máksit* 'pay'

c. . . . evttohusa bidjat johtui **joavkku** . . .
. . . suggestion put motion.ILL group.ACC . . .
'. . . the suggestion to initiate a group . . . '

d. . . . bidjat johtui **boazodoalloskuvlla** Guovdageaidnui
. . . put motion.ILL reindeer.herding.school.ACC Kautokeino.ILL
'. . . intiate a reindeer herding school in Kautokeino'

The semantic distribution of objects of the verb *máksit* is represented in Figure 4.12. The verb *máksit* '1. pay, 2. cost, 3. mean' is polysemous and corresponds to various translation equivalents in English. The semantic prototype category of the object is therefore not only relevant for governor-argument matching, but also to machine translation.

66% of the objects are members of the money, currency, and part prototype categories. Accusative objects of *máksit* meaning 'pay' typically refer to the object or service one pays for (THEME). These can be concrete objects, e.g. *gálvu* 'goods' in ex. (32-a), or services, e.g. *bálvalus* 'service' in ex. (32-b).

(32) a. . . . ja máksit **gálvvuid** báŋkogoarttain
. . . and pay good.ACC.PL bank.card.COM.PL
'. . . and pay for the goods with the bank cards'

b. Muhto go turista máksá **bálvalusaid** ruđain de . . .
but when tourist pays service.ACC.PL money.COM.PL then . . .
'But when the tourist pays the services with money then . . . '

However, objects can also refer to members of other categories: the building-part prototype category if they represent an object or a service one is paying for, e.g. *hoteallalatnja*

'hotel room' in ex. (33-a). Accusative objects can also refer to the amount of money someone pays (INSTRUMENT), typically expressed by a sum in a specific currency (e.g. *2700 ruvnno* '2,700 crowns' in ex. (33-b)) or by a term denoting the particular function of the payment, i.e. *vearru* 'tax' as in ex. (33-c), *doarjja* 'benefit', *sàhkku* 'fine', *láigu* 'rent', or *vealgi* 'debt'. These are members of the money prototype category. When the object is a sum in a specific currency, it can also be translated with 'cost' (apart from 'pay') in certain contexts, cf. ex. (33-d).

(33) a. Widerøe ges biehttalii máksimis **hoteallalanja**.
Widerøe again refused paying hotelroom.ACC
'But Widerøe refused to pay for the hotel room.'

b. In goassege boađe máksit **2700 ruvnno** dan ovddas.
not ever come pay 2,700 crown.ACC it for
'I wouldl never pay 2,700 crowns for it.'

c. ...ja ahte fertet máksit **vearu** ...
...and that have.to pay tax.ACC ...
'...and that you have to pay the taxes ...'

d. CD máksá **250 ruvnno**.
CD costs 250 crown.ACC
'The CD costs 250 crowns.'

Thirdly, the person who is being paid for a particular service can be expressed by an object in accusative case (RECIPIENT), cf. *advokáhta* 'lawyer' in ex. (34-a). When used in its sense 'mean', the verb *máksit* 'mean' occurrs predominantly in *dan ahte* constructions like in ex. (34-b). However, it may also appear together with an object of the time prototype category in accusative case such as *nealgejahki* 'year of hunger' in ex. (34-c). Although theoretically, any semantic prototype category can be used in the object position of *máksit* meaning 'mean', there is no occurrence of the money, currency, or part prototype category in *SIKOR*.

(34) a. Nu ahte mii šaddat ieža máksit **advokáhta**
so that we will ourselves pay lawyer.ACC
'So that we end up paying for the lawyer ourselves'

b. ...mii várra máksá **dan** **ahte** lassi dán ráiddus lea vuordimis.
...which maybe means it.ACC that more this series.LOC is waiting
'...which maybe means that there is more to wait for in this series.'

c. Maiddái dat čieža guoros gordneoaivvi maid nuortabiegga lea
also the seven empty grain.ears that.ACC.PL east.wind has
goldnadan mákset čieža **nealgejagi**.
dried mean seven hunger.year.ACC
'And the seven empty ears dried by the east wind shall be seven years of famine.' [Genesis 41:27]

4.4 Conclusion

This chapter discussed the theoretical basis for lexical semantic categorization, the lexical semantic annotation of the North Sámi lexicon, and an evaluation of this semantic annotation, both in terms of coverage and practical usage. I presented different approaches to semantic categorization, with a focus on defining and distinguishing, of which I chose the distinguishing approach. A semantic tag does not need to comprise a full semantic description of the noun, but rather a semantic generalization that is useful in syntax-based analysis, i.e. grammar checking, disambiguation, semantic role matching, machine translation, and word sense disambiguation.

For my annotation of the North Sámi lexicon, I used Bick's (2000) semantic prototype approach and added semantic tags primarily to nouns, but also to proper nouns, adjectives, and adverbs. I adapted his feature hierarchy to syntactically relevant features in North Sámi. I used a number of tests of syntactic relevance to ensure consistency in annotation. The annotation of members of the human and vehicle prototypes by means of syntactic tests was discussed in detail. This discussion illustrates the difficulty of categorizing peripheral members of a prototype category as they may not behave similarly in these tests, but still behave similarly in other tests. The complete tagset for prototype categories is presented together with certain syntactic tendencies and generalizations that can be drawn. The section on annotation of the North Sámi lexicon also contains a description of the lexicon and handling of the semantic prototype tags with regard to dynamic compounding, resolving issues. While the analysis of right-headed compounds is resolved within the morphological analyzer and the reformatter, compounds that behave differently with regard to their semantic category need to be lexicalized to receive the correct semantic analysis. I pointed out cases where the productivity of certain compound elements may require a rule-based solution of semantic annotation, as not all compounds can be listed in the lexicon. Lastly, I presented different causes of multiple semantic tagging: i.e. homonymy, polysemy, and categorization-related reasons.

In the final section, I evaluated both (lexicon and corpus) coverage of the North Sámi semantic prototype categories and their practical usage. More than 71% of the *nouns.lexc* is annotated with semantic tags, and almost 90% of the nouns in the corpus receive a semantic annotation. Practical usage includes morpho-syntactic disambiguation, error detection, government-argument matching and word sense disambiguation/lexical selection in machine translation. Disambiguation of systematic ambiguities such as adpositions and adverbs can often not be performed based on morpho-syntactic criteria only. However, the distribution of semantic categories associated with the genitive complements of the adposition *ala* 'on' shows clear semantic category preferences when compared to adverbial constructions. Error detection also benefits from the semantic annotation of nouns. Lexical (adposition) errors can be resolved based on the semantic category of the nominal

complement of the adposition, as some adpositions (e.g. *rastá* 'through' and *badjel* 'over') have clear semantic selection restrictions. The disambiguation of accusative objects and genitive modifiers is another challenge that can be improved by means of semantic annotation especially in the case of verbs that preferably appear in idiomatic constructions such as *bidjat johtui* 'initiate'. Semantic annotation can also facilitate lexical selection in machine translation. My evaluation showed that polysemous verbs such as *máksit* '1. pay, 2. cost, 3. mean' often have different semantic selection restrictions for each sense or translation equivalent. I also demonstrated that the semantic prototype categories chosen in *Giella-sme* successfully generalize over semantically similar items in a way that is useful for a number of syntactic tasks. In the next chapter, I will focus on the task of global syntactic error detection applying both the valency tags that were introduced in the previous chapter and the semantic prototype tags discussed in this chapter.

Chapter 5

Semantics and valency in grammar checking

Grammar checking can be about [*syntactic error*] detection, i.e. detection of syntactic errors, or syntactic [*error detection*], i.e. detection of errors by syntactic means. Not all errors that need a syntactic analysis of the context are of syntactic nature. Some errors are typos that result in real words, but to distinguish them from a correct use of the real word and identify them as a typo an analysis of the sentence is necessary.

Syntactic analysis of a sentence with potential grammatical errors needs to overcome a number of difficulties, including homonymy and syntactic ambiguity. It must also reconstruct the intended syntax despite the error and reach a certain depth to be able to match governors with their respective arguments, which is needed for global error detection. North Sámi has 2.6 grammatical possible analyses per word form (Trosterud and Wiechetek, 2007, p.401). Extensive morphological disambiguation (typically based on morpho-syntactic context) is therefore a prerequisite for error detection. Successful error detection does not require a full disambiguation of all words in a sentence, but rather an identification of the context relevant to the error. The context can be local, i.e. restricted to a single phrase, or global, i.e. it can take the entire sentence into account. Atwell (1987, p.42) found an average of 31% of non-words in a sample of 150 errors in English written text. A non-word is a word form that is not in the normative lexicon of a language. In many cases a non-word is the result of a typographical error. Apart from that, 38% of the errors can be found by means of a local syntactic analysis and 31% need a global syntactic analysis and/or a semantic analysis.

In addition to syntactic analysis morpho-syntactic disambiguation is also important for syntactic error detection. In ex. (1), *dán áigodaga* 'this period' can be analyzed as an adverbial or as an object of the verb *suhttat* 'get angry'. The verb *suhttat* 'get angry' typically asks for a THEME in illative case. An object in accusative case such as *dán áigodaga* 'this period' in the verb's immediate context would therefore trigger the valency error detection and correction (i.e. *áigodahkii* 'period (Ill.)'). However, here, the adverbial

reading is the correct one, in which case the annotation of a valency error would result in a false positive.

(1) Tigerat suhttet álkit dán **áigodaga**.
tiger.NOM.PL get.angry easily this period.GEN;ACC
'The tigers get angry easily in/at/during this period.'

Reconstructing the sense and grammar of a sentence with a grammatical error is another challenge in grammar checking. For example, if the finite verb itself contains an error, the whole analysis can crash as the analyzer may identify another ambiguous form in the sentence as the finite verb and associate the erroneous verb's arguments with it. In ex. (2), *bidjui* 'den (Ill.)' is a real word error of *biddjui* 'put (Pass. Prt. Sg3.)'. As the finite verb is missing in the sentence, the analyzer is likely to mistake *dušše* 'only' for a finite verb because it has a less frequent finite verb reading of *duššat* 'perish (Prt. 3Pl.; Prs. 1Du.)'. Only a very robust analyzer can maintain the intended sentence structure in its analysis.

(2) *Láđđi **bidjui** *dušše* násttiid vuollai.
loden den.ILL only;perish.PRT.3PL star.ACC.PL under
'Loden fabric was only placed under star-shaped silver buttons.'

A Constraint Grammar analyzer with its bottom up strategy can work with sentence fragments and output a syntactic analysis despite missing parts. This makes it very robust for the task of error detection. Just like the human brain it manages to reconstruct erroneous parts of the sentence by means of putting together other reliable information (i.e. from the lexicon) in the sentence.

While local syntactic error detection, the "safer" type of syntactic error detection, appears in most full-fledged grammar checkers, state-of-the art grammar checkers very rarely work with global syntactic errors. This chapter deals with modeling a safer way of achieving global error detection by means of semantic prototype tags and verb valency. This work is about ways of modeling a language norm, not about the norm itself. That means that I do not discuss what should be an error and what should not. Instead, I discuss ways of modeling these within grammar checking. The valency errors discussed here are based on the recommendations of *Čállinrávagirji* (2003), current grammars and dictionaries and native speakers' language intuitions. A grammar checker is generally based on an official norm. However, the current officially decided norms for North Sámi are mostly about typesetting, punctuation and spelling, cf. *Riektačállinrávvagat* (2015). Syntactic norms had been discussed in the previous version of the document (*Čállinrávagirji*, 2003), but have not been officially decided yet. This means that in the upcoming process, some of the rules discussed here may be removed, and others may be added.

I will address two types of errors in this chapter, real word errors (cf. ex. (3-a)) and valency errors (cf. ex. (3-b)), both of which may need a full and deep syntactic analysis

of the sentence enhanced by semantic prototypes and valencies.

(3) a. *Lea go imaš ahte balan **jamas** go soames namuha
is it strange that fear.PRS.1SG noise.LOC if someone mentions
skuvlla munnje?
school.ACC I.ILL
'Is it strange that it scares me to death when someone mentions the school to me?'

b. *Son *liikui* erenoamáš bures **sistesihkkelastima** ja danne ...
s/he like.PRT.3SG especially well indoor.biking.ACC and therefore ...
'S/he liked indoor biking a lot and therefore ...'

Real word errors are originally spelling errors resulting in a real word, i.e. Hashemi (2003) even explicitly calls them "Real Word Spelling Errors". In ex. (3-a), the confusion pair members are *jamas* 'noise (Loc.)' and *jámas* 'to death; dead', which are distinguished only by an accent on the first <a>. While *jámas* 'dead' is the intended form here, *jamas* 'noise (Loc.)' is a real word error. Real word error detection in *GoDivvun* makes use of morphological, syntactic and semantic context information depending on the relation of the confusion pair, the rareness/frequency of the forms, their part of speech, etc. While some confusion pairs need global syntactic analysis, others can be resolved in a local context.

Valency errors are errors in the realization of the arguments of a particular governor. They can also involve the governor itself, e.g. if a transitive derivation rather than the intransitive form is used, etc. However, here I will focus on the first type only. They can be case errors (cf. ex. (3-b)), but can also include omitted and redundant subjunctions introducing finite and infinitival arguments, etc. Case errors, on the other hand, can be both local case errors, e.g. case errors within noun phrases, and global case errors. Valency errors are the hardest to detect as the detection process requires a global syntactic analysis, i.e. in the case of ex. (3-b), in which both the finite verb of the sentence and its arguments need to be identified and distinguished from the arguments of other verbal and nominal heads. This requires some knowledge about verb valency and semantic prototype categories.

The form *sistesihkkelastima* 'indoor biking (Gen.;Acc.)' in ex. (3-b) is not a simple spelling error. According to the norm, the verb *liikot* 'like' should have a THEME in illative case, i.e. *sistesihkkelastimii*, rather than accusative case, i.e. *sistesihkkelastima* (*Čállinrávagirji*, 2003, p.87).

Below, I will present some general background on grammar checking focusing first on local and then on global error detection, on different rule-based grammar checkers for Finno-Ugric languages and within the Constraint Grammar framework, and on different approaches to global error detection. I will also give an introduction to the structure, framework, and error types in the North Sámi grammar checker *GoDivvun*. The second

section deals with the use of valencies and semantic prototypes in local and global error detection in *GoDivvun*. I will first describe the system architecture for local error detection and then choose two error types that make use of semantic prototype categories and valencies in error detection, cf. Section 5.2.2. While real word errors make simple references to semantic prototype categories and valencies within error detection rules, local case error detection in adpositional phrases requires the use of semantic prototype categories particularly in the disambiguation rules. I will then describe the system architecture for global error detection and present a detailed valency description of six rection verbs, cf. Section 5.2.3. Valency error detection requires valencies and semantic prototypes in all stages of the error detection process: disambiguation, semantic role analysis, dependency analysis, semantic role annotation and error detection itself. In Section 5.3, I will evaluate all three error types both qualitatively and quantitatively in *SIKOR*.

5.1 Background

5.1.1 General grammar checking

A grammar checker is typically distinguished from a spell-checker by the type of errors it detects and by the context it takes into account to find the error. While a spell-checker corrects non-words, i.e. words that cannot be found in the lexicon, a grammar checker corrects real words, i.e. words that can be found in the lexicon, both those that contain spelling errors and those that contain grammatical errors. The context available to a spell-checker is restricted to the word that contains the error. Therefore, the quality of the spell-checker depends on the quality and size of the lexicon against which the word is checked and from which suggested forms are picked. A grammar checker, on the other hand, looks at a context beyond the word itself in order to identify the error. In addition to dealing with linguistic errors, it can deal with violations in punctuation, capitalization, date formatting, etc. Most grammar checkers are used on top of a spell-checker, where the quality of the latter is improved by including the former, cf. *OrdRet* (Bick, 2006*a*) and *DanProof* (Bick, 2015) for Danish. Good recall and good precision are two contradicting objectives of a grammar checker. However, generally the priority is precision rather than recall as false alarms are more disturbing to the user than undetected errors. If an error is marginal, not agreed on as an error or only detectable at the expense of causing many false alarms, it might not be worth including in the grammar checker.

Uszkoreit (1996) splits up the process of grammar checking into detection, recognition, diagnosis, and correction. Detection means the identification of the erroneous segments in a given text and is according to Arppe (2000) the most difficult task in grammar checking. Recognition refers to the identification of the type of violation. Diagnosis means the identification of the source of the problem and at the same time is a prerequisite for

correction (i.e. reordering, suggesting alternative forms, deleting/adding forms). These steps can, but do not have to, be carried out separately. In *GoDivvun*, I will distinguish only between error detection (including recognition and diagnosis) and error correction.

A typical rule-based grammar checker takes a morpho-syntactically analyzed text as its input before the actual grammar checking takes place. Since detection of grammatical errors requires sentential context, a syntactic analysis and a reliable disambiguation is necessary. Disambiguation in grammar checking differs from disambiguation in parsing. In grammar checking, a full disambiguation is not necessary. In addition, the disambiguator needs to be adapted to potentially erroneous input. A regular parser assumes "*a priori* well-formed sentences" (Arppe, 2000, p.16), but in grammar checking, a disambiguator needs to pay attention to possibly malformed context. Since the disambiguation rules interact with each other, an erroneous form can lead to disambiguation errors of other forms in the sentence. The disambiguation error again can lead to missing context information for the rule that should detect the error itself. However, disambiguation rules cannot be too lax either because finding an error requires a disambiguated context.

Hagen et al. (2001) and Arppe (2000) both relax the rules in their Constraint Grammar disambiguation grammar to adapt them to potentially erroneous input, cf. also Johannessen et al. (2002). Bick (2015, p.56) uses only morphological disambiguation, but several rounds of it, i.e. "first safe error mapping followed by loose morphological disambiguation, then full error mapping followed by strict morphosyntactic disambiguation". Arppe (2000) mentions another technique, i.e. adapting the error detection rules to accept wrongly disambiguated forms when it is clear that they could be wrong analyses of correct forms. In the Basque grammar checker *XUXENg* including modules for determiner error detection (Uria, 2009) and other local error detection (Díaz de Ilarraza et al., 2010), certain disambiguation modules are simply left out (Uria, 2009). Bick (2015, p.56), on the other hand, suggests a more advanced technique, in which disambiguation is run several times both before and after error detection. For the Swedish grammar checker *GRANSKA*, Carlberger et al. (2004) suggests adding the error tag at first and remove the error tag again if the correction is identical to its original form. While rule relaxing is also applied in *GoDivvun*, the process of adapting the disambiguator includes further steps, which are explained in detail in Section 5.2.3.2.

Grammar checking devices for Finno-Ugric languages except for Finnish are still rare, cf. Table 5.1. For Finnish, there are *Kielikone*'s rule-based, but undocumented, tool *Virkku* (Pitkänen, 2006), *Voikko*,[1] and *Lingsoft*'s *FINGRC* implemented in Constraint Grammar.[2] All of the previously mentioned systems only include local, but not global, error detection. The Constraint Grammar grammar checker prototype for Estonian (Liin,

[1] `https://extensions.libreoffice.org/extensions/finnish-spell-checker-and-hyphenator-voikko` (Accessed 2017-06-27)

[2] `http://www.lingsoft.fi/print.php?lang=en&doc_id=458` (Accessed 2017-02-06)

System	Tasks	Implementation	Local synt. errors	Global synt. errors
VIRKKU (Finnish)	grammar checking	Windows	–	–
FINGRC (Finnish)	basic grammar and style checking	MS Office	agreement, missing finite verbs, tense, etc.	-
Voikko (Finnish)	spell and grammar checking, hyphenator	LibreOffice	missing verb, negation verb	-
Lightproof (Hungarian)	spelling and compound errors	LibreOffice/ OpenOffice	-	-
Estonian CG	comma checking	-	-	-

Table 5.1: Grammar checking devices for Finno-Ugric languages

2008) only corrects punctuation errors. There are further Hungarian tools for *LibreOffice*[3] and OpenOffice,[4] which are extended spell-checkers rather than syntactic error detection tools.

Apart from *FINGRC* and the Estonian system, there are several other rule-based grammar checking devices implemented in Constraint Grammar, cf. Table 5.2. *Lingsoft* distributes grammar checkers for the Scandinavian languages,[5] some of which are integrated into *MS Word*, and independent grammar checkers like *Grammatifix* (Arppe, 2000). Newer versions of *MS Word* do not contain an improved grammar checker; in fact, they have actually reduced their amount of error types. The Basque grammar and punctuation checker *XUXENg* includes a number of separate error detection modules that are preceded by Constraint Grammar modules for syntactic parsing and disambiguation. While determiner and postposition error detection is based on constraint grammar rules, agreement errors are detected by means of the UML-based tool *Saroi*, a system taking dependency trees and grammar rules as input and selecting the correct tree based on the conditions in the rules, cf. Oronoz et al. (2010) and Oronoz (2009, pp.167–169). The Danish grammar checkers *OrdRet* (for dyslexic users) (Bick, 2006*a*) and *DanProof* (Bick, 2015) use a number of additional features. *DanProof* (ca. 1,450 rules) is the most advanced of the systems discussed here, as it includes modules for spell-checking, morphological analyis/disambiguation, syntactic analysis, valency tags of the form described in Chapter 3, semantic prototype tags (cf. Chapter 4) and error detection (Bick, 2015,

[3] `https://extensions.libreoffice.org/extensions/magyar-mondatellenorzo` (Accessed 2017-06-27)

[4] `http://extensions.services.openoffice.org/project/lightproof` (Accessed 2017-02-06)

[5] `http://www2.lingsoft.fi/doc/swegc/errtypes.html` (Accessed 2017-02-06)

p.57). In addition, there are systems for learners of Catalan as a second language (*ALLES* Advanced Long-distance Language Learning System (Badia et al., 2004)) and Esperanto (*Lingvohelpilo* (Petrović, 2009)) that implement their error detection rules in Constraint Grammar and use Constraint Grammar syntactic parsing. *Lingvohelpilo* is based on a full vislcg3-parser, *EspGram*, including syntactic analysis, disambiguation, and dependencies. While all of the systems include a somewhat modified disambiguation, partial dependencies are only used in two of the systems. A small set of semantic categories is only used in one of the systems. However, none of the systems makes use of valency information beyond simple transitivity information.

While Table 5.2 shows the kind of linguistic information that is applied in grammar checking, Table 5.3 shows which kind of valency error detection is performed by the respective grammar checkers. Valency errors have been defined in different ways. For Fliedner (2001, p.16), for example, valency errors are missing or redundant governed elements. Wedbjer Rambell (1999) includes erroneous use of governors in her definition of valency errors: e.g. transitive vs. intransitive verb use and erroneous passive verb use. In languages with larger case sets and different infinitival constructions, valency errors can also be case errors, erroneous use of the infinitival form, missing parts in the infinitival construction, etc. Here I will focus on missing and redundant governed elements and errors regarding the form of governed elements. In Wedbjer Rambell's (1999)'s error classification for Swedish, valency errors make up the second largest error type after noun phrase (predominantly agreement) errors, cf. Wedbjer Rambell (1999, p.46). She mentions missing infinitive markers, prepositions, noun phrases and errors in infinitive constructions as possible valency errors. In ex. (4-a), the noun *nytta* 'use' is lacking a prepositional phrase as its argument, i.e. *av honom* 'of him', and is therefore considered ungrammatical. Ex. (4-b) includes a prepositional phrase and is therefore considered grammatical.

(4) a. *Tror inte att jag haft någon mer nytta.
think not that I had any more use
'I don't think that I had any more use.' Wedbjer Rambell (1999, p.49)

b. Tror inte att jag haft någon mer nytta **av honom**.
think not that I had any more use of him
'I don't think that I had any more use of him.' (Ibid.)

Grammar checkers that correct valency errors are rare. The rule-based grammar checker prototype *Scripsi* (for learners of English as a second language) described in Catt's (1988) Master's thesis detects valency errors in very simple English sentences produced by French and Chinese users. It includes case and prepositional errors, e.g. the use of objective case instead of nominative case in ex. (5-a), and the use of a direct object instead of a prepositional phrase in ex. (5-b). Unfortunately, Catt (1988) does not include more complex examples or an evaluation of the rules, suggesting that the system does not

System	SYNTACTIC ERRORS	TECHNIQUES				
		Disambiguation	Explicit dependencies	Semantic prototypes	Valencies	Semantic roles
Grammatifix/ SWEGRC (Swedish)	agreement, infinitive after preposition, constituent order	√	–	–	–	–
FINGRC (Finnish)	agreement (subject-predicate, NP-internal), missing finite verb main clause	√	–	–	–	–
DANGRC (Danish)	agreement (subject-complement, ...), infinitive marker	√	–	–	–	–
NGC/NOBGRC (Norwegian)	agreement (subject-verb, NP-internal), word order	√	√	–	–	–
OrdRet (Danish)	combined spell- and grammar checker, mostly real word errors	√	–	√	√	–
DanProof (Danish)	combined spell- and grammar checker, agreement (subject-subject complement, NP-internal, infinitive marker, subject/object case errors)	√	–	√	√	–
XUXENg (Basque)	agreement (subject, object-verb) complex postpositions determiners	√ √ √	√ ? –	– √ –	– – –	– – –
ALLES (Catalan) Badia et al. (2004)	agreement (NP-internal, subject-verb), word order, valency (direct vs. indirect object)	√	–	–	–	–
Lingvohelpilo (Esperanto)	missing accusative marking, word order, transitivity, tense	√	√	√	√	–

Table 5.2: An overview of Constraint Grammar-based grammar checkers

SYSTEM	VALENCY ERROR TYPES						
	Error regarding the infinitive marker	Verb form errors in non-finite clauses	Missing subjects/objects	Wrong case after preposition	Nominative/accusative error	Direct vs. indirect object error	Errors in finite subclauses
Swedish *Granska*	√	√	–	–	–	–	–
Swedish *Scarrie*	√	√	–	–	–	–	–
Finnish *FINGRC*	–	–	–	–	√	–	√
Swedish *SWEGRC/ Grammatifix*	√	–	–	–	–	–	–
Norwegian *NOBGRC*	√	–	–	√	√	–	–
Danish *DANGRC*	√	–	–	√	–	–	–
Catalan *ALLES*	–	–	–	–	–	√	–
Basque *XUXENg*	–	–	–	–	√	–	–
Danish *OrdRet*	√	√	√	–	–	–	–
Danish *DanProof*	√	–	–	–	√	–	–
Swedish *FiniteCheck*	√	√	–	–	√	–	√
Esperanto *Lingvohelpilo*	–	–	√	√	–	–	–
English *Scripsi*	–	√	–	–	√	–	√

Table 5.3: Valency error detection in grammar checking

work with free input of running text.

(5) a. *__Him__ reads the books. (Catt, 1988, p.55)
b. *This child disobeys __to__ his father. (Catt, 1988, p.58)

While a number of systems intend to detect simple valency errors, only very few of them apply a global syntactic analysis including valency, dependency, and semantic prototype information, which is necessary to find valency errors in running text. However, there are attempts to include semantic categories in real word error detection as some real word errors result in valency errors. Pedler (2007) uses semantic categories derived from *WordNet* in probabilistic real word error detection. She calculates the probability of members of a confusion pair, e.g. (*diary; dairy*) and (*hope; hole*), cf. ex. (6-a), co-occurring with a noun of a certain semantic category in a two- or three-word distance. However, she does not apply a syntactic analysis, and she concludes that semantic categories do not improve the performance of her spell-checker significantly. Banu and Kumar's (2004) real word error detection is based on an algorithm to calculate semantic selection restrictions for governors that appear with arguments that are members of confusion pairs (e.g.

dessert; desert), cf. ex. (6-b) and ex. (6-c). There is probably a connection between their poor precision of 10% and recall of 19% and the lack of syntactic analysis. Selection restrictions alone are not of much use in grammatical error detection. They need to be paired with syntactic preferences of verbs and a syntactic analysis of the whole sentence.

(6) a. It is my sincere **hole (hope)** that you will recover swiftly. (Pedler, 2007, p.39)
b. The cook served the **dessert**. (Banu and Kumar, 2004, p.131)
c. *The cook served the **desert**. (Ibid.)

Bick (2015) includes semantic categories and valencies (cf. Section 3.1) in *DanProof*. His system achieves a recall of 65.1%, a precision of 91.7% and an F-score of 76.1%: its performance is significantly higher than that of the *DANGRC*-based grammar checker in *MS Word 2007* (recall 20.8%, precision 54.6%, F-score 30.1%) (Bick, 2015, p.60). To my knowledge, his approach is the only one that makes use of semantic categories and valencies in error detection. Even though researchers and developers of grammar checkers frequently mention the necessity of systematic encoding of semantic and valency information in the lexicon to perform successful real word error and valency error detection (Fliedner, 2001, p.173), other commercial grammar checkers do not make use of valencies. Hagen and Lane (2001) mention that missing words cannot be found by a grammar checker without semantic knowledge. Even seemingly local syntactic errors like determiner errors often require a global syntactic analysis or fine-grained semantic categories (Díaz de Ilarraza et al., 2010).

5.1.2 North Sámi grammar checking

The North Sámi grammar checker *GoDivvun* is based on a prototype of *grammarchecker.cg3* (Wiechetek, 2012). The target group are native speakers of North Sámi who write or publish Sámi text for personal or professional use. Restricting the target group is important to achieve good precision and recall as the types of errors largely depend on the language proficiency of the writers. The assumption is that while most of the grammatical errors of language learners are proficiency errors, native speakers tend to make more typographical or copy-paste errors resulting in grammatical errors.

GoDivvun is part of *Giella-sme* and has access to the same lexica and descriptive (in addition to the normative) morphological analyzers and compilers. *GoDivvun* contains rule-based Constraint Grammar modules for error detection and correction (*grammarchecker.cg3*) and syntactic analysis/disambiguation (*disambiguator.cg3*) and is compiled with the vislcg3-compiler.[6] Finite state transducer (fst)-compilers are used for morphological analysis.[7] It uses the descriptive morphological analyzer *tokeniser-gramcheck-*

[6] `http://beta.visl.sdu.dk/cg3.html` (Accessed 2017-02-06)
[7] `https://hfst.github.io/` (Accessed 2017-02-06)

gt-desc.pmhfst. The morphologically analyzed text serves as an input for the disambiguator and the error detection module. A newer version of the grammar checker contains a number of other modules for tokenization and generation, e.g. the simple disambiguation module *mwe-dis.cg3*, which performs compound-error detection and is applied before disambiguation.[8] The combination of finite state automatons and constraint grammar has been successful for the previously mentioned grammar checkers of Swedish, Finnish, Danish and Norwegian bokmål (and newer versions).

GoDivvun can be tested as an online tool and as a command-line tool and is in the process of being integrated in *LibreOffice*.[9] Figure 5.1 illustrates the valency error in ex. (3-b). The valency error *sistesihkkelastima* 'indoor biking (Acc.)' is identified via a blue line below the word that includes the error (as opposed to a red line marking a spelling error). A click on the error produces a message with a diagnosis and a suggestion of the correct form. The error message includes the diagnosis, cf. ex. (7), and a suggestion of the correct form, i.e. *sistesihkkelastimii* 'indoor biking (Ill.)'.

(7) Iskka geavahit illatiivahámi alege akkusatiivahámi
'Try to use the illative form instead of the accusative form'

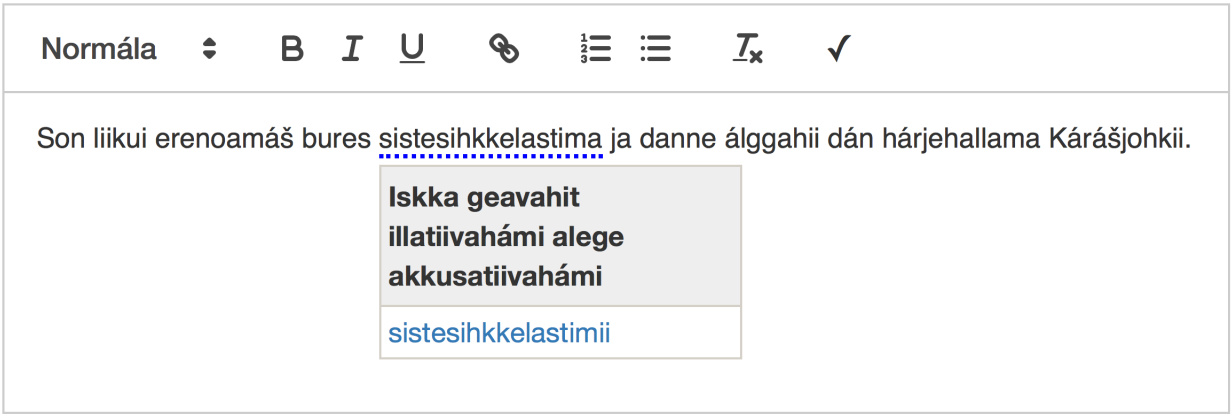

Figure 5.1: Error detection and correction by the *GoDivvun* online tool

The error detection and correction module *grammarchecker.cg3* is based on Constraint Grammar *ADD*-rules for error detection and *COPY*-rules for error correction, cf. also Chapter 2. The *ADD*-rule in Figure 5.2 adds the error tag *&msyn-valency-ill-acc* to an accusative form (*Acc*) in a particular syntactic context. The *COPY*-rule, on the other hand, replaces the accusative tag with an illative tag in a given tag sequence thereby producing the input to the normative morphological generator, *generator-gt-norm.hfstol*,

[8] version r156914 (Accessed 2017-09-13)
[9] `http://gtweb.uit.no/gc/` (Accessed 2017-06-28)

which generates the correct forms based on the tag sequence given by the *COPY*-rules. The suggested form is marked by a special tag, &*SUGGEST*. The online tool further matches each error tag with specific feedback that can be accessed by the user.

```
ADD (&msyn-valency-ill-acc) TARGET Acc IF SYNTACTIC CONTEXT ;

COPY (Ill &SUGGEST) EXCEPT (Acc) TARGET &msyn-valency-ill-acc ;
```

Figure 5.2: Simplified *vislcg3* error detection and correction rules

GoDivvun distinguishes between six general error types: real word errors, compound errors, morpho-syntactic errors, syntactic errors, lexical errors, and punctuation errors, cf. Table 5.4. When classifying error types, one can base the classification on causes or outcome of the error. While a cause can be a typo, the outcome can be a real word error or a syntactic error. The error types in *GoDivvun* are mainly based on the outcome and the analysis that is necessary to identify the error. While this is most important for error detection, identifying the causes of the error is most relevant for the feedback given to the writer. The form *várri* 'mountain (Nom.)' instead of *vári* 'mountain (Gen.; Acc.)' can be a grammatical error, i.e. use of nominative instead of genitive/accusative case, and may be a result of lacking language proficiency in the use of these cases. However, for a native speaker it is most likely to be either a typo (single consonant rather than double consonant) or based on the phonetics of their local dialect (i.e. missing accusative plural ending *-id*) as it is silent. While the syntactic context can be used to produce the correct form, i.e. either accusative plural or accusative singular, the feedback needs to be given based on the cause of the error, i.e. double-consonant error rather than case error for a grammar checker with a target group of native speakers.

5.2 Valencies and semantic prototypes in *GoDivvun*

Finding a (global) syntactic error in a syntactically unreliable context is one of the most challenging tasks in grammar checking. As syntactic analysis with grammatically correct input is already challenging because of morphological homonymy and syntactic ambiguity, syntactic analysis of potentially ill-formed input needs to take into account both spelling and grammatical errors as well, adding to the number of possible readings. While syntactic analysis only can be insufficient in syntactic error detection, valencies and semantic prototype categories can make the context more reliable and facilitate the analysis of the sentence and error detection. Adding another level to the linguistic analysis makes it more robust.

ERROR TYPE	SENTENCE WITH AN ERROR	CORRECTED SENTENCE
Real word errors	lea **áibbaš** dárbbašlaš ahte . . . 'it is miss necessary that . . . '	lea **áibbas** dárbbašlaš ahte . . . 'it is completely necessary that . . . '
Compound errors	lean álo liikon [. . .] **jurddašan vuohkái** 'I have always liked the way of thinking'	lean álo liikon [. . .] **jurddašan-vuohkái** 'I have always liked the way of thinking'
Lexical errors	Jus telef[o]vnna **badjel** máksá gir[o], de . . . 'If one pays the invoice over the telephone, then . . . '	Jus telef[o]vnna **bokte** máksá gir[o], de . . . 'If one pays the invoice over the telephone, then . . . '
Morpho-syntactic errors	ollosat liikojedje **šuoŋa** ja **muitalusa** 'many liked the song (Acc.) and the story (Acc.)'	ollosat liikojedje **šukŋii** ja **muitalussii** 'many liked the song (Ill.) and the story (Ill.)'
Syntactic errors	Vars, liiko maid gullá. 'Vars likes what s/he hears'	Vars, liiko **dasa** maid gullá. 'Vars likes what s/he hears'
Punctuation errors (e.g. missing commata)	Son livčči gal viššal **{ }** muhto sus ii leat goassege dilli bargat. 'S/he would be diligent but s/he never has the time to work.'	Son livčči gal viššal**,** muhto sus ii leat goassege dilli bargat. 'S/he would be diligent, but s/he never has the time to work.'

Table 5.4: The six general error types in *GoDivvun*

In *GoDivvun*, valency and semantics are used in disambiguation and error detection of grammatical error types, i.e. real word, lexical, compound, morpho-syntactic, syntactic, and punctuation errors, both within local and global rules. While local rules refer to semantic prototypes and valencies directly, more global rules (e.g. valency error detection rules) have access to a dependency and semantic role analysis and can refer to semantic prototypes and valencies in specific argument positions. In the following, I will show error types of different degrees of locality/globality in relation to the type of linguistic information and depth of linguistic analysis needed. I will start with very local error detection, which can be resolved without testing any syntactic context, purely based on the fact that there is a similar and better alternative to the form. Secondly, I will look at real word errors that are based on idiosyncratic relations between the confusion pair members and which typically require local error detection. Thirdly, I will discuss more systematic errors in local contexts, which do not involve a confusion pair based on a lemma, but can be reduced to morpho-syntactic tag sequences like case errors in adpositional phrases. Lastly, I will focus on global errors and their detection, in particular valency errors. These are based on the analysis of the whole sentence, and their detection includes several modules that will be explained in detail, i.e. disambiguation, dependency analysis, semantic role mapping and finally valency error detection and correction.

5.2.1 Very local error detection

While a spell-checker generally detects non-words, a grammar checker detects real words that appear in the wrong context. In her analysis of North Sámi text by proficient writers (40,736 words), Antonsen (2013, pp.7–8) finds that there are 4% spelling errors in words, based on both non-words and real words, of which 78% are identified by the spell-checker. Most of the 22% undetected errors are based on real words (some are norm-specific difficulties). They are not detectable based on the word context only, and are therefore left to the grammar checker. For an automatic spell-checker, a non-word is any form that cannot be found in the lexicon. As the lexicon cannot be expected to be complete, existing words can also erroneously be marked by the spell-checker. While lemmata are listed, word formation processes such as compounding, inflection and derivation are modelled by a morphological analyzer. *GoDivvun* includes both a normative and a descriptive analyzer. The normative morphological analyzer only recognizes word forms that are accepted by the norm. Other commonly used forms that are not listed as normative forms in the lexicon are considered non-words. The descriptive analyzer can store subforms, erroneous forms, and dialectal forms and tag them both with error tags and dialect tags. *GoDivvun* applies descriptive morphological analyzers in order to provide as much context information as possible to find an error. Morphological processes can produce forms that are possible from a grammatical point of view but rare in *SIKOR*, cf. Table 5.5. The normative morphological analyzer *sme-norm.fst*[10] provides an analysis for the following forms, some of which are considered non-words by a newer version of the normative morphological analyzer *analyser-gt-norm.hfstol*.[11] Compounding produces nonsense compounds like *nammalassii* 'name threshold (Ill.)' in ex. (8-a) and *sihkarastit* 'secure have time' in ex. (8-b). However, *SIKOR* does not provide any examples in which the form is correct, i.e. all examples are real word errors. Derivation can also produce a number of nonsense forms such as the denominal derivation *billehuvvet* 'become without a flute' in ex. (8-c). There are further sequences of passive, causative and frequentative derivations that produce forms that are only real word errors in *SIKOR*, e.g. *čohkkohalle* 'sharpen (Caus. Pass. Freq. Prt. 3Pl.)' instead of *čohkohalle* 'they sit comfortably' in ex. (8-d). A number of inflectional forms are also rare or restricted, i.e. the biblical connegative form as in *vahko* 'become stronger (Imp. ConnegII.)' in ex. (8-f), which is confused with the frequent noun *vahku* 'week'. A second example involves the possessive suffix first person singular forms, e.g. *bidjon* 'den (Px1sg.)', cf. ex. (8-e). Both forms are only real word errors in *SIKOR*.

In her study of rare inflectional forms, Antonsen (2014) suggests restricting the morphological analyzer as some inflected forms can be found only in very restricted morpho-

[10] version r53455 (Accessed 2012-01-31)
[11] version r106600 (Accessed 2014-12-23)

Rare form	***SIKOR***	**Correction**	***SIKOR***
COMPOUNDING			
nammalassii namma 'name' + lassa 'threshold' (Sg. Ill.)	537	*namalassii* (Adv.) 'namely'	4,198
ruovttuluotta ruovttu 'home' (Gen.) + luodda 'track' (N. Sg. Gen.)	174	*ruovttoluotta* ruovttoluotta (Adv.) 'back'	6,066
sihkarastit sihkar 'secure (A.)' + astat 'have time' (V. Imp. Pl2.)	432	*sihkkarastit* sihkkarastit (Inf.) 'secure'	6,967
DERIVATION			
inflected forms of *billehuvvat* bille 'flute' (deverbal) (Denom. Inf.) 'become without a flute'	8	inflected forms of *billahuvvat* billahuvvat (V. Inf.) 'be destroyed'	416
inflected forms of *čohkkohallat* čohkat 'sharpen' (V. Caus. Pass. Freq. Inf.)	30	*čohkohallat* čohkohallat 'sit comfortably' (V. Inf.)	292
INFLECTION			
bidjon biedju 'den' (N. Sg. Acc. PxSg1.)	136	*biddjon* bidjat 'put' (Pass. V. IV. PrfPrc.)	2,828
vahko vahkat 'become stronger' (V. IV. Imprt. ConNegII.)	32	*vahku* vahkku 'week' (N. Sg. Gen.)	4,714

Table 5.5: Real word errors in *Giella-sme* that are caused by morphological overgeneration

logical contexts in *SIKOR*. She notes that these forms are overgenerations and cover up for spelling errors in frequent existing forms. This strategy is an alternative to specifying error detection rules for these forms. While it simplifies error detection and morpho-syntactic analysis by reducing homonymy, the advantage of specifying error detection rules is that these forms can still receive an analysis. Only the forms that are similar to a form that is a better alternative (i.e. a confusion pair counterpart) are marked as errors. These include various possessive forms, certain imperative forms, etc.

(8) a. *. . . seamma doibmii maid son ieš bargá,
. . . same activity.ILL which s/he herself;himself works,
nammalassii silba- ja čoarvedáiddárin.
name.threshold.ILL silver- and horn-artist.ESS
'. . . same profession that s/he carries out, namely silver and horn artist.'

b. Dat lea prošeakta mas lea ulbmil hukset ovttasbargoguimmiid našunála
this is project that has objective build cooperation.partners national
ja gaskariikkalaš dásis, **sihkarastit** riektevuođu ja vuođu
and international level, secure.have.time.INF law.basis.ACC and basis.ACC
ovdánit mearrasámi guovlluin.
develop.INF coastal sámi area.LOC.PL
'It is a project whose aim is to establish cooperation partners on a national and

international level to preserve the legal foundation and promote development in coastal Sámi areas.'

c. Biepmut eai ábut vurkko[d]uvvot nu guhk[á] ahte
foods not pay.off store.PASS.PRS.3PL as long as
billehuvvet.
flute.CAR.DENOM.PRS.3PL
'It is not worth it to store the food a long time until it goes bad.'

d. Muhtomin ledje golbma olbmo geat **čohkkohalle**
sometimes were three people that sharpen.CAUS.PASS.FREQ.PRT.3PL
beaŋkkas.
bench.LOC
'Sometimes there were three people that sat comfortably on the bench.'

e. Ja daid oktavuođas leat digaštallamat mat gusket ekonomalaš
and this context.LOC are debates that concern economical
váikkuhusaide **bidjon** váttisvuohtan.
impact.ILL.PL den.PXSG1 problem.ESS
'And in this context, discussions that concern the economical impact are presented as problems.'

f. Mannan **vahko** čoagganedje 160 olbmo ...
last become.stronger.IMP.CONNEGII gathered 160 people ...
'Last week 160 people gathered ...'

If treated by the grammar checker rather than the morphological analyzer, these forms can be discarded without any syntactic context conditions, i.e. the error detection rule tags any instance of the word as an error and replaces it with the desired form without referring to a syntactic context. The following rule in *grammarchecker.cg3* marks the compound *ruovttuluodda* 'home track' as an error if it appears in genitive or accusative singular case. The rule does not specify any syntactic or semantic context. It only refers to a similar and better alternative to the target, i.e. a confusion pair counterpart such as the adverb *ruovttoluotta* 'back'.

```
ADD (&real-ruovttoluotta) TARGET N IF (0 ("ruovttu#luodda") LINK 0 Gen OR Acc);
```

Rules of this type are specified in the beginning of *grammarchecker.cg3* and are applied directly after disambiguation. While these errors can be ruled out by the grammar checker based on the word context only, below I will focus on syntactic errors that require both local and global syntactic error detection.

5.2.2 Local error detection

Local grammatical error detection that is based on context conditions typically requires the identification of a smaller part of the sentence, i.e. more than just the targeted word, but not the whole sentence. This section discusses how semantic tags can be used to resolve local syntactic errors. Whereas a spell-checker takes into account only the word context itself, local error detection rules refer to local syntactic contexts. While global error detection rules require an analysis of the whole sentence, local error detection rules mostly refer to adjacent word forms and elements within the same noun phrase or adpositional phrase. Local error detection rules resolve both idiosyncratic and systematic local errors. These can be e.g. noun phrase internal agreement errors or case errors in adpositional phrases. Local error detection is performed after syntactic analysis and disambiguation in the beginning of *grammarchecker.cg3* (after very local error detection), cf. Figure 5.3. Local syntactic errors can be lexical, real word, morpho-syntactic, and syntactic errors. However, semantic prototypes and valencies are predominantly used in real word error detection, morpho-syntactic error detection, and lexical error detection. Most syntactic errors that regard e.g. agreement or comparation of adjectives can be resolved by means of syntactic and morphological constraints only, i.e. without semantic prototypes and valency information. Lexical error detection rules generally concern the erroneous use of a lexeme, such as *badjel* 'over' instead of *bokte* 'via' in ex. (9-a)–(9-b). Choosing the correct postposition or adverb often depends on the semantic category of the nominal head, which is why many of the lexical rules refer to semantic prototype categories. However, because of their marginality in *GoDivvun* and their idiosyncrasy, they will not be discussed further in this chapter.

(9) a. *Jus telef[o]vnna **badjel** máksá gir[o], de ...
if telephone.GEN over pay invoice.ACC, then ...
'If one pays the invoice over the telephone, then ...'

b. Jus telef[o]vnna **bokte** máksá gir[o], de ...
if telephone.GEN over pay invoice.ACC, then ...
'If one pays the invoice over the telephone, then ...'

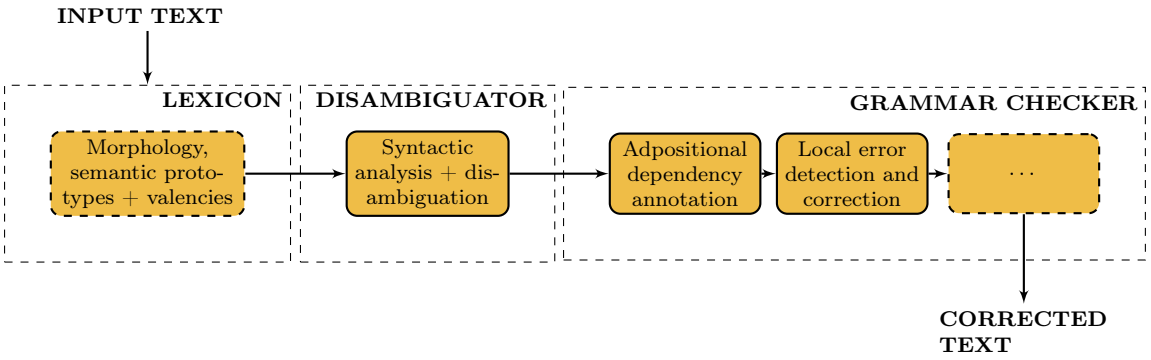

Figure 5.3: The system architecture of all local error detection in *GoDivvun*

5.2.2.1 Real word errors

Misspellings that result in real words make up ∼22% of undetected spelling errors in a study of North Sámi by Antonsen (2013) and are very common. In the discussion that follows, I will distinguish between spelling errors that result in different words with different lemmata and parts of speech, and spelling errors that result in different forms of the same lemmata. While the first type is based on an idiosyncratic relation between two forms, the second type is typically based on a systematic relation between a whole set of lemmata that can be generalized by different morphological tag sequences. Here, I will only consider the first type a "real word error". As opposed to Antonsen (2013), the second type will be considered either a morpho-syntactic or a syntactic error, and will not be discussed in this section.

A double consonant error in the consonant center like *iskkan* 'try (1Sg.)' in ex. (10) results in a systematic syntactic error. Here, the form should only have a single consonant: it should be *iskan* 'try (PrfPrc.)' instead. This error is possible for all verbs with double consonants in their consonant center ending in *-at*. They are specified in the set *DOUBLE-CONSONANT-AT-VERBS* below:

```
LIST DOUBLE-CONSONANT-AT-VERBS = (".*hkat"r V) (".*rtat"r V)
                                 (".*skat"r V) (".*tkat"r V);
```

The following rule adds the error tag *&syn-prfprc-not-prssg1* to verbs with double consonants ending in *-at* in the first person singular present tense form unless they co-occur with a pronoun in first person singular in nominative case.

```
ADD (&syn-prfprc-not-prssg1) TARGET DOUBLE-CONSONANT-AT-VERBS IF (0 (Ind Prs Sg1))
(NEGATE *-1 ("mun" Pron Pers Sg1 Nom) BARRIER GRAMCHK-S-BOUNDARY)
(*-1 ("leat") BARRIER GRAMCHK-S-BOUNDARY);
```

(10) *Isaksen lohká iežaset **iskkan** beassážiid áigge buollin
Isaksen claims oneself.ACC.PXPL3 try.PRS.1SG Easter.GEN time.GEN fire

sihkkarvuođa ...
safety.system.ACC ...
'Isaksen claims that he tried the fire safety system during Easter time ...'

Real word errors are based on confusion pairs. Confusion pairs consist of at least two similar forms, of which one can, but does not have to, be more frequent than the other. The reason that confusion pairs are not grammatically predictable is that they often result from phonological or graphemic similarities between two related or unrelated forms. Different types of real word errors and the type of spelling error causing them are shown in Table 5.6. Causes for real word errors are typically typos of the following kinds: accent errors, double consonant errors, diphthong errors, vowel errors and other errors that are caused by the divergence between phonological form and its representation in writing (e.g. caused by an unpronounced endings in certain dialects). Some typos are caused by the proximity of the letters on the keyboard. Most of the errors have an edit distance of 1 or 2 (i.e. 1 or 2 characters need to be changed). However, phonological errors caused by an unpronounced ending can have a larger edit distance.

Table 5.7 presents a number of confusion pairs along with their frequencies in *SIKOR* and the frequencies of real word errors related to them. These will be evaluated in Section 5.3. While some confusion pairs have one rare and one frequent member, others have two equally frequent members. There are three verb pairs (sometimes groups of verb forms based on the same lemma), two verb-noun pairs, and one adverb-noun pair. The noun-adverb pair is the following: (*várra*;*varra*). The verb pairs include several confused forms of *áddet* 'understand' and *addit* 'give', i.e. (*áddet*;*addet*), (*ádde*;*adde*), and (*ádden*;*adden*). Confused forms of *sáhtašit* 'give a ride' and *sáhttit* 'can' include any inflected form of *sáhtašit* 'give a ride' and any inflected conditional form of *sáhttit* 'can'. Confused forms of *čohkket* 'collect' and *čohkat* 'sharpen' include the following confused forms: (*čohkket*;*čohket*), (*čohkke*;*čohke*), and (*čohkken*;*čohken*). The verb-noun pairs are the following: (*čohkká**;*čohkka**) and (*biddjui*;*bidjui*). Confused forms of *čohkkát* 'sit' and *čohkka* 'mountain top' include the following confused forms: (*čohkká*;*čohkka*), (*čohkkán*;*čohkkan*), (*čohkkát*;*čohkkat*), (*čohkkába*;*čohkkaba*), (*čohkkába*;*čohkkaba*), and (*čohkkame*;*čohkkáme*).

Most real word errors are found for forms that should be forms beginning with *sáhtáš-* 'can' instead of forms beginning with *sáhtaš-* 'give a ride' (500) and forms that should be *várra* 'maybe', instead of *varra* 'blood' (164). Their counterparts, i.e. forms of *sáhtaš-* and *várra* 'maybe' do not have any instances of real word errors. However, there are correctly spelled examples that show real context of forms beginning with *sáhtaš-* (26) and *varra* (266). Other frequent confusion pairs (*jamas;jámas*) with high error rates have been discarded for this study because one of their confusion pair members does not have a single correctly spelled example. For the confusion pair (*čohkke**;*čohke**), the rare forms *čohket* 'sharpen (Prs. 3Pl.)', *čohke* 'sharpen (Prs. 1Du.)', etc. make up only 0.38% of

Confusion pair member A	**Confusion pair member B**	**Morph. tags**	**Sem. tags**	**Val. tags**
Consonant errors				
biddjui 'put (Pass. Prt. 3Sg.)'	*bidjui* 'den (Ill.)'	X	X	X
biddjon 'put (Pass. Prs. 1Sg.)'	*bidjon* 'den (Nom. PxSg1.)'	X	-	-
dohko 'there'	*dohkko* 'clump (Prt. 3Pl.)'	X	-	X
duodjái 'handicraft (Ill.)'	*duoddjái* 'productive'	X	-	X
ádjáid 'grandparents (Acc. Pl.)'	*áddjáid* 'time-consuming (Acc. Pl.)'	X	-	-
lohkat 'read'	*lohkkat* 'lock (Nom. Sg. PxSg2.)'	X	-	-
measta 'almost (Po)'	*meastta* 'puree, mush'	X	-	-
vuodjit 'drive'	*vuoddjit* 'drivers (Nom. Pl.)'	X	-	-
nuoran 'blunt (Prs. Sg1.)'	*nuorain* 'youth (Com.)'	X	-	X
Accent errors				
joatkkan 'continuation (Ess.)'	*joatkkán* 'continue (Prs. 1Sg.)'	X	-	X
jámas 'to death; dead'	*jamas* 'noise (Loc.)'	X	-	X
čohkká 'sit (3Sg. Prs.)'	*čohkka* 'mountain top'		X	-
čohkat 'sharpen'	*čohkkát* 'sit'	X	X	-
vahku 'week (Gen.)'	*váhku* 'fish broth (Nom.)'	X	-	-
várra 'possibly'	*varra* 'blood, danger'	X	X	-
Vowel errors				
álo 'always (Adv.)'	*alu* 'of the height (Po)'	X	X	-
dihtii 'because of (Po.)'	*dihti* 'sparrowhawk (Gen.)'	X	X	-
áddet 'understand'	*addet* 'give (Prs. 3Pl.)'	X	X	-
jearrá 'ask (Prs. 3Sg.)'	*jeara* 'yeast'	X	-	-
fitnet 'visit (Prs. 3Pl.)'	*fidnet* 'get (Inf.)'	X	-	-
Aspiration errors				
atte 'give (Prs. ConNeg.)'	*ahte* 'that'	X	-	-
dakko 'there'	*dahko* 'do (Imprt. ConNegII.)'	X	-	-
Phonological errors				
vuvdiid 'seller (Acc. Pl.)'	*vuvdii* 'abdominal cavity (Ill.)'	X	-	-
Type errors				
bokte 'via'	*bohte* 'come (Prs. 1Du.)'	X	X	-
Vowel + consonant errors				
lasse '(to) lock (Prs. 3Sg.)'	*láse* 'window'	X	-	-

Table 5.6: Real word errors in *Giella-sme* according to their cause

Confusion pair (common; rare)	***SIKOR***	**Common real error**	**Rare real error**
Adverb vs. noun			
várra; varra ‘maybe, care; blood’	5,020 (varra:430; várra:4,590)	varra: 164	várra: 0
Verb vs. verb			
adde*; ádde* ‘give; understand’	5,703 (ádde*:3,474; adde*:2,229)	adde*:22	ádde*:10
sáhtáš*; sáhtaš* ‘can; give a ride’	4,334 (sáhtaš*:526; sáhtáš*:3,818)	sáhtaš*:500	sáhtáš*:0
čohkke*; čohke* ‘collect; sharpen’	2,601 (čohke*:10, čohkke*:2,591)	čohke*:9	čohkke*:0
Verb vs. noun			
biddjui; bidjui ‘put (Pass. Prt. 3Sg.); den (Ill.)’	611 (bidjui:66, biddjui:545)	bidjui:29	biddjui:0
čohkká*; čohkka* ‘sit; mountain top’	1,810 (čohkka*:69, čohkká*:1,741)	čohkka*:56	čohkká*:0

Table 5.7: The distribution of correct instances and real word errors in confusion pairs in *SIKOR*

all the occurrences of the confusion pair. For the confusion pair (*ádde*;adde**), on the other hand, the distribution is more even. Forms of *áddet* ‘understand’ make up 60.92%, and forms of *addit* ‘give (Prs. 3Pl.)’ make up 39.08%. Confusion pairs with an equal distribution of both forms require more careful rules than confusion pairs with one common and one rare member. In addition, the (morphological, syntactic, semantic, and valency-related) similarity of the forms and their contexts are relevant for the construction of error detection rules. While real word error detection rules frequently use semantic tags, there are two ways for valency to be relevant. The first type regards governors with different valencies that appear in otherwise similar syntactic contexts, as in the case of, for example, the verbs *čohkket* ‘collect’, *čohkat* ‘sharpen’, and *čohkkát* ‘sit’. The second type regards confusion pairs, where one member is the potential argument of a governor (typically in adverbial case), e.g. the illative *bidjui* ‘den (Ill.)’, which is confused with *biddjui* ‘put (Pass. Prt. 3Sg.)’. The real word error rules for (*biddjui;bidjui*) and (*varra;várra*) deal with confusion pair members that are not of the same part of speech. In the case of the confusables *biddjui* ‘put (Pass. Prt. 3Sg.)’ and *bidjui* ‘den (Ill.)’, *biddjui* is nine times as frequent as *bidjui* ‘den (Ill.)’. The forms are not related to each other, have a different part of speech and also differ from each other syntactically and semantically. Even if the forms differ from each other in many respects, the real word error itself can lead to disambiguation errors of other forms in the sentence. This can lead to an analysis of the sentence in which the erroneous form may seem correct. The form *bidjui* ‘den (Ill.)’ is

mostly used in a context of animals (cf. ex. (11-b)) or also humans that can enter in a den. The following simplified rule tests therefore if a noun of the animal prototype can be found in the close context (*NEGATE *-1 Sem/Ani*). The noun *biedju* 'den' is of the *place* prototype and in illative case a potential DESTINATION. Therefore, the rule tests for a verb with a DESTINATION in illative case in its valency (*<TH-Acc-Any><DE-Ill-Any> OR <TH-Acc-Any><SO-Loc-Any><DE-Ill-Any> OR <DE-Ill-Plc>*), which can be satisfied by a noun in illative case unless there is a finite verb in the context (*NEGATE *0 VFIN*) such as *bidjat johtui* 'launch', *bidjat eret* 'put away', etc.

```
ADD (&real-biddjui) TARGET ("bidju" N Sg Ill) OR ("biedju" N Sg Ill)(NEGATE *0 VFIN)
(NEGATE *0 <TH-Acc-Any><DE-Ill-Any> OR <TH-Acc-Any><SO-Loc-Any><DE-Ill-Any>
OR <DE-Ill-Plc>)(NEGATE *-1 Sem/Ani);
```

As neither a member of the animal prototype nor a verb with a DESTINATION in illative can be found in the context of *bidjui* in ex. (11-a), the rule annotates a real word error to *bidjui* 'den (Ill.)'.

(11) a. *Dat maid **bidjui** gulaskuddamii boazoorohahkii ...
this which den.ILL hearing.ILL reindeer.district.ILL ...
'This which was referred to the reindeer district for discussion ...'

b. Muhto rieban láve čoaggit stuorát návddiid bázahusaid ja doalvut
but fox use.to collect bigger predator carcasses and bring
daid **bidjui**.
it.ACC.PL den.ILL
'But the fox usually collects the carcasses of bigger predators and brings them into the den.'

Real word error detection for the forms *áddet* 'understand' and *addet* 'give (Prs. 3Pl.)' is more challenging as the forms are almost equally distributed and morpho-syntactically similar. The forms are unrelated, but have the same part of speech. Therefore, a set of quite a few rules referring to both semantic prototype tags and valency tags is necessary to detect possible errors. While the form *addet* can be a second person singular past tense form, a present tense third person plural form or a second person plural imperative form of the verb addit 'give' (cf. ex. (12-a)), it can also be a real word error for the infinitive, a third person present tense form or a second person plural imperative form of the verb *áddet* 'understand' (cf. ex. (12-b)).

(12) a. Galggat servvoštallat dakkár ustibiiguin geat **addet**
should hang.out that.kind.of friend.COM.PL that give.PRS.3PL
dutnje movtta.
you.ILL encouragement.ACC
'You should hang out with the kind of friends that encourage you.'

b. *Muitalusat **áddet** midjiide jurddašeami ...
stories understand we.ILL thinking.ACC ...

'The stories make us think . . . '

The following simplified rule &*real-addit* adds an error tag to the form *áddet* if there is a plural noun (i.e. a potential subject), an object (*@<OBJ OR @-F<OBJ*) and a human illative (*Ill + Sem/Hum*, . . .) anywhere in the sentence.

```
ADD (&real-addit) TARGET ("áddet") IF (0 (V TV Inf))
(*0 (Pl Nom) BARRIER GRAMCHK-S-BOUNDARY)
(*0 @<OBJ OR @-F<OBJ BARRIER GRAMCHK-S-BOUNDARY)
(*0 Ill + Sem/Hum OR Ill + Sem/Fem OR Ill + Sem/Mal OR Ill + Sem/Sur
OR Ill + Sem/Org OR Ill + Pers);
```

The verb *áddet* 'understand' has a number of typical objects, which are untypical for *addit* 'give', like e.g. members of the language prototype, cf. *suomagiela* 'Finnish (Acc.)' and *kvenagiela* 'Kven (Acc.)' in ex. (13). The real word error rules for a form of *áddet* discards objects of the human, language, text or state prototypes. The rules also search for an illative argument of the human, body, animal prototypes, and destination adverbs/adpositions, which are potential arguments of the verb *addit* 'give'.

(13) . . . ahte áddet jogo **suomagiela** dahje **kvenagiela**
. . . that understand either Finnish.ACC or Kven.ACC
'. . . that you understand either Finnish or Kven'

The rules also discard typical subclause arguments with *ahte* 'that', which appear with *áddet* 'understand', but not with *addit* 'give', and MANNER-arguments like *bures* 'well', *boastut* 'wrongly', *vearrut* 'wrongly', etc. They search for objects of the prototype categories that are concrete, but neither animate nor place, e.g. text, currency, plant (i.e. concrete objects). In addition, the set *ADDIT-OBJ* is specified to identify the objects of idiomatic RECIPIENT-less constructions with *addit* 'give' specific verbs, e.g. *addit ándagassi* 'forgive', *addit ráđi* 'give advice', etc. Verbs with the valencies *<TH-Inf>*, *<Inf>*, and *<TH-Acc-Any><TH-Inf>* can also be governors of the infinitive form *áddet* 'understand' as opposed to *addet* 'give (Prs. 3Pl.)' and are used in the respective rules. In ex. (14), the illative *dutnje* 'you (Ill.)' is an argument of *addá* 'give (Prs. 3Sg.), and not *addet* 'give (Prs. 3Pl.)', which should be *áddet* 'understand'. The real word error detection of *addet* 'give (Prs. 3Pl.)' could be improved by means of governor-argument matching. If the illative is mapped to another verb, the RECIPIENT of *addet* 'understand' does not have a potential form in the sentence and the error can be recognized. However, at the moment dependency annotation is done after local error detection.

(14) *Dasgo Hearrá addá *dutnje* jierpmi **addet** buot.
because Lord gives you.ILL mind.ACC give.PRS.3PL everything
'Because the Lord gives you the mind to understand everything.'

Here, dependency annotation would be useful to establish a governor-argument relation.

Real word error rules for *ádden* ‘understand (Prs. 1Sg.; PrfPrc.)’ use semantic prototype tags. While *addit* ‘give’ typically appears with an argument in accusative case and a human argument in illative case, there are many exceptions. In ex. (15-a) the illative argument is of the body part prototype category and *addit* means ‘hit’ rather than ‘give’. In ex. (15-b) the illative argument is missing, which can be the case in constructions with concrete nouns in accusative case that are not animate, or of the place, currency, or plant prototype like *lieđážiid* ‘flower (Acc. Pl.)’.

(15) a. *...ja de **ádden** luosa *oaiváii* šluppohiin ...
...and then understand.PRS.1SG salmon head.ILL club.COM ...
‘...and then I hit the salmon on the head with the club ...’

b. *Mu **ádden** *lieđážiid* dan dihte go mu
I.ACC understand.PRS.1SG flower.ACC.PL it because because my
mielas son dárbbašii daid
opinion.LOC s/he needed them
‘I gave her/him the little flowers because in my opinion s/he needed them’

Often local contexts are used as clues to identify either member of the confusion pair, i.e. coordination, premodification, subject contexts, object contexts, adverbial contexts, verbal contexts, subclause contexts. Clues can also be more idiosyncratic, as in the sequence *gal varra* ‘definitely blood’ in ex. (16-a), which should be *gal várra* ‘most probably’ and can be used in the real word error detection rule *real-várra*. However, the local context is not always reliable. The form *varra* ‘blood’ typically occurs as the subject of certain verbs (e.g. *golgat* ‘flow’, *boahtit* ‘come’, etc.) in their third person singular form. However, in ex. (16-b), *varra* ‘blood’ is not the subject of *boahtit* ‘come’, but a real word error of *várra*, which is a sentence-initial adverbial. Even though the members of the confusion pair are unrelated and of different parts of speech, they can occupy similar contexts locally. When local contexts of confusion pair members are similar, a global analysis including valencies and semantic prototype categories is necessary.

(16) a. *Nuorta-Finnmárkkus **gal** varra lei veadje[]meahttun ...
East-Finnmark.LOC definitely blood was impossible ...
‘In East-Finnmark, it was most probably impossible ...’

b. *Varra **boahtá** dálá presideanta Egil Olli oidnot dán
blood come.PRS.3SG current president Egil Olli see.PASS.INF this
láhkkái
way
‘The current president Egil Olli will probably appear this way’

When global context is missing or too vague, there can be real ambiguity. In ex. (17-a), *sáhtašeimme* ‘give a ride (Prt. 2Du.)’ can be correct but can also be a real word error for *sáhtášeimme* ‘can (Pot. 2Du.)’. The generic object *dan* ‘it’ cannot be associated with any semantic prototype category and be an object of both verb readings, in the case of *sáhttit*

'can' an elliptical reading. In ex. (17-b), syntactically, *várra* 'possibly' can be correct, but can also be a real word error for *varra* 'blood', even if in the context of a fairy tale (i.e. *Cinderella*), the real word error reading *varra* 'blood' is preferred. The local context can be thought of as a subjectless sentence embedded in an imperative clause in the first case, where *várra* 'possibly' functions as a sentence adverbial. Alternatively, it can be interpreted as a sentence where the subject *varra* 'blood' agrees with the third person singular verb *lea* 'be (Prs. 3Sg.)'.

(17)

a. Na dan gal **sáhtašeimme**.
so s/he.ACC definitely give.a.ride.PRT.2DU
'We definitely gave her/him a ride.'
'This we definitely could (do).'

b. Gea go lea **várra** golleskuova siste
look if is possibly golden.shoe.GEN inside
'Look to see if there is blood in the golden shoe'

The real word error rules for the confusion pair (*sáhtáš*;sáhtaš**) heavily rely on syntax as *sáhttit* 'can' is an auxiliary while *sáhtašit* 'give a ride' is a main verb. However, the verbs *sáhtašit* 'give a ride' and *sáhttit* 'can' can also clearly be distinguished by their valencies. In contrast to *sáhttit* 'can', the verb *sáhtašit* 'give a ride' typically appears with objects of the human prototype category and/or a DESTINATION-argument. These are specified by means of semantic prototype specifications in the real word error rules.

The real word error rules for the confusion pair (*čohkke*;*čohke**) refer to semantic prototype categories and valencies. Rules identifying a form of *čohkat* 'sharpen' that should actually be a form of *čohkket* 'collect' specify negative conditions to the accusative argument of the verb, i.e. they should not be members of the *WOODEN-THINGS*-set or of the tool prototype category. In its infinitive form *čohkket* 'collect' can also be distinguished from the form *čohket* 'sharpen (Prs. 3Pl.)' by means of a potential governor with a *<TH-Inf>*-valency.

The confusion pair (*čohkká*;*čohkka**) includes several verb and noun forms. The rules that identify a real word error for *čohkka* 'mountain top' when it should be *čohkká* 'sit (Prs. 3Sg.)' refer to human and animal subjects and search for a LOCATION-argument in locative case of the following types: furniture, vehicle, building part, building, place, group, organization. In ex. (18) *čohkka* 'mountain top' is correctly used. The clue is a modifier of the place prototype category, i.e. *Stetind* (famous mountain in Northern Norway). The real word error rule for *čohkka* therefore specifies a negative condition for modifiers of the place prototype category.

(18) Stetind nammasaš **čohkka** manná njuolga bajás gitta 1.400 meht[e]ra
Stetind named mountain.top goes straight up until 1,400 meter.GEN

allodahkii.
height
'The mountain top called Stetind goes straight up until an altitude of 1,400 meters.'

Semantic prototype tags and valency tags are used to a different extent to resolve real word errors. There are different strategies depending on the part of speech of the real word error. Verbal real word errors typically test for potential subjects, i.e. nouns in nominative case, belonging to a particular semantic prototype category. The following rule part tests for the human prototype:

```
*0 (Nom) BARRIER GRAMCHK-S-BOUNDARY LINK 0 Sem/Hum OR Sem/Mal OR Sem/Fem OR Sem/Sur
```

Other rules refer to potential objects, i.e. accusative nouns, that are members of a particular semantic prototype category. While *áddet* 'understand' and *sáhtašit* 'give a ride' can have a human object, *ráhkadit* 'make, prepare' can have an object of the food prototype. The following rule part tests for an accusative of the human prototype category:

```
(*0 Acc BARRIER GRAMCHK-S-BOUNDARY OR GRAMCHK-VFIN LINK 0 Sem/Ani OR Sem/Hum OR
Sem/Ani OR Sem/Org OR Indef OR Refl OR Pers - ("dat") LINK NOT 1 Inf)
```

Verb error detection rules test not only the semantic prototype categories of potential subjects or objects, but also adverbials. The verb *čohkkát* 'sit' can have a LOCATION-argument, which is an adverbial of the place prototype category. Noun error detection rules often test for potential governors by referring to their valencies. The form *bidjui* 'den (Ill.)', which can be confused with *biddjui* 'put (Pass. Prt. 3Sg.)', is a potential DESTINATION of a verb with a DESTINATION-argument in its valency. The condition below tests for a context that does not include any of the following valencies: *<TH-Acc-Any><DE-Ill-Any>*, etc.

```
(NEGATE *0 <TH-Acc-Any><DE-Ill-Any> OR <TH-Acc-Any><SO-Loc-Any><DE-Ill-Any>
OR <DE-Ill-Plc> BARRIER GRAMCHK-S-BOUNDARY OR ("de"))
```

Other error detection rules test the semantic prototype category in coordination. The real word error rule *real-várra*, for example, includes a negative condition for coordination with nominative nouns of the semantic categories body, animal product, and substance. Genitive modifiers of the types human, animal and language are also excluded.

5.2.2.2 Local case errors

This section deals with local case errors in adpositional phrases. I distinguish between real word errors, which I consider to be based on idiosyncratic relations between the confused forms, and (morpho-)syntactic errors, which I consider to be based on systematic relations between the confused forms, the former of which were treated in the preceding sections. Both are considered to be real word errors ("duohtasánimeattáhus") by Antonsen (2013, p.11). Case error detection relies on the context needed to resolve them. Whereas global case errors require an analysis of the argument structure of the sentence, local case errors can be resolved locally. Global case errors, which rely on valency structures, will be treated in Section 5.2.3.

Morpho-syntactic errors also involve the confusion of two real word forms. However, because of the systematic relation between the confused forms, general, rather than idiosyncratic, rules that refer to morpho-syntactic characteristics can be used. A typical morpho-syntactic error involving systematically confused forms is a form in nominative case that is confused with a form in genitive/accusative case. Alternating double and single consonants are possible nominative vs. genitive/accusative case distinctions (cf. ex. (19-a)). Local contexts for these forms are adpositional phrases, which involve a pre- or postposition and a dependent genitive case. Genitive case is governed by the pre- or postposition, which is typically adjacent or separated from the genitive noun/pronoun/adjective only by nominal modifiers. Below, I will show how local case errors in adpositional phrases benefit from semantic prototype tags and valency tags, both in disambiguation and error detection.

Case errors can have the same causes as real word errors. They are typically typos resulting in consonant errors such as in ex. (19-a) and in ex. (19-b). In ex. (19-a), the nominative form *dievvá* 'hill' should be a genitive form with only one consonant <v>, i.e. *dievá*. In ex. (19-b), the object of *ráhkadit* 'make' should be an accusative form with a double consonant <tt>, i.e. *goanstta* 'trick (Acc.)', rather than *goansta* 'trick (Nom.)'. Nominative and genitive/accusative forms of nouns with an even number of syllables, aside from contracted stems, only differ in the consonant gradation of the central consonants. For a large number of nouns this means only a quantitative difference, i.e. where the alternation is between single and double consonants. The comitative singular form *áššiin* 'issue (Com.)' in ex. (19-c) is the result of a typo, resulting in two vowels rather than one, i.e. <ii> instead of <i>. Essive and comitative singular forms of nouns ending in *-i* with geminates in the consonant center and consonant gradation between the second and third grade are only distinguished by a double or single <i>. In certain dialects of spoken language, such as Guovdageaidnu, this difference can be difficult to hear.

(19) a. Sii bidje bálgá mielde **dievá/*dievvá** badjel.
they went path along hill.GEN/*hill.NOM over

'They went along the path over the hill.'

b. Jagi 2004:is ráhkadii Gaup ***goansta** mii lea ain dál okta dan
year 2004 made Gaup trick.NOM that is still now one those
goansttain FMX:s.
tricks FMX.LOC
'In the year 2004 Gaup invented a trick that is still one of the tricks in FMX.'

c. ... dát hástalus lea ovddiduvvon ođđa ***áššiin** ...
... this challenge is put.forward.PASS.PRFPRC new issue.COM ...
'... this challenge has been put forward as a new issue ...'

Below, I will focus on case errors in adpositional phrases like the one shown in ex. (19-a). In post- and prepositional phrases, the post- and prepositions require a dependent in genitive case. However, most post- and prepositions are homonymous with adverbs, which do not require a dependent at all. Therefore, case error detection in adpositional phrases depends predominantly on a successful disambiguation of the adposition- and adverb-reading. In regular parsing, the disambiguation grammar *disambiguation.cg3* chooses the adposition- over the adverbial-reading based on the genitive case of the preceding or following noun. Genitive case is disambiguated from accusative case by checking the context for a transitive verb requiring an object in accusative case. However, in grammar checking, the genitive case of the adposition complement cannot be used to disambiguate adpositions and adverbs as the error itself would discard the adpositional reading and thereby eliminate the only hint that could help to detect the case error. Therefore, disambiguation of adpositions and adverbs in *disambiguator.cg3*[12] is based on a set of idiosyncratic rules for each adposition/adverb pair referring to semantic prototype categories and valencies, cf. Table 5.8.

There is a set of 57 disambiguation rules for 23 common adpositions. Altogether there are 305 adpositions in the lexicon, including compound adpositions. There are 1,089 possible analyses of these adpositions, i.e. 3.6 possible analyses per adposition. Further, 61 existing general disambiguation rules from *disambiguation.cg3* are modified to suit the process of error detection. The rules for five of the adpositions include both prototype categories and valencies. The rules for six of the adpositions include only prototype categories, and the rules for three other adpositions include only valencies. Adpositions are often three-way-ambiguous. They have a preposition-, postposition-, and adverb-reading. While disambiguation between adverb- and adpositional readings is systematic, there can also be idiosyncratic homonymies of other parts of speech involved. Nominal or pronominal homonyms of adpositions are usually rare forms or can be disambiguated by valency information. One idiosyncratic homonymy can be found in the form *alddis*, which can be analyzed as the third person possessive locative form of the reflexive pronoun *ieš* 'own' and the third person possessive form of the adverb *alde* 'on'. However, the reflexive

[12] version r116737 (Accessed 2015-06-30)

Postposition /adverb	**Prototype categories**	**Valencies**
Prototype categories and valencies		
ala, nala 'on'	electrical object	<TH-Acc-Any><ala>, <TH-Acc-Clth><ala>
badjel 'over'	measure, money, place	<badjel>
badjelii 'on'	clothes, jewelry	<TH-Acc-Any><badjelii>
mielde 'with'	concrete, human, place, route	<TH-Acc-Any><mielde>, <mielde>
sisa 'inside'	building, container, body of water, substance, cloth object	<TH-Acc-Any><sisa>
Prototype categories		
áigi 'ago'	time	-
alde, nalde 'on'	clothes, jewelry, electrical object	-
vuollel, vuollil 'under'	measure, money	
bokte 'via'	animal	-
maŋŋil 'after'	event, time, organization	-
rastá 'through'	place	-
Valencies		
birra 'about, around'	-	<birra>
bálddas 'next to'	-	<LO-Loc-Plc>
báldii 'next to'	-	<LO-Ill-Plc>

Table 5.8: Semantic prototypes and valencies in disambiguation rules of adpostitions in *disambiguator.cg3* version r116737

pronoun reading occurs in very limited and specific contexts, e.g. as part of a multi-word expression such as *ieš alddis* 'by herself/himself', cf. ex. (20). The disambiguation rule below therefore removes the adverbial reading in the context of *ieš* 'oneself'.

```
REMOVE:GramPo ("alde") IF (0 PxSg3) (-1 ("ieš"));
```

(20) Muhto eai leansmánnit ge nagot **ieš alddis** fuobmát buot
but not sheriff.NOM.PL either manage themselves detect all
monnesuollagiid.
egg.thieve.ACC.PL
'But not even the sheriffs manage to detect all the egg thieves by themselves.'

Other homonymies with adpositions include verb forms like *alde* 'get closer (Prs. 1Du.; Prt. 3Pl.)', *luhtte* 'trust (Prs. 1Du.; Prt. 3Pl.)', and *bokte* 'wake up (Prs. 1Du.; Prt. 3Pl.)', cf. ex. (21).

(21) Háliidan dákko **bokte** giitit buohkaid ...
want this via;wake.up.PRT.3PL thank all.ACC.PL ...
'Hereby, I want to thank everybody ...'

Generally, polysemous adpositions/adverbs require more disambiguation rules as they need to refer to the possible contexts of each sense. Disambiguation rules of the highly polysemous adposition/adverb *badjel* 'over (preposition); more than (+numeral); after (preposition, +temporal expression); over (adverb +temporal expression); away (adverb, as in: the pain went away)', cf. Sammallahti and Nickel (2006, p.37), refer to both semantic prototypes and valencies, cf. Figure 5.4.

```
1  SELECT (Adv) IF (0 ("badjel") LINK 1 Num OR ("čuohti") OR MEASURE OR
2  MEASURE2 OR Sem/Measr OR TIME-QUANT OR ("logenear") OR Sem/Money) ;
3
4  REMOVE (Adv) IF (0 ("badjel") LINK NOT *0 <badjel-V> BARRIER S-BOUNDARY) ;
5
6  REMOVE (Adv) IF (0 ("badjel") LINK -1 Sem/Plc)(NEGATE 1 Num OR N) ;
```

Figure 5.4: Disambiguation rules for *badjel* 'over' in *disambiguator.cg3*

The first rule in Figure 5.4 selects an adverbial reading (i.e. 'more than'), cf. ll.1–2, when *badjel* 'over' is used with a numeral, a measure expression or a noun of the money category, cf. ex. (22-a). Certain idiomatic expressions with verbs that involve *badjel* pick the adverbial reading as well (cf. ex. (22-b)), which is why the tag *<badjel-V>* is used in a negative condition for picking an adverbial reading in the second rule, cf. l.4. The third rule in l.6 removes the adverbial reading of *badjel* in the context of a noun that is a member of the place prototype category expression, as in ex. (22-c).

(22) a. . . . muhtin oassi sáddejuvvon gálvvus lea badjel **vahkku** orron
. . . some part sent goods.LOC has over week.ACC stayed
galbma . . .
frozen . . .
'. . . some part of the sent goods has stayed frozen for over a week . . . '

b. Sus lei visot suhttu mannan **badjel**.
s/he.LOC has all anger gone over
'All her/his anger went away.'

c. *Jiehtanas ii beassan **várri** badjel.
giant not got mountain.NOM over
'The giant did not get over the mountain.'

The following disambiguation rule for *mielde* 'with' refers to both semantic prototypes and valencies. The postpositional reading is chosen in the context of a verb that includes the adverbial reading of *mielde* in its valency (*<TH-Acc-Any><mielde>* and *<mielde>*). In ex. (23), the expression *leat mielde* 'have with' requires an adverbial reading of *mielde*. The postpositional reading is not selected as the verb *leat* has the valency tag *<mielde>*.

```
SELECT Po IF (0 ("mielde") LINK NEGATE -1 <TH-Acc-Any><mielde> OR <mielde>)
(NEGATE -1 Sem/Concrete-NotHuman-NotPlace LINK *-1 (Sem/Hum Loc) OR Prop
BARRIER S-BOUNDARY LINK *0 ("leat" Pl3) OR ("leat" Sg3))
(NEGATE -1 Loc LINK 0 Sem/Org OR Sem/Build OR Sem/Plc);
```

(23) *Áillus lei sihke **loavdda** ja **lávvo-muorat** mielde.
Áilu.LOC was both tarp.NOM and tent-pole.NOM.PL with
'Áilu had both the tarp and tent poles with him.'

Typical modifications to existing disambiguation rules that select alternative (particle- or adverb-) readings to adpositional readings include constraints that prevent the rule from discarding an adpositional reading if present. This is done by conditions of the type *(NEGATE 0 Po LINK -1 Gen)* or the stricter version *(NEGATE 1C Po)*. Since the original rules assume correct input, the modifications prevent correct readings in an erroneous text from being discarded.

After disambiguation, the error detection module *grammarchecker.cg3*[13] deals with case errors in adpositional phrases in the following way: A set of seven dependency rules sets the dependency relation of genitive nouns, pronouns or numerals to an unambiguous postposition following the genitive. Successful disambiguation is clearly necessary for establishing dependency links as only fully disambiguated postpositions are associated with their dependents. The following rule sets the parent of a genitive to a fully disambiguated postposition to the right of it **1C Po* unless it is separated from it by a member of the parameterized set *S-BOUNDARY*, including global conjunctions, subjunctions, relative pronouns, etc.

[13]version r53901 (Accessed 2012-02-10)

```
SETPARENT Gen TO (*1C Po BARRIER S-BOUNDARY);
```

In a final step, an error detection rule maps an error tag *&msyn-gen-before-postp* to the potential dependent of a postposition (noun, pronoun, numeral, adjective) unless it is in genitive case and a dependent of the adposition.

```
ADD:gen-before-postp (&msyn-gen-before-postp)
TARGET NP-HEAD - ABBR IF (NOT 0 Gen)(1C Po)(NEGATE 1 N);
```

The rule that attaches the error tag to a noun, pronoun, numeral or adjective in front of a postposition refers to a clearly disambiguated postposition, i.e. *1C Po*. A number of negative conditions specified after the context operator *NEGATE* make sure that the noun is not a part of another noun phrase, etc.

```
ADD (&msyn-gen-before-postp) TARGET NP-HEAD - ABBR IF (NOT 0 Gen)(1C Po)(NEGATE 1 N)
(NEGATE 1 (&syn-ollis-not-miehtá)) (NEGATE 1 Adv LINK 1 CS)
(NEGATE -1 ("ovdal") OR ("maŋŋil") LINK -1 ("dego") OR ("dugo") OR ("nugo"))
(NEGATE -1 ("ovdal") OR ("maŋŋil") LINK -1 ("go") LINK -1 ("nu"))
(NEGATE 0 (Pron Indef Attr))
(NEGATE 0 ("dat") + (Sg Loc) LINK 1 ("ovdal") LINK -2 Sem/Time LINK -1 Num)
(NEGATE 0 ("ieš") LINK 1C NOT-IESJ-PP)(NEGATE 0 Com + Sem/Animate LINK -1 Gen OR Acc);
```

5.2.2.3 Summary: Local error detection

Local error detection in *grammarchecker.cg3* includes rules for both real word errors and case errors in adpositional phrases. Both semantic prototype categories and valencies are available to local error detection rules and can be used in simple context conditions. Global error detection rules, on the other hand, are preceded by dependency and semantic role analysis and can refer to semantic prototype categories and valencies in specific argument positions. Relations between real word error confusion pair members differ with regard to their similarity in terms of part of speech, syntactic context, valency, and frequency of their correct and erroneous use. Similar confusion pair members typically require more precise real word error rules that also refer to semantic prototype categories and valencies. I chose six frequent real word error confusion pairs that are represented by correct uses of both members in *SIKOR* to illustrate the use of semantic prototype categories and valencies. The examples include members of the same and different parts of speech, i.e. verbs, adverbs, and nouns. Semantic prototype categories are specified for subjects, objects, adverbials, genitive modifiers and coordinated items of the real word error. Valencies are specified where a confusion pair member is the potential argument of a certain governor in the context. Governor-argument dependency relations are not available to local error

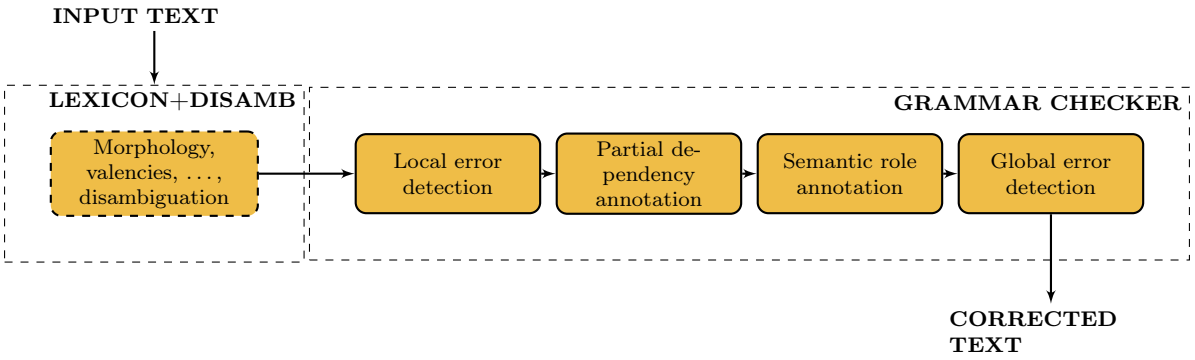

Figure 5.5: The system architecture of global error detection in *GoDivvun*

detection. However, one can specify conditions that refer to a verb with a certain valency in the left or right context of the real word error and specify barriers to limit the range.

While real word error detection rules directly refer to valencies and semantic prototype categories, local case error detection rules for adpositions only refer to morpho-syntactic constraints. They typically include a negative condition for a genitive dependent of the adposition. However, local case error detection rules for adpositions rely on the exact disambiguation of each adposition in question. In order to successfully disambiguate adpositional and adverb readings, both semantic categories and valencies are necessary. Both candidates for an adpositional phrase are checked, i.e. the adposition and a candidate with certain morphological (i.e. part of speech) and semantic characteristics are checked. The alternative adverbial reading is discarded based on typical context conditions for an adverbial reading. Case error detection of the respective noun is only performed if a full disambiguation has taken place.

5.2.3 Global error detection

Global error detection is needed for errors that require a syntactic analysis of the whole sentence. A deep syntactic analysis includes a dependency annotation and a semantic role annotation, which again requires valencies and semantic prototype tags. In this section, I will discuss a certain type of global syntactic errors, i.e. valency errors. The process of valency error detection is complex as most governors have more than one valency. At first, the context of a particular governor is tested to see if any of these valencies are satisfied by correct realizations of the arguments. Additionally, other governors in the sentence are matched with their arguments to discard them as potential valency errors of other governors. While local error detection rules are performed directly after valency annotation and syntactic analysis/disambiguation, cf. Figure 5.3, global error detection takes place after local error detection, governor-argument dependency analysis, and semantic role mapping, cf. Figure 5.5.

Both local and global error detection rules make use of a morphologically and syntactically analyzed and disambiguated input. The lexicon provides semantic prototype tags for potential arguments, i.e. nouns and other parts of speech, and valency tags are added via substitutions in *valency.cg3* to potential governors, cf. Chapter 3. Dependency and semantic role analysis, on the other hand, are directly integrated in the grammar for error detection and correction *grammarchecker.cg3*. Mapping dependencies of governors and their grammatically correct arguments provides an initial identification of the correct forms in a sentence. The rules aspire to test the full valency potential of a governor. Semantic roles are then mapped to the arguments to distinguish between different types of arguments of the governor and mark correct forms of arguments with the equivalent semantic role. These can later be referred to in the error detection rules. Error detection rules then search for a particular form in the sentence that does not correctly satisfy the valency conditions among the forms that have not received a semantic role label. Unless the governor already has a daughter with the required semantic role, the rules add the respective error tag to the unassociated form.

Valency errors are syntactic errors regarding the relation between a governor and its arguments. Valency errors include missing or redundant arguments, morphological case errors, missing or wrong use of subjunctions, erroneous use of non-finite clauses, etc. Valency errors are often standardized to a lesser degree than other morphological or syntactic errors. As mentioned above, the North Sámi norm decided and described in *Riektačállinrávvagat* (2015) does not refer to syntactic errors at all. Its predecessor *Čállinrávagirji* (2003, pp.87–88), on the other hand, specifies a number of both grammatical and ungrammatical valencies.[14] Standard (descriptive) grammars typically provide an incomplete list of grammatically correct valencies, but seldom discuss incorrect valencies. Native speakers, on the other hand, often do have strong intuitions about acceptable and unacceptable valencies. Here, I will therefore follow the norm in the cases where there is one. Otherwise, I will follow the linguistic intuitions of informants *H* and *N*. Possible future syntactic norms will eventually call for enhancements of the grammar checker. The previous section deals with local case errors in adpositional phrases. However, case errors can also be global errors and be related to the valency restrictions of a governor. While the original cause of the error can be the same as for local case errors, e.g. a typo, global error detection requires the grammar checker to perform a global analysis and identify the relation between the verbal governor and the targeted form. The accusative/nominative rule in ll.1–3 adds an error tag, *&syn-acc-not-nom*, to a noun in a nominative singular form. The noun is a member of the set *DOUBLE-CONSONANT-NOUNS*, which includes nouns that alternate between single and double consonants in nominative and accusative case. To the left of the noun there should be a verb with an unsatisfied THEME

[14]"Muhtun sánit gáibidit ahte nubbi eará sátni lea dihto kásushámis, omd. mun beroštan dus (iige *mun beroštan dutnje), mun liikon dutnje (ii ge *mun liikon dus)."

in accusative case in its valency. The essive/comitative rule in ll.5–7 adds an error tag, *&msyn-ess-not-com*, to a noun in comitative case without a semantic role tag. To the left of it there should be a verbal governor with an argument in essive case in its valency, *<TH-Acc-Any><RO-Ess-Any>*, cf. ex. (19-c) in Section 5.2.2.2.

```
1  ADD (&syn-acc-not-nom) TARGET (N Sg Nom) IF
2  (*-1 <TH-Acc-Any> BARRIER NPNHA LINK NONE c §TH)
3  (0 DOUBLE-CONSONANT-NOUNS)(NEGATE 0 §ANYROLE - §AG) ;
4
5  ADD  (&msyn-ess-not-com) TARGET (N Com) IF (NOT 0 Sem/Hum OR Sem/Org OR Sem/Ani
6  LINK *-1 <TH-Acc-Any><RO-Ess-Any> BARRIER NPNH LINK NOT 0 <Com-*Ani>)
7  (NEGATE 0* <Com-*Ani> BARRIER GRAMCHK-S-BOUNDARY)(NEGATE 0 §ANYROLE) ;
```

Case errors can be typos based on phonetic or graphemic similarities of the confused forms, e.g. accusative/nominative and essive/comitative forms. However, there can also be several synonymous valencies in use, of which only one is considered normative. This is the case for the illative rection verb *liikot* 'like', which is also used with THEME-arguments in locative or accusative case in *SIKOR*, cf. also Kittilä and Ylikoski (forthcoming).

The valency rules below focus on errors in passive constructions. The first rule in ll.1–3 adds an error tag, *&syn-illative-agent-with-hallat-passive*, to animate nouns in locative case if the parent is a passive verb that has an AGENT in its valency. The AGENTS in these constructions should be an illative and not a locative form, i.e. *vielljasis* 'brother (Ill. PxSg3.)' instead of *vieljastis* 'brother (Loc. PxSg3.)', in ex. (24-a).

The second rule in ll.5–6 adds an error tag, *&syn-no-agent-with-ot-passive*, to an animate noun in locative case if the parent is an AGENT-less passive verb, i.e. typically *-uvvot* or *-ot*-passive. This is the case for the locative forms *olbmos* 'person (Loc.)', *vieljažiin* 'brother (Loc.)', *oappás* 'sister (Loc.)' and *neabis* 'nephew (Loc.)' in ex. (24-b).

```
1  ADD (&syn-illative-agent-with-hallat-passive) TARGET Loc IF
2  (0 Sem/Hum OR Sem/Ani OR Sem/Org)
3  (NEGATE 0 §SO OR §LO)(p (Der/h Der/alla) OR (Der/halla) OR (Der/adda));
4
5  ADD (&syn-no-agent-with-ot-passive) TARGET Loc IF (0 Sem/Hum OR Sem/Ani OR Sem/Org)
6  (p Der/Pass);
```

(24) a. *Prospero lea herttot Milanos Itálias, muhto rivvehallá
Prospero is duke Milan Italy, but rob.PASS.PRS.3SG
vieljastis válddi.
brother.LOC.PXSG3 power
'Prospero is a duke in Milan in Italy, but is robbed of his power by his brother.'

b. *...eaiggáduvvo viđa **olbmos**, namalassii dan golbma
...own.PASS.PRS.3SG five people.LOC, namely the three
vieljažiin, daid **oappás** ja sin **neabis**.
brother.LOC.PL, their sister.LOC and their nephew.LOC

'. . . it is owned by five people, namely the three brothers, their sister and their nephew.'

Valency error detection relies on a process of government-argument matching. The full valency potential of a governor is tested before searching for an error. While *liikot* 'like' typically has a THEME in illative case, its THEME-argument can also be realized in other ways.

In ex. (25-a), the verb *liikot* 'like' is followed by the possible argument candidates: an adjacent infinitive, *leat* 'be', a noun in locative case, *luonddus* 'nature (Loc.)', and a pronoun in illative case, *sidjiide* 'they (Ill.)'. Only the illative form and the infinitive are correct realizations of a THEME. However, a locative form is a typical valency error. The illative form *sidjiide* 'they (Ill.)' is discarded as an argument of *liikot* 'like', because its actual governor, the verb *addit* 'give', is much closer to it. In this case the adjacent infinitive *leat* 'be' is the argument of *liikot* 'like'. While linear closeness is an important factor in governor-argument matching, the closest candidate is not always the correct candidate. In ex. (25-b), the verb *liikot* 'like' is closer to the accusative pronoun *maid* 'it (Acc.)' than the verb *bargat* 'do, work'. However, *bargat* 'do, work', and not *liikot* 'like', is the governor of *maid*. The following section will therefore deal with possible valency errors and the full valency potential of six verbal governors before discussing the complex process of governor-argument matching and error detection.

(25) a. Oahpahus galgá veahkehit ohppiid liikot *leat* luonddus ja addit
teaching shall help students like be.INF nature.LOC and give
sidjiide vejolašvuođa ovdánahttit fantasiija . . .
they.ILL possibility develop imagination . . .
'The teaching should help the students to like being in nature and give them the chance to develop their imagination'

b. **Maid** don liikot *bargat* friddjabottuin?
what.ACC.PL you like do.INF break.LOC.PL
'What do you like to do on breaks?'

5.2.3.1 Valency errors

The construction of error detection rules requires an overview of the full valency potential in order to associate the governors with their correct arguments before searching for potentially incorrect arguments. This section discusses the valency errors of six North Sámi rection verbs in the context of their full valency potential and their distribution in *SIKOR*.

There are no explicitly stated norms for valencies, and what is considered grammatically correct and what is not is often disputed. Finding a norm for valencies is not the topic of this dissertation. Rather, the purpose is to investigate to what extent it is possible to distinguish between grammatical and ungrammatical sentences by means of rule-based

grammars. Therefore, the distinctions between grammatical and ungrammatical sentences presented in this section will be assumed for the construction of grammar rules. The flexible nature of the grammar checker will allow for possible changes according to future norms. That means that constructions that are labelled as errors here, may be considered correct in the future, and the other way around. In addition, I will show disambiguation errors and real word errors involving instances of the verb. The verbs *liikot* 'like', *luohttit* 'trust', *suhttat* 'get angry', *ballat* 'fear', *beroštit* 'care', and *dolkat* 'get fed up' are listed as rection verbs in a number of Sámi grammars and grammatical descriptions, cf. e.g. Mikalsen (1993, pp.49–74), Pope and Sárá (2004, pp.251–252), and Sammallahti and Nickel (2006). While *liikot* 'like', *luohttit* 'trust', and *suhttat* 'get angry' are described as verbs with an illative-rection ("illatiivarekšuvdna"), *ballat* 'fear', *beroštit* 'care', and *dolkat* 'get fed up' are mentioned as verbs with a locative-rection ("lokatiivarekšuvdna"). Rection verbs are suitable for this study as they prefer a construction with a THEME as opposed to a THEME-less construction. They also preferably appear with their THEME-argument in an adverbial case, i.e. not accusative case. Finding accusative case errors is not primarily related to valency errors as accusative and genitive are homonymous and occupy not only the object position of a verb, but also frequently that of a premodifier. As global errors, in particular valency errors, are the focus of this study, for the most part I will limit my discussion to verbal governors that are frequently involved in valency error constructions.

The selected verbs are represented by an average of 16.3 valency frames in *SIKOR* (excluding different realizations of subjects), of which an average of 8.5 frames are considered grammatical in this work. The frames include arguments of different types, i.e. noun phrases of different morphological case, adpositional phrases, idiomatic constructions with particular adverbs, non-finite constructions of different types and finite subclause arguments.

SIKOR is used both for developing and testing valency error detection rules. Half of the sentences containing instances of the respective verb are used for developing rules; the other half are used for testing. As sentences can include more than one instance of the verb, the corpora for testing and developing differ slightly in size.

5.2.3.1.1 Valencies of *liikot* 'like'

The verb *liikot* 'like' is listed as an illativ verb by Nickel and Sammallahti (2011, pp.233–234), Sammallahti and Nickel (2006, p.436), Mikalsen (1993, p.49), Nielsen (1926-1929, p.525), and *Čállinrávagirji* (2003, p.87). Mikalsen (1993, p.49) and Sammallahti and Nickel (2006, p.436) further name its infinitive valency. Mikalsen (1993, p.49) also mentions subclause arguments introduced by *ahte* 'that' and THEMES realized as noun phrases in accusative case. However, she does not include these in the valency of *liikot* 'like'. Nielsen (1926-1929, p.525) also argues that it can be used in "locative construction[s] as when answering the question 'whither', like being somewhere". This probably refers to

constructions with place nouns where *leat* 'be' is omitted, cf. ex. (26-b). *Čállinrávagirji* (2003, p.87) considers locative constructions ungrammatical, cf. ex. (26-a), without being specific about the type of locative constructions. Both locative and accusative arguments are often considered interference from constructions in the Norwegian and Finnish language and are therefore considered ungrammatical. According to Kittilä and Ylikoski (forthcoming, p.11), "[f]rom the non-prescriptivist point of view", both accusative and locative are grammatical constructions. However, here I will follow the recommendation of *Čállinrávagirji* (2003).

(26) a. *mun liikon **dus**
I like you.LOC
'I like you' (*Čállinrávagirji*, 2003, p.87)

b. ?...liiko nu bures **Sámis** ahte iiba hálitge vuolgit.
...likes so well Lappland.LOC that not want leave
'...s/he likes being in Lappland so much that s/he does not even want to leave.'

In *SIKOR*, there are examples for 14 different valency frames among the 3,801 occurrences of the verb *liikot* 'like', cf. Table 5.9.

Six valencies of the verb *liikot* 'like' are considered grammatical in this work. These make up 93.08%. Grammatical constructions include both THEME-arguments realized as forms in illative case (51.67%), infinitives (30.91%), subordinate clauses (8.26%) and THEME-less constructions (2.24%). In addition to THEMES realized as subclauses with *go* 'that, when' and *ahte* 'that', I also include subclauses with *jus/jos* 'if' in the valency – I consider them arguments – cf. ex. (27), as they frequently appear in constructions where other realizations of a THEME are missing. They can alternatively be thought of as adjunct subclauses. Considering these subclauses THEMES, however, is useful in *grammarchecker.cg3* as they receive a role and are recognized as correct constructions, which again prevents the valency error detection rules from searching for an error.

(27) Kobra lea dego earáge gearbmašat dat ii liiko **jus** dan
cobra is like any.other snake.NOM.PL it not like if it.ACC
fallehit dehe dulbmot.
attack.PRS.3PL or step.PRS.3PL
'The cobra is like other snakes; it does not like to be attacked or stepped on.'

Eight other argument realizations that are considered ungrammatical by informant *H* and/or *N* make up 6.92% of the instances of *liikot* 'like' in *SIKOR*. These include arguments in accusative, locative, nominative or comitative case, and adpositional phrases that are used instead of an argument in illative case. Comitative plural forms are frequently used as locative plural forms in the Eastern dialect, but according to *Čállinrávagirji* (2003, p.83-84) this does not follow the norm. In ex. (28-a) the THEME is realized as a form

Valency	***SIKOR***	**Example**
Grammatical constructions (as defined in this system)		
TH-Ill	1,964	Gonagas gal liikui **dasa**. 'The king liked it.'
TH-Inf	1,175	Son ii liikon **borrat** guoros čoavjái 'S/he did not like to eat on an empty stomach'
TH-go	219	Mii borgalottit liikot oba bures **go** albma ládje borgá 'We blizzard birds like it very much when there is a real snowstorm'
TH-jus/jos	29	ii liiko **jus** dan fallehit dehe dulbmot. 's/he does not like when s/he is attacked or stepped on.'
TH-ahte	66	In liiko **ahte** buot sámegiel báikenamat eai leat fárus. 'I do not like that they have not included all Sámi place names.'
TH-0	85	Mun liikon buoremusat. 'I like (it) best.'
Ungrammatical constructions and corrections (as defined in this system)		
TH-Acc	189	*Son liikui erenoamáš bures **sistesihkkelastima** 'S/he liked indoor biking quite a lot.'
(Ill.)		Son liikui erenoamáš bures **sistesihkkelastimii**
TH-Loc	45	* inge liiko **dilis** mas mii leat. 'I don't like the situation we are in.'
(Ill.)		inge liiko **dillái** mas mii leat.
TH-FS	14	*olbmot liikojit **maid son lea designen**. 'people like what s/he has designed'
(+ *dasa*)		olbmot liikojit **dasa** maid son lea designen.
TH-Adv-loc	9	?Mun liikon **dáppe** badjin 'I like (it) up here'
(+ *leat* 'be')		Mun liikon **leat** dáppe badjin 'I like to be up here'
TH-Nom	3	*Liikon **dat** mii lea simpal ja vulgeara 'I like things that are simple and unrefined'
(Ill.)		Liikon **dasa** mii lea simpal ja vulgára
TH-Com	1	*son lea álo liikon **ivnniguin** ja daid son lea málen dávváliin 's/he has always liked colors and s/he has painted them on the blackboard'
(Ill.)		son lea álo liikon **ivnniide** 's/he has always liked color'
TH-ovddas	1	*Boares áhkku nu liikui veahki **ovddas** 'The old woman liked the help so much'
(Ill.)		Boares áhkku nu liikui **veahkkái**
TH-AktioEss	1	ii liiko **leamen** guovddážis. 's/he does not like to be in the center'
(Inf.)		ii liiko **leat** guovddážis. 's/he does not like to be in the center'
Total	3,801	

Table 5.9: The valency distribution of *liikot* 'like' in *SIKOR*

in accusative case, *dili* 'situation (Acc.)'. In ex. (28-b), there is a locative THEME, *dilis* 'situation (Loc.)'. Relative clauses are typical contexts for valency errors. The cases of the referents of the relative clause and the matrix clause are easily confused. In ex. (28-c), the THEME of *liikot* 'like' is realized by a form in nominative case matching the case of the subsequent subject, *mii* 'that' of the relative clause. However, it should be realized by a form in illative case as it is an argument of *liikot* 'like'.

Other ungrammatical constructions are cases of ellipses, where a full version is preferred in written text. In ex. (28-d), on the other hand, the infinitive *leat* 'be' is omitted in a construction with a location adverb, e.g. *dáppe* 'here'. Informant *H* suggests adding the infinitive *leat* 'be' to correct the sentence. In ex. (28-e), the referent of the matrix verb *liikot* 'like' is missing, leaving a (finite) relative clause without an explicit referent, unlike constructions like ex. (28-f), where the referent (*dasa* 'it (Ill.)') of the relative pronoun is expressed. Informant *H* suggests adding the illative referent of the relative clause to correct the sentence.

(28) a. *. . . gii ii loga liikot **dili** nu go dál lea.
. . . who not say like.INF situation.ACC such as now is
'. . . who says s/he doesn't like the situation as it is now.'

b. *. . . inge liiko **dilis** mas mii leat.
. . . not like situation.LOC which.LOC we are
'. . . I don't like the situation we are in.'

c. *Liikon **dat** mii lea simpal ja vulg[á]ra
like.PRS.1SG it.NOM that.NOM is simple and unrefined
'I like simple and vulgar things unrefined'

d. ?Mun liikon **dáppe** badjin
I like here up
'I like (it) up here'

e. *. . . olbmot liikojit **maid** son lea designen.
. . . people.NOM.PL like what.ACC s/he has designed
'. . . people like what s/he has designed.'

f. sii liikojit **dasa** maid besset vásihit
they like.PRS.3PL it.ILL that.ACC got experience.INF
'they liked what they got to experience'

5.2.3.1.2 Valencies of *luohttit* 'trust'

The verb *luohttit* 'trust' is listed as an illative rection verb by Sammallahti and Nickel (2006, p.450), Mikalsen (1993, p.50), Nielsen (1932-1960*b*, p.586), and *Čállinrávagirji* (2003, p.87). Sammallahti and Nickel (2006, p.450) further mention its use with adpositional phrases with *ala* 'at', cf. ex. (29-a). Mikalsen (1993, p.50) also mentions subclauses with *ahte* 'that' and accusative + infinitive constructions, cf. ex. (29-b).

(29) a. luohttit soapmása **ala**
trust someone.GEN on

'to trust someone' (Sammallahti and Nickel, 2006, p.450)

b. Mii luohttit **du** *nagodit* visot akto bargat.
we trust you.ACC manage.INF everything alone do.INF
'We trust you to manage to do everything by yourself.' (Mikalsen, 1993, p.50)

Table 5.10 shows the representation of the valencies of *luohttit* 'trust' in *SIKOR*.

Of the 1,163 occurrences of *luohttit* 'trust' in *SIKOR*, 830 (71.36%) are considered grammatical in this study. I exclude 175 occurrences of *luohttit* 'trust' from the valency analysis, i.e. 14.96%. These include certain derivations with other valencies (2) like the deverbal noun *luohttima* 'trust (Acc.)' in ex. (30-a) and disambiguation errors (154) of forms like **luhtte** in ex. (30-b). I also exclude real word errors (18) like *luohte* 'trust' (Imp. 2Sg.) in ex. (30-c), which should be a one-word compound with the adjacent word, i.e. *luohteárbevierru* 'tradition of joik'.

(30) a. dakko bokte f[a]rggabut vuitet mánáid **luohttima**
that through sooner win.PRS.3PL children.GEN.PL trust.ACC
'through that they gain the children's trust sooner'

b. Fitnat dáidd[á]ra **luhtte**, teáhteris dahje čájáhusas
visit artist.GEN trust.PRS.2DU;at, theater.LOC or exhibition.LOC
'Visit the artist, the theater or the exhibition'

c. ***Luohte** árvevierru lea su geasuhan
trust.IMP.2SG tradition has s/he.ACC attract.PRFPRC
'The joiking tradition has attracted him/her'

The verb *luohttit* 'trust' is typically used with a THEME-argument; only 1.55% of the cases are used without a THEME. Without disambiguation and real word errors, which make up a significant 15.04%, there are 998 cases of *luohttit* 'trust' in *SIKOR*. The most frequent valency includes a THEME in illative case (70.25%), such as *guhtet guimmiidasamet* 'each other (Ill.)' in ex. (31). The THEME can also be realized as a postpositional phrase with *ala/nala* 'on' (1.03%).

(31) Jos duostat luohttit **guhtet guimmiidasamet**, de mii ollet
if dare.PRS.1PL trust.INF each other.ILL, then we reach
guhkkelii.
further
'If we dare to trust each other, then we get further.'

Ungrammatical constructions (14.02%) include predominantly subordinate clauses with *ahte* 'that' (9.37%), cf. ex. (32-a), which according to *N* should be preceded by a pronoun in illative case, i.e. *dasa* 'it (Ill.)'. The second largest group of ungrammatical constructions are THEME-less constructions that should have a THEME in illative case. Infinitival constructions include accusative + infinitive constructions, cf. ex. (32-b), simple infinitival constructions and non-finite actio locative – or according to Ylikoski

Valency	***SIKOR***	**Example**
Grammatical constructions (as defined in this system)		
TH-Ill	818	luohttit **guhtet guimmiidasamet** ‘trust **each other**’
TH-(n)ala	12	luohttit guhtet guimmiidasaset **nala** ‘trust **on** each other’
Ungrammatical constructions (as defined in this system) and corrections		
TH-ahte	109	*luohttit **ahte** sin ságat dollet deaivása. ‘trust **that** their stories add up.’
(+ *dasa*)		luohttit **dasa** ahte sin ságat dollet deaivása.
0	18	*Diggi ii luohttán. ‘The court does not trust’
(+ *dasa*)		Diggi ii luohttán **dasa**.
Acc + Inf	9	*ferte luohttit **gánddaid nagodit** rahčat. ‘needs to trust that **the boys will manage** to make an effort’
(*dasa ahte* + Nom + VFIN)		ferte nagodit luohttit **dasa ahte gánddat rahčet**.
TH-Inf	4	álbmot mii luohttá **gábidit** iežas vuoigatvuođaid ‘*a nation that trusts **to ask** for its rights’
(lexical error)		álbmot mii *duostá* gábidit iežas vuoigatvuođaid
TH-Aktioloc	2	In sáhte luohttit **addimis** hálldašeami dakkar searvái ‘I cannot trust such an association with our administrative work’
(Inf.)		In sáhte luohttit, **inge addit** hálldašeami dakkar searvái
TH-FS-Qpron	3	*iige dušše luohttit **mii doppe lei daddjon** ‘trust **what has been said there**’
(+ *dasa*)		iige dušše luohttit **dasa** mii doppe lei daddjon
TH-FS	2	Mun luohtán **Kárášjoga Sámiid Searvi gávdná buori kandidahtá** ‘I trust **Kárášjohka’s Sámi Community will find a good candidate**’
(+ *dasa ahte*)		Mun luohtán **dasa ahte** Kárášjoga Sámiid Searvi gávdná buori kandidahtá
Ill + *ahte*	3	*mun sáhtán sidjiide luohttit **ahte doibmet 100 proseantta** ‘*I can trust them **to function 100 percent**’
(+ *dan ektui*)		mun sáhtán sidjiide luohttit **dan ektui ahte doibmet 100 proseantta**
TH-Acc	7	*Dál orut beare haga luohttime **dovdduidat** ‘Now you seem too to be trusting **your feelings** too much’
(Ill.)		Dál orut beare haga luohttime **dovdduidasat**
TH-Com	1	*Ii oro gal hotealla **dáinna báhpiriin** stuorrát luohttimin ‘The hotel does not seem to trust **these papers** very much’
(Ill.)		Ii oro gal hotealla **dáidda báhpiriidda** stuorrát luohttimin
RS-Ess	4	?Ii áhkku luohte aitto **danin**. ‘Grandmother does not seem to trust (i.e. have faith) just **because of that**’
		Ii áhkku luohte aitto **dasa**.
Disambiguation errors and real word errors		
derivations	2	go sii dakko bokte f[a]rggabut vuitet mánáid **luohttima** ‘because they gain the children’s **trust** sooner’
disambiguation error	154	Fitnat dáidd[á]ra **luhtte**, teáhteris dahje čájáhusas ‘Visit the artist at the theater or exhibition’
real word error	18	***Luohte** árvevierru lea su geasuhan ‘**trust** tradition has attracted him/her’
		Luohteárbevierru lea su geasuhan
Total	1,163	

Table 5.10: The valency distribution of *luohttit* ‘trust’ in *SIKOR*

(2009, p.36) second infinitive – constructions. Accusative + infinitive constructions are corrected to finite subclauses introduced by *ahte* 'that' and preceded by *dasa* 'it (Ill.)' by *H* and *N*. (Magga, 1986, p.169), on the other hand, does not mark accusative + infinitive constructions such as the ones in ex. (32-c) as ungrammatical. In addition, there are finite subclause constructions that are introduced by question pronouns, cf. *mii* 'what' in ex. (32-d), and those that are introduced by neither a question pronoun nor a subordinating conjunction. These should be preceded by a pronoun in illative case, i.e. *dasa* 'it (Ill.)', according to *N*. Ungrammatical constructions also include simple noun phrases in accusative and comitative case, which should be realized by illative case. Altogether, the errors make up 13.72%.

(32) a. *guldaleaddjit galget ain boahtteáiggis luohttit **ahte** sin ságat dollet
listeners should still future.LOC trust that their stories are
deaivása.
true
'the listeners should still trust in the future that their stories are true.'

b. *Šaddá váttis čiekčan, ferte luohttit **gánddaid nagodit**
becomes difficult football, needs trust.INF boy.ACC.PL manage.INF
rahčat.
make.an.effort.INF
'It is a difficult football match, one needs to trust that that the boys will manage to make an effort.'

c. Mii luohttit du nagodit visot akto bargat
we trust you.ACC manage everything alone do.INF
'We trust that you manage to do everything by yourself' (Magga, 1986, p.169)

d. *iige dušše luohttit **mii doppe lei daddjon**
not.either only trust what.NOM there has say.PRFPRC
's/he does not just trust what has been said there either'

5.2.3.1.3 Valencies of *suhttat* 'get angry'

The verb *suhttat* 'get angry' is listed as an illative rection verb by Sammallahti and Nickel (2006, p.676), Mikalsen (1993, p.54), Nielsen (1932-1960*c*, p.609), and *Čállinrávagirji* (2003, p.87). Sammallahti and Nickel (2006, p.676) and Nielsen (1932-1960*c*, p.609) also mention THEMES realized as adpositional phrases with *ala* 'at' and THEME-less constructions with past participle forms of *suhttat* 'fear'.

Table 5.11 shows the valency distribution for *suhttat* 'get angry' in *SIKOR*. It is less frequent than *liikot* 'like' and *luohttit* 'trust'. A number of forms and uses are excluded from the analysis (6.92%). These are, for example, attributive uses of the perfect participle form *suhttan* 'angry', as in ex. (33-a). There are only two instances of real word errors, concerning the form *suhtastallama* 'getting angry (Gen.)', shown in ex. (33-b), which is confused with *suohtastallama* 'having fun (Gen.)'.

Valency	***SIKOR***	**Example**
Grammatical constructions (as defined in this system)		
TH-0	431	Naba go beana lea suhttan? 'And when the dog is angry?'
TH-Ill	152	olmmoš suhtai nuppi **olbmui** 'a person got angry **at another person**'
RS-Loc	8	Máhtte suhtai **das**. 'Máhtte got angry **because of it**.'
RS-Com	7	mii leat máŋgasat geat leat suhttan dáinna **barg[o]vugiin** 'many of us have gotten angry because of this **working method**'
TH-ala	13	ledjen suhttan iežan ja daid earáid **ala**. 'I was angry **at** myself and the others.'
RS-dihte	18	suhttan heajos čázi **dihte** 'angry **because of** the water'
RS-geažil	5	eai oro suhttame nu olu láigolihtu **geažil** 'they do not seem to get angry so much **because of** the rent agreement'
TH-go	129	ja dál leat olbmot suhttan **go** Radio Golli ii leat dan dieđihan ovdal. 'and now people have gotten angry **that** Radio Golli did not report it earlier.'
Ungrammatical constructions (as defined in this system) and corrections		
TH-Inf	1	*ja suhttan **gullat** olbmuid šláddariid sudno birra 'and gotten angry **hearing** people's gossip about them' ja suhttan **go gulan** olbmuid šláddariid sudno birra
Disambiguation, syntactic uses and real word errors		
syntax	56	Movt gulahallat **suhttan** mánáin? 'How to deal with an **angry** child?'
real word	2	***Suhtastallama** duođalaš beali birra 'About the serious side **of getting angry**'
error		**Suohtastallama** duođalaš beali birra 'About the serious side **of having fun**'
Total	838	

Table 5.11: The valency distribution of *suhttat* 'get angry' in *SIKOR*

(33) a. Movt gulahallat **suhttan** mánáin?
how communicate.INF angry child.COM
'How to communicate with an angry child?'

b. ***Suhtastallama** duođalaš beali birra logaldallá NTNU professor
getting.angry.GEN serious side about lectures NTNU professor
'The NTNU professor lectures about the serious side of having fun'

91.05% of all valencies of *suhttat* 'get angry' in *SIKOR* are considered grammatical constructions by *N* and in this work. 51.43% appear without a THEME. 18.14% appear with a THEME in illative case, cf. ex. (34-a), and only 1.55% with a THEME realized as an adpositional phrase with *ala* 'at', cf. ex. (34-b). The material also includes postpositional phrases with *dihte* 'because of', and *geažil* 'because of'. 0.95% have a REASON-argument in locative or comitative case, cf. ex. (34-c).

(34) a. Dolin jus olmmoš suhtai nuppi **olbmui**, de ...
in.old.days if person get.angry.PRT.3SG other person.ILL, then ...
'In the old days if a person got angry at another person, then ...'

b. ledjen suhttan iežan ja daid earáid **ala**.
have get.angry.PRFPRC myself and the other.GEN.PL on
'I have gotten angry at myself and the others.'

c. mii leat máŋgasat geat leat suhttan dáinna
we are many that have get.angry.PRFPRC this.COM.PL
barg[o]vugiin
working.method.COM
'many of us have gotten angry because of this work routine'

15.39% appear with a THEME realized as a *go*-subclause, cf. ex. (34-b). There is only one instance of an infinitive argument (0.12%), which, according to *N*, should be a *go*-subclause, cf. ex. (35-b).

(35) a. ja dál leat olbmot suhttan **go** Radio Golli ii leat dan
and now have people get.angry.PRFPRC that Radio Golli not have it
dieđihan ovdal.
told earlier
'and now people have gotten angry because Radio Golli did not report it earlier.'

b. *Soai leaigga sihke dolkan ja suhttan **gullat**
they had both get.fed.up.PRFPRC and get.angry.PRFPRC listen.INF
olbmuid šláddariid sudno birra
people.GEN.PL gossip.ACC.PL they.GEN about
'They had gotten fed up and angry listening to people's gossip about them'

5.2.3.1.4 Valencies of *beroštit* 'care'

The verb *beroštit* 'care' is listed as a locative verb by Sammallahti and Nickel (2006, p.73), Mikalsen (1993, p.71), Nielsen (1932-1960*a*, p.154), and *Čállinrávagirji* (2003, p.87). Nielsen (1932-1960*a*, p.154) and Mikalsen (1993, p.71) also mention *beroštit* 'care' with an infinitive THEME, cf. ex. (36-a). Sammallahti and Nickel (2006, p.73) also mention a THEME-less construction, cf. ex. (36-b). *Čállinrávagirji* (2003, p.87) considers an illative THEME ungrammatical, cf. ex. (36-c). Kittilä and Ylikoski (forthcoming, p.16) also describe realizations as adpositional phrases with *birra* 'about' and accusative case. However, they explicitly apply an non-prescriptivist point of view.

(36) a. don gal gusto beroštat **čuovvut**
you certainly obviously care follow.INF
'you obviously care to follow' (Nielsen, 1932-1960*a*, p.154)

b. ale beroš!
don't bother
don't bother! (Sammallahti and Nickel, 2006, p.73)

c. *mun beroštan **dutnje**
I care.about you.ILL
'I care about you' (*Čállinrávagirji*, 2003, p.87)

In *SIKOR*, there are examples for 15 different valency frames among the 3,076 occurrences of the verb *beroštit* 'care', cf. Table 5.12. I excluded instances of the non-finite (abessive) form *beroškeahttá* 'careless' from *SIKOR* because it differs in its valency distribution from other forms of *beroštit* 'care'. 1.72% make up other derivations (e.g. causative derivations and nominalizations) that I have not included in the evaluation as they significantly changed the valency preferences. In the relative construction in ex. (37), the passive is used with both an EXPERIENCER in nominative case, *politiijat* 'police (Nom. Pl.)', and a THEME in locative case, *geain* 'who (Loc. Pl.)', and the sentence does not make sense. Instead, an active form of the verb should be used.

(37) *sáhttá identifiseret olbmuid **geain** politiijat leat
can identify people.ACC.PL who.LOC.PL police have
beroštuvvon
care.PASS.PRFPRC
'*can identify people who the police have been cared about'

Of 3,801 occurrences altogether, eight valency frames (92.71%) are considered grammatical and 5.6% ungrammatical by *H* and also in this work. In the case of *beroštit* 'care', grammatical valency frames include locative arguments, infinitives or subordinate clauses with *ahte* 'that' or *go* 'that, when'. Seven valency frames are considered ungrammatical. These include accusative, illative, comitative, and nominative THEMES that should be in locative case, cf. ex. (38-a). There are also instances of adpositional phrases with *birra* 'about' that should be in locative case, cf. as *man birra* 'about which' in ex. (38-b). In addition, there are finite subclauses that should be preceded by a referent in locative clause, cf. ex. (38-c). Compared to the verb *liikot* 'like', there are less possible governed cases of the THEME, resulting in more varied case errors, i.e. accusative, illative and comitative THEMES, which should realized by means of locative case, cf. ex. (38-a).

(38) a. *suohkan berošta iežas **nuoraiguin** ja háli[i]da sin ruoktot
municipality cares their youth.COM.PL and want them home
boahtit bargat
come.INF work.INF
'the municipality cares about its young people and wants them to come home to work'

b. *Árbeášši ii leat gal dat **man** **birra** orru Heaika eanaš
inheritance not is certainly that which.GEN about seems Heaika mostly
berošteamen ...
caring ...
'Inheritance is certainly not the thing that Heaika cares most about ...'

c. *Lean eambbo beroštan **mii muitaluvvo** **luođis** ...
have more care.PRFPRC that tell.PASS.PRS.3SG joik.LOC ...
'I have cared more about what is told in the joik ...'

Valency	***SIKOR***	**Example**
Grammatical constructions (as defined in this system)		
TH-Loc	2,113	Dásge lea sáhkan beroštit **nuppis**. 'care about **the other one**'
TH-Inf	399	geat beroštedje **oahppat** 'who cared **to learn**'
TH-Aktioloc	42	Biret Ánná beroštišgođii **luonddudálkkodeamis** 'Biret Ánná started to care about/got interested in **natural medicine**'
TH-go	27	ii olus beroštan **go** bealli manai gildii dienasin 's/he didn't care at all **that** one half went to the municipality as profit'
TH-ahte	24	eai oro berošteamen stuorrát **ahte** skuter bisánii 'they do not seem to care a lot **that** the scooter stopped'
TH-jus	4	duostat beroštit **jus** oidnet soapmasa gii dárbbaša veahki 'dare to care **if** they see someone who needs help'
TH-0	243	Berrebehtet beroštit. 'You should care.'
Ungrammatical constructions (as defined in this system) and corrections		
TH-FS	56	*Lean eambbo beroštan **mii muitaluvvo luođis** 'I have cared more about **what is told in the joik**'
(+ *das* 'it (Loc.)')		Lean eambbo beroštan **das** mii muitaluvvo luođis
TH-Acc	50	*iige beroštan **olbmuid** iige **biillaid** 's/he did not care about neither **people** or **cars**'
(Loc.)		iige beroštan **olbmuin** iige **biillain**
TH-birra	25	*Suoma bealde fas beroštit **fitnodatoagguma birra** 'In Finland, they care **about commercial fishing** again'
(Loc.)		Suoma bealde fas beroštit **fitnodatoaggumis**
TH-Ill	20	*lohká dehálažžan ealgabivdit beroštišgohtet **bivdogeahččalemiide** 'elk hunters start to care **about hunting exams**'
(Loc.)		lohká dehálažžan ealgabivdit beroštišgohtet **bivdogeahččalemiin**
TH-Com	18	*suohkan berošta iežas **nuoraiguin** 'the municipality cares **about** its **young people**'
(Loc.)		suohkan berošta iežas **nuorain**
TH-Nom	1	ii berre beroštit **dat** geat eai geardda gullat sámegiela 'one should not be bothered **by those** that do not tolerate to hear Sámi'
(Loc.)		ii berre beroštit **dain** geat eai geardda gullat sámegiela
TH-PrfPrc	1	*gielda ii leat beroštange **fitnan** guorahallamin su dárbbuge. 'the municipality hasn't even bothered to **pay** him **a visit** in order to look into his needs'
(Inf.)		gielda ii leat beroštange **fitnat** guorahallamin su dárbbuge.
Derivations with other valencies		
derivations	50	mii eanemusat beroštahtii **dološáigásaš geađgedáiddáriid**. 'that first got **ancient stone artists** interested'
nominalization	3	jus dát hehtte su **beroštemiid**. 'if this goes against his/her **interests**.'
Total	3,076	

Table 5.12: The valency distribution of *beroštit* 'care' in *SIKOR*

5.2.3.1.5 Valencies of *ballat* 'fear'

The verb *ballat* 'fear' is listed as a locative verb by Sammallahti and Nickel (2006, p.44), Mikalsen (1993, p.69), Nielsen (1932-1960*a*, p.125), and *Čállinrávagirji* (2003, p.87). Sammallahti and Nickel (2006, p.44) also mention THEME-less constructions with MANNER-adverbials like *issorasat* 'awfully', cf. ex. (39-a), subclause constructions with *ahte* 'that' and non-finite (actio essive, i.e. progressive) THEMES. Nielsen (1932-1960*a*, p.125), on the other hand, mentions generic accusatives, cf. ex. (39-b), and subclause constructions with *ahte* 'that'. Kittilä and Ylikoski (forthcoming, p.20) also describe uses of accusative THEMES, but again from a descriptive point of view.

(39) a. ballat **issorasat**
fear awfully
'be awfully afraid' (Sammallahti and Nickel, 2006, p.44)

b. **dan** in bala
it.ACC not fear
'I do not fear it (happening)' (Nielsen, 1932-1960*a*, p.125)

In *SIKOR*, there are examples for 37 different valency frames among the 5,695 occurrences of the verb *ballat* 'fear', cf. Table 5.13 and Table 5.14. This makes it the most versatile verb of this analysis. 0.93% are certain derivations with other valencies, syntactic uses, disambiguation errors, and real word errors of the verb *ballat* 'fear', which I excluded from the valency analysis. Derivations like the non-finite (actio locative) *ballamis* 'there is the fear' are often used adverbially and have other valency preferences than the regular use of the verb, cf. ex. (40-a). Also certain syntactic functions, such as attributive uses of the perfect participle form *ballan* 'afraid' like the one in ex. (40-b), are excluded. Disambiguation errors include homonymous forms of, for example, *ballu* 'fear', the surname *Ballo*, *balláde* 'ballade', and *ballát* 'get scared'. Real word errors include confusions with *ballát* 'get scared' and *bállet* 'be in peace'. In ex. (40-c), the form *ballat* is a real word error and should be *ballát* 'get scared'. The verb *ballát* 'get scared' can appear with a DESTINATION-argument in illative case or with the adverb *eret* 'away'.

(40) a. **ballamis** lea ahte dárogiella sáhttá vuoitit
fear.ACTIO.LOC is that Norwegian can win.INF
'there is the fear that Norwegian can dominate'

b. . . . lei nieida suorganan dego **ballan** njoammil.
. . . has girl get.scared like fear.PRFPRC rabbit
'. . . the girl got scared like a frightened rabbit.'

c. *. . . dat mat ovddemus bohte guvlui eai dárbbaš **ballat** eret . . .
. . . that which first arrive area.ILL not need get.scared away . . .
'. . . the first ones that arrived to the area do not need to get scared away . . . '

Of a total of 5,656 instances of the verb, 23 different valency frames are categorized as valid valencies (85.65%), cf. Table 5.13. The verb is typically used with a THEME,

Valency	*SIKOR*	Example
Grammatical constructions (as defined in this system)		
TH-Loc	1,256	iige bala **barggus** 's/he is not afraid of work either'
TH-Acc-ok	35	- **Dan** in bala. 'I am not afraid of it (happening)'
TH-Com-ok	17	muhto ii **dainna** dárbbat ballat. 'but one does not have to worry about that.'
TH-ahte	1,495	Balan **ahte** lea juo menddo maŋŋit. 'I am afraid that it is already too late'
TH-go	94	ja balan **go** mánáide biddjojuvvo ovddasvástádus 'and then I am afraid that the responsibility is given to the children'
TH-jus	35	Ii ábut ballat **jus** álggos dadjá sániid áibbas endorii. 'It is not worth it to be afraid if in the beginning one says the words utterly wrong'
TH-FS-comma	83	**Muđui sáhttet sámit báhcit dušše historjan**, ballá son. 'Otherwise, the Sámi may soon be a thing of the past, s/he fears.'
TH-FS-Qpron	74	ballat **goas dal de heaittihit fálaldaga**. 'be afraid of what will happen when they finish the offer'
TH-FS-Qst	3	ballat **lea go áhku árbi duššan agibeaivái**. 'they fear that their grandmother's heritage has been wasted forever'
TH-Aktioloc	377	ballat **buohccámis** 'be afraid **of getting sick**'
TH-Inf-ok	1	Son ballá boastut **sáhttit** mannat 'S/he is afraid that it could go wrong'
Acc + Inf	672	balai **daid šaddat** heajos ovdamearkan ássiide 's/he was afraid of them becoming a bad example for the inhabitants'
Acc + PrfPrc	46	**maid** son ballá **suoláduvvon** 'which s/he is afraid has been stolen'
Acc + AktioEss	14	**Albmá** balle **gárremi[i]n** 'they were afraid that the man was acting drunk'
Acc + Ess	13	**maid** mii ballat dábáleamos **ággan** 'which we fear (is) the most common excuse'
Acc + Loc	10	Sii balle eambbo **miinnaid šiljus**. 'They were afraid there (were) more mines in the yard'
Acc + Hab	3	muhto ballá **direktoráhtas eará plánaid**. 'but is afraid of the director (having) other plans'
Acc + VGen	1	ballat garra guohtunnákkuid **boađi boađi**. 'fear that hard grazing disputes are just around the corner'
TH-dihte	23	balai heakkas **dihte** 's/he was worried **about** his/her life'
TH-geažil	4	ballá erenoamážit daid doalloheaittihemiid **geažil** 'fears especially **because of** the shutting down of industries'
TH-beales	3	Son balai ođđa fitnodaga boahtteáiggi **beales**. 'S/he was worried **about** the new company's future'
TH-danne	2	**danne** eai dárbbat ballat '**because of this** they do not need to worry'
TH-0	617	Ale bala! 'Don't be scared!'

Table 5.13: The valency distribution of *ballat* 'fear' in *SIKOR* (part 1)

which is realized as a nominal argument, a non-finite argument or a finite subclause. The nominal and non-finite arguments are typically in locative case (33.47%). In ex. (41-a), the THEME of *ballat* 'fear' is realized as the non-finite actio locative form *buohccámis* 'get sick (Actio. Loc.)'. Accusative case (0.72%) is restricted to very general expressions, cf. ex. (39-b). Comitative arguments denoting the indirect cause of actions are mentioned by Nielsen (1926-1929, p.348), cf. ex. (41-b), where *ballat* can be translated as 'be worried about' (0.12%). The verb *ballat* 'fear' also appears with THEMES realized as adpositional constructions with *geažil* 'because of', *beales* 'on someone's account' or *dihtii/dihte* 'because of'. Of the 4,879 correct instances of *ballat* 'fear', 12.63% are used without a THEME- or REASON-arguments; some of them express the extent of fear, cf. ex. (41-c).

(41) a. Ii oktage bargi galgga ballat **buohccámis**
not any worker should fear get.sick.ACTIO.LOC
'No worker should be afraid of getting sick'

b. Muhtin váhnemat dán nástegovas ballet maid hirbmadit **dainna**
some parents this zodiac.sign.LOC fear also extremely it.COM
maid sin mánát fuobm[á]jit bargat.
what their children come.up do.INF
'Some parents with this zodiac sign are extremely worried about the mischief their children could think up.'

c. ... go hupmen journalisttain ballen **veahá**.
... when talk.PRT.1SG journalist.COM.PL get.scared a.little
'... when I talked to the journalists I got a little scared.'

Other very frequent realizations of the THEME-argument are subclauses with *ahte* 'that' (26.26%), *jus/jos* 'if' and *go* 'that'. Subclauses can also be introduced by a relative pronoun or a question adverbial, cf. ex. (42-a), include a finite verb with a question particle, cf. ex. (42-b), or be followed by a comma-separated finite verb and optional subject, similar to a direct speech construction, cf. ex. (42-c).

(42) a. ballat **goas** dal de heaittihit fálaldaga.
fear when now then finish.PRS.3PL offer.ACC
'be afraid of what will happen when they finish the offer.'

b. ... ja ballat **lea go áhku árbi duššan**
... and fear is Q grandmother.GEN heritage waste.PRFPRC
agibeaiváí.
forever.ILL
'... they fear that their grandmother's heritage has been wasted forever.'

c. **Ja dat gal veadjá boahtit**, balan mun.
and this certainly can come, fear I
'And this certainly may come, I fear.'

THEMES realized as non-finite constructions include simple actio locatives, or infinitives

in impersonal constructions, cf. ex. (43-a). Additionally, there are non-finite clauses with accusative subjects, where the non-finite verb form can be an infinitive, a participle perfect or an actio essive, i.e. progressive, form (cf. Nickel and Sammallahti (2011, pp.262–265)), as in ex. (43-b). Non-finite forms of *leat* 'be' can also be omitted in elliptical accusative + essive, and certain accusative + locative constructions. In ex. (43-c), the elliptical clause involving a human locative argument, *sápmelaččain*, and an argument in accusative case, *noaiddástallanmáhtu*, can be considered the THEME of *ballat* 'fear'. Other elliptical constructions involve LOCATION-arguments of the place and time prototype category.

(43) a. Son ballá boastut **sáhttit** mannat muhtin beaivvi muhtin
s/he fears wrong can.INF go.INF some day some
guorbmebiilavuddjiin
truck.driver.COM
'S/he is afraid that someday things might go wrong with some truck driver'

b. Muhto son goitge balai **daid šaddat** heajos ovdamearkan
but s/he anyway feared they.ACC become.INF bad example.ESS
ássiide
inhabitant.ILL.PL
'But s/he was afraid of them becoming a bad example for the inhabitants'

c. ... dáčča t balle **sápmelaččain noaiddástallanmáhtu**
... Norwegians fear Sámi.LOC.PL magic.knowledge.ACC
'... Norwegians fear that Sámi have knowledge of magic'

13 different valency frames are categorized as ungrammatical valencies (12.40%), cf. Table 5.14. The majority of them are simple infinitive constructions, cf. ex. (44-a), which can be corrected by replacing the infinitive with an actio locative, or alternatively by adding a generic accusative, e.g. *dan* 'it (Acc.)' or *iežas* 'oneself (Acc.)', assuming that the sentence is an incomplete accusative + infinitive construction. Actio essive, i.e. progressive, and past participle constructions can also be incomplete when the accusative argument is missing. Accusative + non-finite form constructions can also include typos, leading to a nominative form instead of an accusative form, like *nieida* 'girl (Nom.)' instead of *nieidda* 'girl (Acc.)' in ex. (44-b). Alternatively, the verb form can be a finite verb instead of an infinitive as in the accusative + finite form construction in ex. (44-c). The second most frequent ungrammatical construction involves THEMES realized as finite subclause arguments of *ballat* 'fear', cf. ex. (44-d), which can be corrected by adding the subjunction *ahte* 'that'. Other ungrammatical constructions include THEMES in accusative, illative, locative and nominative case, which should be locative arguments. In addition, there are single instances of both infinitives combined with a subordinator and locative arguments combined with adpositions.

(44) a. *Ollugat garvet doaimmaid/lágidemiid go ballet
many avoid activity.ACC.PL/arrangment.ACC.PL because fear

Valency	***SIKOR***	**Example**
Ungrammatical constructions (as defined in this system) and corrections		
TH-Inf (Actio Loc) (+ Acc)	325	*ballet **olgguštuvvot**. ‘because they are afraid of being excluded.’ go ballet **olgguštuvvomis**. *ballá **čuohcat** garrasit sámegiela positiivvalaš ovdáneapmái ballá **dán čuohcat** garrasit sámegiela positiivvalaš ovdáneapmái
TH-Aktioess (+ Acc)	2	*ballat **čalmmostahttimin** fitnodagaid ‘are afraid of focusing on companies’ ballat **čalmmustahttimis** fitnodagaid
TH-PrfPrc (+ Acc)	9	*ballet **ožžon** dávdda ‘*they are afraid of having gotten the illness’ ballet **iežaset ožžon dávdda**
Nom + Inf	28	*ballat seamma **dilit čuožžilit** eará báikkiin ‘fear the same situations will arise in other places’ ballat seamma **diliid čuožžilit** eará báikkiin
Nom + PrfPrc (Acc + Prf-Prc)	3	*Mearrasámi **nieida** ballet **heavvanan** ‘*They fear the coastal Sámi girl to **have drowned**’ Mearrasámi **nieidda** ballet **heavvanan**
Acc + VFIN (Acc + Inf)	2	*balai **hearggi** su **fálleha**. ‘was afraid that the reindeer would attack her/him’ balai **hearggi** su **fallehit**
TH-FS (+ *ahte* ‘that’)	254	*ballá **sámegiella šaddá gievkkangiellan** ‘s/he is afraid Sámi becomes a language confined to the home’ ballá **ahte sámegiella šaddá gievkkangiellan**
TH-Acc (Loc.)	65	*Ballet **heakkaset** ‘Fear for one’s life’ Ballet **heakkaineaset**
TH-Ill (Loc.)	8	*Ii bala politihkalaš **mearrádussii**. ‘S/he is not afraid of the political decision.’ Ii bala politihkalaš **mearrádusas**
TH-Com (Loc.)	6	*ja seammas ballat **ođđa dieđuiguin**. ‘and at the same time they are worried about the news.’ ja seammas ballat ođđa **dieđuin**.
TH-Nom (Loc.)	1	*Veahá ballen **diekkár** ‘I was a little afraid of this’ Veahá ballen **diekkárin**
Nom + Loc (Acc + Loc)	1	*ballet menddo olu **soahkemuorra** Finnmárkkus ‘researchers fear there (is) too much birch wood in Finnmark’ dutkit ballet menddo olu **soahkemuora** Finnmárkkus
Loc + *ovddas* (Loc)	2	*dárbbaša **sajistis ovddas** ballat ‘need to worry about his position’ dárbbaša **sajiinis** ballat
Excluded forms		
derivations	46	**ballamis** lea ahte dárogiella sáhttá vuoitit ‘there is the fear that Norwegian can dominate’
syntax	7	dego **ballan** njoammil ‘like a frightened rabbit’
disambiguation error	50	**ballá** geahnoheabbu eret ‘the weaker one gets scared away’
real word error	8	*eai dárbbaš **ballat** eret ‘they do not need to get scared away’ eai dárbbaš **ballát** eret
Total	5,695	

Table 5.14: The valency distribution of *ballat* ‘fear’ in *SIKOR* (part 2)

olgguštuvvot.
exclude.PASS.INF
'Many avoid activities/arrangements because they are afraid of being excluded.'

b. *Mearrasámi **nieida** ballet **heavvanan**
coastal.sámi girl.NOM fear.PRS.3PL drown.PRFPRC
'They fear the coastal Sámi girl has drowned'

c. *balai **hearggi** su **fálleha**
fear.PRT.3SG reindeer.ACC s/he.ACC attack.PRS.3SG
'was afraid that the reindeer would attack her/him'

d. *ballá sámegiella **šaddá** gievkkangiellan jus
fear.PRS.3SG Sámi.NOM become.PRS.3SG kitchen.language.ESS if
gielddat jorgališgohtet unnit go maid dál dahket.
municipalities translate less than what now do
's/he fears Sámi will become a language confined to the home if the municipalities start translating less than they do now.'

5.2.3.1.6 Valencies of *dolkat* 'get fed up, be sick of'

The verb *dolkat* 'get fed up' is listed as a locative rection verb by Sammallahti and Nickel (2006, p.207), Mikalsen (1993, p.74), Nielsen (1932-1960*a*, p.561), and *Čállinrávagirji* (2003, p.87), cf. ex. (45-a). Sammallahti and Nickel (2006, p.207) also mention THEME-less constructions with past participle forms of *dolkat* 'get fed up', cf. ex. (45-b), and THEMES realized as infinitives. Mikalsen (1993, p.74) also discusses illative THEMES, cf. ex. (45-c).

(45) a. Mánát leat dolkan **guolis**.
children have get.fed.up.PRFPRC fish.LOC
'The children have gotten sick of the fish.' (*Čállinrávagirji*, 2003, p.87)

b. leat dolkan
be fed.up.PRFPRC
'be fed up' (Sammallahti and Nickel, 2006, p.207)

c. Mun lean dolkan **dutnje**
I am fed.up.PRFPRC you.ILL
'I am fed up with you' (Mikalsen, 1993, p.74)

Table 5.15 shows the distribution of *dolkat* 'get fed up' in *SIKOR*. With 681 occurrences, it is less frequent than *liikot* 'like', *luohttit* 'trust', *beroštit* 'care', and *ballat* 'fear'. Two instances of real word errors are excluded from the anlysis, one of which is *dolkomii* 'getting fed up (Ill.)', which is confused with *dulkomii* 'interpretation (Ill.)' in ex. (46).

(46) *Čuoládagat leat čavga čadnojuvvon govaid **dolkomii** ja ...
prints are closely related images interpretation.ILL and ...
'The prints are closely related to the interpretation of images and ...'

Valency	***SIKOR***	**Example**
Grammatical constructions (as defined in this system)		
TH-Ill	184	Eva lea veahá dolkan **skuvlii** 'Eva is a bit fed up with the school'
TH-Loc	91	Lea dábálaš dolkat **stohpobargguin** 'it is common to get fed up with housework'
Loc-aktio	89	várra dolkkai **gullamis** mu 's/he possibily got sick of **listening** to me'
RS-go	30	muhto várra dolkkai **go** ii ožžon makkárge dávástusa. 'but s/he possibly got fed up because s/he didn't get any response.'
TH-0	151	muhtomin dolká ge. 's/he also gets fed up sometimes.'
Ungrammatical constructions (as defined in this system) and corrections		
TH-Acc	2	*Dál gal leat nu hirbmasit dolkan iđit eahket campingvovnna-beatnagiid **čiellama** 'Now they are very fed up with the dogs in the caravans barking in the morning and evening'
(Loc.)		Dál gal leat nu hirbmasit dolkan iđit eahket campingvovnna-beatnagiid **čiellamis**
TH-ahte	2	*Sara lea dolkan **ahte** ealáhus šaddá guoddit jahkásaččat giellása nama. 'Sara is fed up that the industry is always seen as being full of liars.'
(+ *das*)		Sara lea dolkan **dasa ahte** ealáhus šaddá guoddit jahkásaččat giellása nama.
TH-Inf	124	*man dolkan Ole Niklas lei **johtit** biillain. 'how sick Ole Niklas was of traveling by car.'
(Actio. Loc.)		*man dolkan Ole Niklas lei **johtimis** biillain.
TH-Com (Loc.)	4	*leat dolkan **festiválaiguin** 'be fed up with the festivals' leat dolkan **festiválain**
TH-Ess	1	*Grand Prix-Máhtte dolkan **čájáhussan** '*Grand Prix-Máhtte is fed up as an exhibition'
(+ *leamis* Loc.)		Grand Prix-Máhtte dolkan **leahkimis čájáhusas** 'is fed up with being exhibited'
TH-ovddas	1	*leai dolkan filbmakamera **ovddas** birrajándoriid 's/he was fed up with (being) in front of the film camera the whole day'
(+ *leahkimis*)		leai dolkan **leahkimis** filbmakamera ovddas birrajándoriid
Real word errors		
real word error	2	leat čavga čadnojuvvon govaid **dolkomii** 'are closely connected to *getting fed up (Ill.)* with the pictures' Čuoládagat leat čavga čadnojuvvon govaid **dulkomii** 'are closely connected to the pictures' interpretation'
Total	681	

Table 5.15: The valency distribution of *dolkat* 'get fed up, be sick of' in *SIKOR*

22.17% of the occurrences appear without an argument. Realizations of the THEME include locative, illative, non-finite constructions, and finite subclauses. While only locative rection is considered grammatical, the distribution between locative (including actio locative forms) and illative case arguments is almost even: 27.02% of the occurrences of *dolkat* 'get fed up' appear with an argument in illative case, cf. ex. (47-a), and 26.43% appear with an argument in locative case, cf. ex. (47-b). 4.41% appear with a *go*-subclause, which can be thought of as the REASON-argument. However, *dolkat* 'get fed up, be sick of' frequently appears with a REASON-argument where a THEME is missing, cf. ex. (47-c). When searching for erroneous arguments in error detection, the REASON-argument can therefore be an indicator for a correct THEME-less construction.

(47) a. *Eva lea veahá dolkan **skuvlii**
Eva is a.bit fed.up school.ILL
'Eva is a bit fed up with the school'

b. Lea dábálaš dolkat **stohpobargguin**
is common get.fed.up house.work.LOC.PL
'It is common to get fed up with housework'

c. muhto várra dolkkai **go** ii ožžon makkárge
but possibly get.fed.up.PRT.3SG because not get.PRFPRC any
dávástusa.
response.ACC
'but s/he possibly gotten fed up because s/he didn't get any response.'

Erroneous constructions make up only 19.68% if illative constructions are not included, and 46.70% if illative arguments are included. An extensive study on grammaticality judgments of these valencies is an interesting research topic in itself but falls outside the scope of the present dissertation. The most frequent errors include infinitive constructions (18.21%), cf. ex. (48-a). These are also mentioned in Mikalsen's (1993) study of rection verbs (p.11). Ungrammatical constructions also include a THEME in accusative and comitative case, which should be realized by locative case, cf. ex. (48-a). Other ungrammatical constructions are subclauses with *ahte* 'that' that should be preceded by a locative antecedent.

(48) a. *muitala ge man dolkan Ole Niklas lei **johtit** biillain.
tells also how fed.up Ole Niklas was travel.INF car.COM
'tells also how fed up Ole Niklas was of traveling by car.'

b. *it vuos leat dolkan **festiválaiguin**
you first not fed.up.PRFPRC festival.COM.PL
'you have not got fed up with the festivals yet'

Additionally, there are elliptical constructions with an argument in essive case, cf. ex. (49-a), or an adpositional phrase with *ovddas* 'in front of', cf. ex. (49-b). In both cases the construction is missing a non-finite actio locative form of *leat* 'be', i.e. *leahkimis* 'be

Valency	**liikot**	**luohttit**	**suhttat**	**beroštit**	**ballat**	**dolkat**
Dominant	51.67% (Ill.)	70.25% (Ill.)	18.14% (Ill.)	69.7% (Loc.)	26.26% (*ahte*)	27.02% (Ill.)
2nd dominant	30.91% (Inf.)	1.03% (*ala*)	15.39% (*go*)	12.97% (Inf.)	22.06% (Loc.)	26.43% (Loc/Actio. Loc.)
0-valency	2.24%	1.55%	51.43%	6.42%	10.82%	22.17%
Total errors	6.92%	14.02%	0.12%	5.6%	12.40%	46.70%
Total instances	3,801	1,163	838	3,076	5,695	681

Table 5.16: Valencies of *liikot, beroštit, ballat, luohttit, suhttat, dolkat* in *SIKOR*

(Actio. Loc.)'.

(49) a. *Grand Prix-Máhtte dolkan **čájáhussan**
Grand Prix-Máhtte fed.up.PRFPRC exhibition.ESS
'Grand Prix-Máhtte is fed up with being exhibited'

b. *leai dolkan filbmakamera **ovddas** birrajándoriid
was fed.up.PRFPRC film.camera.GEN in.front.of whole.day.ACC.PL
's/he was fed up with (being) in front of the film camera the whole day'

5.2.3.1.7 Summary: Valencies

The six rection verbs differ in their frequency in *SIKOR*, as well as in the distribution of the typical valency described in the literature, cf. Table 5.16. Of the valency analysis of the six rection verbs, *ballat* 'fear' (5,695) is the most frequent verb, followed by *liikot* 'like' (3,801) and *beroštit* 'care' (3,076). The verb *dolkat* 'get fed up' has the highest percentage of errors, i.e. 46.70%, followed by *ballat* 'fear' (12.40%), and *luohttit* 'trust' (14.02%).

All of the six verbs are described as rection verbs and they are most relevant for error detection as they potentially trigger valency errors. However, some of the verbs frequently occur in THEME-less constructions: *suhttat* 'get angry' occurs without a THEME-argument in more than 50% of the cases. While the dominant valencies of most of the verbs coincide with the grammatical descriptions, some verbs occur more frequently in other constructions. The verb *beroštit* 'care' occurs more often in a subclause construction with *ahte* 'that' (26.26%) than with a locative argument (22.06%). The verb *dolkat* 'get fed up' occurs slightly more frequently with an argument in illative case (27.02%) than in locative case (26.43%). Mikalsen (1993, p.74) mentioned this tendency towards illative THEME early on.

As we saw in this section, an analysis of valencies interacts with morphological constraints (i.e. certain derivations that change the valency of a verb), syntactic constraints (e.g. attributive forms), disambiguation errors and real word errors in the targeted form.

The following section deals with disambiguation and methods that can be used to adapt the disambiguator for valency error detection.

5.2.3.2 Adapting disambiguation

Valency error detection requires correctly disambiguated input. In order to disambiguate potentially erroneous input, the following adaptations are made to the disambiguator of regular input. Syntactic context is necessary for finding syntactic errors, which is why correct disambiguation is crucial for error detection, both local and global. Although it is not necessary for the whole sentence to be analyzed correctly, a disambiguation error of the target itself or the governor of a particular argument can lead to false positives and false negatives. Birn (2000, p.33) comments on the relationship between disambiguation and grammatical error detection:

> On the one hand, it is obvious that disambiguation is a prerequisite for any effort at precise error detection. On the other hand, a grammar error may disturb the disambiguation, with either a disambiguation error or remaining ambiguity as a consequence, and this in turn may disturb the error detection.

The regular procedure consists of relaxing the rules of the disambiguation grammar and using homonymy sets for the disambiguation of particular ambiguities. Birn (2000, p.34) avoids extensive adaptations of the disambiguator by making form-specific error detection rules and accepting systematic ambiguities in the error rules. He further adjusts disambiguation rules where they are not specific enough by means of sets referring to forms with specific homonymies that should receive a certain analysis in a specific context. With respect to modifying the disambiguation grammar, I apply a more elaborate strategy including the following steps:

1. Remove default rules
2. Relax systematic homonymy rules
3. Include semantic and valency tags
4. Write more specific disambiguation rules for idiosyncratic homonymies

Default rules are rules that apply in the absence of any context condition. For a regular disambiguator it makes sense to remove a number of readings at the end of the disambiguation process if no particular conditions apply, i.e. establish a default. This is the case for comitative singular and locative plural forms, which are systematically homonymous. The homonymy is difficult to resolve, which is why there is a default rule, *KillCom*, as illustrated below, following a set of comitative-locative rules and selecting a locative if no other rule applies. However, for a grammar checker, the objective is not to reduce analyses, but rather to find the error despite existing ambiguities. This rule is therefore removed from *disambiguator.cg3*.

```
SELECT:KillCom (Pl Loc) IF (0 (Sg Com));
```

In ex. (50), the coordinated comitative argument of *gulahallan* 'understanding', *guolástuskomitéain* 'fishery committee (Com.)', is erroneously disambiguated as a locative, cf. l.15 of Figure 5.6. The noun *gulahallan* 'understanding' can have a comitative argument, which can only be mapped to its governor if the comitative is not disambiguated. As the comitative *komisearain* 'Commissioner (Com.)' in l.4 is a real word error and confused with *kommisearain*, and the comitative *Damanaki:n* in l.6 is not recognized by the morphological analyzer, the identification of *komisearain* 'Commissioner' as a comitative is particularly important. The homonymous form *guolástuskomitéain* 'fishery committee (Com.)' is wrongly disambiguated here. The comitative reading preceding a semicolon is discarded (l.16) and only the locative reading (without a semicolon) remains. Dependency analysis can therefore not be applied, and is analyzed as a default dependency pointing to itself (20->20). It should however point to *gulahallan* 'understanding' (20->14) and be annotated as a THEME. Without this annotation, an error detection rule can easily falsely annotate a valency error. Therefore, it is important that the default *KillCom*-rule is removed in *disambiguator.cg3*.

(50) ...gulahallan komisearain Damanaki:n ja Eurohpá
...communication Commissioner.COM Damanaki.COM and Europe
parlameantta **guolástuskomitéain**, ...
parliament fishery.committee.COM ...
'...communication with Commissioner Damanaki and European Parliament's fishery committee ...'

In a second step, disambiguation rules with context conditions for systematic, as opposed to idiosyncratic, homonymies need to be relaxed. These include homonymies between accusative and genitive case forms, locative plural and comitative singular forms, and infinitives and first person plural present tense verb forms. If one of the analyses of the systematic homonyms is correct in a certain context, while the other one is an error, false positives can be avoided when the forms are not disambiguated.

The first two steps lead to a reduction of disambiguation and an increase of possible analyses that are linked to a particular form. The third and fourth steps to modifying the disambiguator concern the enhancement and specification of disambiguation rules, this time to reduce the number of possible analyses for forms that are relevant as constraint conditions in error detection. The rules are enhanced by means of additional linguistic information, in particular semantic prototype categories and valency information, and general rules are split up into more idiosyncratic rules to improve their robustness. This is relevant for rules that are based on one particular constraint that is sensitive to the error itself. Disambiguation rules for adverbs and postposition/prepositions rely on a genitive to

```
"<gulahallan>"
        "gulahallan" N <CO-Com-Hum> Sem/Act Sg Nom #14->14
"<komisearain>"
        "komi#searra" Sem/Hum N Sg Com @ADVL> #15->14
"<Damanaki:n>"
        "Damanaki:n" ? #16->16
"<ja>"
        "ja" CC @CNP #17->17
"<Eurohpá>"
        "Eurohpá" N Prop Sem/Plc Sg Gen @>N #18->18
"<parlameantta>"
        "parlameanta" Sem/Build_Org §TH N Sg Gen @>N #19->22
        "parlameanta" Sem/Build_Org §TH N Sg Acc @OBJ> #19->22
"<guolástuskomitéain>"
        "guolástus#komitéa" Sem/Org N Pl Loc @ADVL>  #20->20
;       "guolástus#komitéa" Sem/Org N Sg Com
```

Figure 5.6: Ex. (50) syntactically analyzed by *GoDivvun*

the left (for postpositions) or right (for prepositions) and favor a preposition/postposition over the adverb reading, cf. the disambiguation rules below in ll.1–2. If the required genitive form has a typo, disambiguation cannot be performed successfully. Therefore, the disambiguation module of the grammar checker specifies idiosyncratic disambiguation rules for each adposition, and enhances them with semantic prototype tags and valency tags. The rule in l.4 selects a postpositional reading of *(n)alde* 'on' if it is directly preceded by a pronoun or a noun of the place prototype category. The rule in l.5, on the other hand, discards an adverbial reading of *(n)alde* 'on', unless a verb with a LOCATION-argument of the place prototype category in its valency (*<LO-Loc-Plc>*) can be found in the same context.

```
REMOVE Pr (NOT *1 Gen BARRIER NPNH) ;
REMOVE Po (NOT -1 Gen) ;

SELECT:GramPo (Po) IF (0 ("alde") OR ("nalde") LINK -1 Sem/Plc OR Pron) ;
REMOVE:GramPo (Adv) IF (0 ("nalde") OR ("alde"))(NEGATE *0 <LO-Loc-Plc>) ;
```

Homonymies of core elements involved in the global analysis of a sentence, especially, need to be resolved carefully with the help of semantic tags and valency tags in addition to syntactic context information. In the case of valency error detection, not only does the context of the governor need to be analyzed correctly, but also the governor itself. As the valency analysis showed, many of the verbs have homonymous forms, which can lead to falsely identified errors (false positives) if the verbs are wrongly disambiguated.

The previous section has shown that both disambiguation and real word errors influence the detection of valency errors. Most of the verbs have a number of unexpected

homonymies in particular forms, which is why it is important to find confusion pairs for disambiguation errors of significant parts of speech, i.e. governors of arguments that can produce valency errors. 154 of the occurrences of *luohttit* 'trust' are disambiguation errors (13.24%) and should be analyzed as *luhtte* 'at (Po.)' or another confused form. The distribution of the verb *ballat* 'fear', on the other hand, includes 50, i.e. only 0.88%, disambiguation errors, where the form should be traced back to another verb, or even another part of speech. However, disambiguation errors can lead to false positives in error detection, as the grammar checker assumes that the wrongly disambiguated form is the governor of certain arguments in the sentence, and consequently no correct forms can be identified.

In the case of ex. (51-a), *Ballo* is a surname and not a form of *ballat* 'fear', and consequently does not require the same arguments. In ex. (51-b), *luhtte* 'at' is a postposition and not a form of *luohttit* 'trust'.

(51) a. **Ballo** maid čuoččuha ahte Sámedikke presideanta ...
Ballo also claims that Sámi.Parliament president ...
'Ballo also claims that the president of the Sámi Parliament ...'

b. Sámi árra[n]iid **luhtte** gávdne maiddái asbe[a]sta keramihka ...
Sámi fireplace.GEN.PL by find also asbestos ceramics ...
'Asbestos ceramics are also found by Sámi fireplaces ...'

Certain forms of *ballat* 'fear' can also be confused with forms of the verb *ballát* 'get scared'. The following disambiguation rule selects a reading of *ballat* 'fear' instead of *ballát* 'get scared' unless it is followed by the adverb *eret* 'away':

```
SELECT ("ballat" TV) IF (0 ("ballát"))(NEGATE *1 ("eret") BARRIER NPNHA - Acc - Gen);
```

Disambiguation is followed by a set of local error detection rules and hence by partial dependency annotation to match correct arguments and their governors before searching for valency errors.

5.2.3.3 Governor argument dependency annotation

A syntactically analyzed and partly disambiguated sentence is the input to *grammarchecker.cg3*. The grammar checker runs real word error detection rules first, followed by dependency annotation rules before semantic role mapping and finally valency error detection is performed. Dependencies are not mapped to all parts of the sentence but mainly to correct arguments of verbal governors. This is done on the basis of valency tags, morpho-syntactic disambiguation and semantic prototype tags. For the dependency rules, I reuse a small set of dependency rules from *dependency.cg3* that map dependencies to arguments that are realized as finite subclauses and non-finite clauses.

Distance		**In between**	**Example**
3	*bargat* + NP	matrix verb	**Maid** don liikot *bargat* friddjabottuin? 'What (Acc.) do you like to do (Inf.) on your breaks?'
3	*ballat* + Inf	adverbials	Iige loga *ballat* gal TV:s **leat**. (cf. ex. (52-a))
11	*ballat* + FS	adverbials, relative clause, punctuation	**Balan** nu go Guovdageainnus, gos lea garra [le]stadi[á]nalaš osku, **lohket** mu jallas báhppan (cf. ex. (52-b))
12	*ballat* + NP	complex subject, matrix verb	Muhto **das** eaba loga Leme[t] Ánte Buljo ja Thor Thrane galgat sámi artisttaid *ballat*. (cf. ex. (52-c))

Table 5.17: Examples of North Sámi verbs and their arguments with different linear distances

New specific dependency rules map governors to arguments that are realized as nominal phrases, adpositional phrases, different types of subclauses, and simple and complex non-finite constructions. The rules are ordered according to linear closeness of the arguments. Dependencies are not mapped to subjects because they attach to the finite verb, which is not necessarily the semantic governor. The dependency rule set below starts out with simple rules that define an argument to the direct right (1) or left (-1) of their governor, e.g. an infinitive argument to the direct right of a verb with a *<TH-Inf>*-valency, cf. l.1. It is followed by another rule, cf. l.2, searching for an infinitive two tokens to the right of a potential governor unless there is an adjacent infinitive. General rules, cf. l.3, are specified later on, searching for an infinitive anywhere in the sentence unless another infinitive is closer to the governor.

```
1 SETCHILD <TH-Inf> TO (1 Inf);
2 SETCHILD <TH-Inf> TO (2 Inf)(NOT 1 Inf);
3 SETCHILD <TH-Inf> TO (*0 EOS LINK -1 ("?") LINK *-1 BOS LINK *1 Inf BARRIER Inf);
```

Constraint grammar rules have to take into account the linear structure of a sentence. The larger the linear distance between a governor and its argument, the more challenging correct dependency annotation is as other potential heads can be between the governor and its argument, and have to be ruled out, cf. Table 5.17.

If an argument has a larger linear distance to its governor than another form that is closer to the governor, the task becomes more challenging. In ex. (52-a), the THEME-argument of *ballat* 'fear' is not the locative *TV:s* 'on TV', but the infinitive *leat* 'be' (which should be an actio locative form, *leahkimis*, in written language according to *H*). However, if the locative is mapped to *leat* 'be' first, it becomes unavailable to *ballat* 'fear'. Therefore it is beneficial for a robust analysis to establish as many valency relations between governors in a sentence and their governors as possible. In addition, there can be

several potential arguments satisfying the valency restrictions of a governor. In ex. (52-b), the governor (*balan* 'fear (Prs. 1Sg.)') of the finite verb *lohket* 'claim (Prs. 3Pl.)' of the subsequent subclause is separated from its governor by other arguments, i.e. an inserted relative clause, *gos lea garra lestadiánalaš osku* 'where the Lestadian faith is strong'. In the accusative + infinitive construction in ex. (52-c), the verbal governor, *ballat* 'fear', is separated from its argument, *das* 'it (Loc.)', by a complex coordinated subject, i.e. *Leme[t] Ánte Buljo ja Thor Thrane* 'Leme[t] Ánte Buljo and Thor Thrane', the object *sámi artisttaid* 'Sámi artists (Acc.)', and the matrix verb *eaba loga*.

(52) a. *Iige loga *ballat* gal TV:s **leat**.
not claim fear.INF definitely TV.LOC be.INF
'S/he claims that s/he does not fear is not afraid of being on TV.'

b. **Balan* nu go Guovdageainnus, gos lea garra [lestadiánalaš]
fear.PRS.1SG just like Guovdageaidnu.LOC, where is strong Lestadian
osku, **lohket** mu jallas báhppan ...
faith, claim me crazy priest.ESS ...
'I'm afraid that they will say I am a crazy priest just like in Guovdageaidnu, where the Lestadian faith is strong ...'

c. Muhto **das** eaba loga Leme[t] Ánte Buljo ja Thor Thrane galgat sámi
but it.LOC not say Leme[t] Ánte Buljo and Thor Thrane should Sámi
artisttaid *ballat*.
artists fear.INF
'But Leme[t] Ánte Buljo and Thor Thrane say that Sámi artists should not be afraid of that.'

Table 5.18 shows the linear proximity between verbal governors and their arguments, i.e. their direct dependents. Since the results are based on a running dependency analysis of *SIKOR*[15] instead of the gold standard, the results need to be taken with a pinch of salt.

The results show that non-finite arguments are closest to their governors; 58% are directly adjacent. Their mean distance is 1.92. The mean distance of nominal objects to their verbal governors is 2.3, i.e. less than the mean distance between nominal adverbials and their governors, which is 2.89. More than 85% of all nominal objects are either directly adjacent to or separated by 2–3 tokens from their governors. These tokens can be adverbials, but they can also simply be parts of complex objects. Right hand nominal objects (2.1) are also closer (or less complex) than left hand nominal objects (2.51). However, right hand nominal adverbials (3.09) are more distant than left hand nominal adverbials (2.28). The mean distance of finite subclause arguments, on the other hand, is 5.02. 49% of all instances of finite subclause arguments have a distance of at least 5 tokens to their verbal governors. The distance is naturally larger as the finite verb of the

[15] Accessed 2014-11-17

Syntactic label	**1 (close)**	**2–4**	**5–>5 (far)**	**No/wrong analysis**	**Mean distance**
Objects					
right hand nominal objects (@<OBJ)	45.6%	44.9%	6.3%	3.1%	2.10
left hand nominal objects (@OBJ>)	31.3%	49.6%	9.2%	10.0%	2.51
left non-finite objects (@-FOBJ>)	95.7%	0.3%	0.1%	3.9%	1.01
right non-finite objects (@-F<OBJ)	52.9%	40.9%	2.2%	4.1%	1.70
non-finite clause (@ICL-OBJ)	0	91.2%	8.8%	-	2.84
finite subclause objects (@FS-OBJ)	0.1%	51.3%	48.9%	-	5.02
Adverbials					
right nominal adverbials (@<ADVL)	29.1%	50.2%	28.8%	18.4%	3.09
left nominal adverbials (@ADVL>)	48.9%	38.8%	9.0%	3.4%	2.28
unspecified nominal adverbials (@ADVL)	14.7%	19.4%	8.8%	57.2%	3.30
right non-finite adverbials (@-F<ADVL)	26.7%	54.4%	7.0%	0.6%	3.02
left non-finite adverbials (@-FADVL>)	56.5%	35.9%	6.3%	1.4%	1.94
left finite subclause adverbials (@FS-ADVL>)	1.2%	25.1%	61.9%	11.9%	6.58
right finite subclause adverbials (@FS-<ADVL)	0.1%	44.4%	55.6%	-	5.39
Generalizations					
nominal objects	38.4%	47.3%	7.8%	6.6%	2.30
nominal adverbials	30.9%	36.1%	15.5%	26.3%	2.89
non-finite arguments	58.0%	32.8%	3.9%	2.50%	1.92
subclause arguments	0.4%	40.3%	55.4%	32.1%	5.66
all arguments	33.6%	37.9%	20.3%	7.7%	3.16

Table 5.18: The average linear distance between verbs and their arguments in *SIKOR*
0 = pointing to itself
1 = pointing to the adjacent word to the left or right
2 = pointing to the second word to the left or right
etc.

subordinate clause is counted as the dependent of the governor. In the sentence it can be separated from its governor by a conjunction, subject, adverbials, etc. The column titled *No/wrong analysis* in Table 5.18 includes instances where the annotation has not been successful, and the dependency annotation is set to default. That means that the token is pointing to itself, e.g. 1->1 or 2->2, i.e. the first or second token in a sentence has itself as its syntactic governor. The column also includes combinations of syntactic labels pointing to a governor to the left and dependency annotations pointing to a governor to the right, and vice versa. The latter are excluded from the calculation of the mean distance between a governor and its arguments.

About 70% of the verbal governors are at least fairly close to their arguments, i.e. directly adjacent (34%) or at a distance of up to four (= three tokens between the governor and its argument) (38%). The mean distance to their nominal, non-finite and finite subclause arguments is 3.16 tokens.

Dependency rules not only map correct arguments to the governors in question, they also interact with each other. Even if a certain governor is not associated with any argument (the argument may be distant), a false positive can be avoided if other dependency relations between governors and their arguments in a sentence are annotated. A noun phrase becomes unavailable for error detection if it is the dependent of another governor. The dependency rule below sets the child of the locative verb *čohkkát* 'sit' in ex. (53) to the locative noun *sugadanstuolus* 'rocking chair (Loc.)', as it is to its direct left and a member of the furniture prototype category. The locative noun is therefore unavailable to an error detection rule that marks ungrammatical arguments of the illative noun *liikot* 'like'. This densification of linguistic analysis within a sentence makes the grammar checker more robust.

```
SETCHILD <LO-Loc-Plc> OR <LO-Loc-Any> TO (-1 Sem/Place + Loc OR Loc + Sem/Furn OR
Loc + Sem/Build OR Loc + Sem/Org);
```

(53) Mun nu liikon dán sugadanstuolus **čohkkát**.
I so.much like this rocking.chair.LOC sit.INF
'I like sitting in this rocking chair so much.'

Dependency mapping rules for accusative + non-finite verb constructions, cf. Figure 5.7, are the most elaborate as they have to consider two arguments that are not restricted in their linear organization and that can be separated from their governor by many other parts of the sentence. The most difficult dependency annotation involves non-finite constructions with both an infinitive/perfect participle or actio essive and an accusative subject as arguments of the verb, mainly because the accusative can be in different positions. The matrix verb is typically followed by the accusative and then the infinitive. However, the accusative subject can also be found after the infinitive (and after

the matrix verb), cf. ex. (54-a)–(54-b). In this case the first two rules, ll.1–4, annotate the required dependencies. If the infinitive is a transitive verb, the accusative can be either the subject or the object of the infinitive, which is another challenge in dependency annotation. In constructions like the one in ex. (54-c), the matrix verb can also be preceded by the accusative and followed by the infinitive. In its most extreme version it is separated from the infinitive not only by the matrix verb, but also by the subject of the matrix verb (cf. ll.6–10).

```
1  SETCHILD (V <TH-Acc-Any><TH-Inf>) (NEGATE 1 @-FOBJ> LINK p Inf OR PrfPrc) TO
2  (*1 Inf + IV BARRIER NOT-ADV LINK *1 Acc OR Gen BARRIER NPNHA);
3  SETCHILD (V <TH-Acc-Any><TH-Inf>) (NEGATE 1 @-FOBJ> LINK p Inf OR PrfPrc) TO
4  (c Acc OR Gen LINK *-1 Inf + IV BARRIER NPNHA);
5
6  SETCHILD (V <TH-Acc-Any><TH-Inf>) (NEGATE 1 @-FOBJ> LINK p Inf OR PrfPrc) TO
7  (*1 Inf + IV BARRIER NOT-ADV LINK *-1 @SUBJ> LINK *-1 Acc OR Gen BARRIER NPNHA);
8  SETCHILD (V <TH-Acc-Any><TH-Inf>) (NEGATE 1 @-FOBJ> LINK p Inf OR PrfPrc) TO
9  (c Acc OR Gen LINK *1 @SUBJ> BARRIER NPNHA LINK *1 <TH-Acc-Any><TH-Inf>
10 LINK *1 Inf BARRIER NOT-ADV);
```

Figure 5.7: Dependency rules for governors with accusative + infinitive valencies in *grammarchecker.cg3*

(54) a. . . . ii go *bala* **luottahuhttit čáhcelottiid**.
. . . not Q fear extinguish.INF waterbird.ACC.PL
'. . . is not afraid of water birds going extinct.'

b. . . . go ballet šaddat menddo olu **luossabivdiid**.
. . . because fear become.INF too much salmon.fisher.ACC.PL
'. . . because they fear that there will be too many salmon fishermen.'

c. Dan **maid** politiij[a] ja Suodjalus *balaiga* **leat** várálaš
it that.ACC police and military fear be.INF dangerous
soahtebázahussan
war.artifact.ESS
'What the police and military fear are dangerous artifacts from the war'

Dependency analysis is followed by semantic role mapping to unambiguously identify each argument in the valency of a governor.

5.2.3.4 Semantic role mapping

In *grammarchecker.cg3*, partial semantic role mapping is performed directly after dependency annotation as a further step to identify correct arguments and their governors. Semantic role mapping is partial as full semantic role mapping is not required in grammar checking. It is predominantly used within valency error detection. The rules target pronouns, nouns, verbs, adjectives, and numerals. Syntactically, mainly objects and adverbials are annotated. Determiners, nominal modifiers, and adpositional complements, on the other hand, are not targeted by the rules. Semantic roles are mapped to dependents of governors based on the governors' valency frames. As a governor typically can have different valency frames and a specific role can be realized in different ways, mapping semantic roles can generalize over different ways to satisfy the valency restrictions of a governor. And because only grammatical realizations of arguments receive a semantic role, semantic role specifications serve as constraints for valency error detection rules. The more roles that can be mapped correctly, the more robust the error detection process.

The following two rules map a THEME (*§TH*) to a noun/pronoun in illative case with a verbal parent that has the valency *<TH-Ill-Any>*.

```
SUBSTITUTE N (§TH N) TARGET N IF (p (V <TH-Ill-Any>))(0 Ill) ;
SUBSTITUTE Pron (§TH Pron) TARGET Pron IF (p (V <TH-Ill-Any>))(0 Ill) ;
```

Another set of rules annotates adjunct labels to secure adjuncts. This is beneficial for error detection, as these adjuncts can be discarded as targets of error detection rules. The rule below adds the adjunct label §*TIME-ADJUNCT* to a noun of the time prototype category if it is directly preceded by a genitive demonstrative pronoun or an adjective typically modifying a time expression (*TIME-A*).

```
SUBSTITUTE N (§TIME-ADJUNCT N) TARGET N IF (0 Sem/Time + Gen
LINK -1 (Pron Dem Gen) OR TIME-A) ;
```

In ex. (55-a), the adjunct label is added to *áigodaga* 'period (Acc.)'. This is relevant for valency error detection with regard to the illative rection verb *suhttat* 'get angry' as an accusative is a typical valency error in the context of illative rection verbs. However, here *suhttat* 'get angry' is used in a THEME-less construction. Time expressions in accusative case are not necessarily adjuncts, which is why the context needs to be specified in adjunct rules. In ex. (55-b), for example, the potential time adjunct *bearjadaga* 'Friday (Acc.)' is the subject of *šaddat* 'become'.

(55) a. Tigerat suhttet álkit dán **áigodaga**.
tigers get.angry easily this period.GEN;ACC
'Tigers get angry easily during this period.'

Potential argument	***liikot*-corpus**	***luohttit*-corpus**	***suhttat*-corpus**	***ballat*-corpus**	***beroštit*-corpus**	**Total**
nominal objects	58.17%	56.80%	59.81%	61.43%	62.18%	60.47%
pronominal objects	52.66%	60.85%	57.89%	63.78%	58.02%	54.16%
nominal adverbials	39.97%	37.66%	38.70%	33.03%	39.29%	43.53%
pronominal adverbials	50.42%	57.26%	51.75%	49.65%	45.86%	55.57%
Total						
	50.31%	53.14%	52.04%	51.97%	51.34%	53.43%

Table 5.19: Coverage of semantic role mapping for objects and adverbials in corpora of five rection verbs analyzed by *grammarchecker.cg3* version r116225

b. . . . ja nu ballat **bearjadaga** ge šaddat.
. . . and so fear Friday.ACC also become.INF
'. . . and we are afraid that Friday will also be like that.'

Table 5.19 shows the semantic role coverage for nominal and pronominal objects and adverbials analyzed by *grammarchecker.cg3*[16] in the corpora of the rection verbs in question. This version of *grammarchecker.cg3* contains 288 dependency rules and 112 rules adding semantic role tags to potential arguments, which are predominantly nominal, but also include adpositions, adverbs, numerals and adjectives. 52.04% of all objects and adverbials receive a semantic role, which means they are no longer possible targets of valency error detection rules. While technically all objects, as long as they are annotated correctly, should receive a semantic role, not all adverbials, e.g. sentence adverbials that do not occur in the valency of a governor, should receive a semantic role. This fact is also mirrored in the coverage, as object coverage is generally higher than adverbial coverage. For nominal arguments, there is a difference of 20% between object coverage (59.81%) and adverbial coverage (38.7%). For pronouns the coverage of semantic roles annotated to objects (57.89%) and adverbials (51.75%) only differs by 6%. Semantic role coverage for objects can definitely still be improved to provide optimal conditions for valency error detection, which will be discussed in the following section. Dependency and semantic role mapping is followed by valency error detection and correction rules.

[16] version r116225 (Accessed 2015-06-23)

5.2.3.5 Valency error detection and correction

The previous sections about dependency and semantic role mapping dealt with finding grammatically correct arguments of a governor. While finding correct forms is based on matching the valency tags of a governor to their respective arguments, finding incorrect forms involves a search for an error among the forms that have not been matched. However, due to partial dependency analysis and incomplete valency annotation not all correct forms are matched with their governors. Sentential adjuncts, subjects, postnominal modifiers, etc. are systematically not matched with their governors. Dependency analysis reduces the potential targets of error detection rules, and error detection rules need to find the ill-formed target based on this pre-selection and other context information. Error detection rules are typically based on a pairing of potentially ill-formed input and a well-formed alternative. Valency rules, as opposed to real word error rules, refer to tag sequences rather than to lemma-tag sequences. They target, for example, an accusative form that should be an illative form in a particular context. Table 5.20 gives an overview of possible Constraint Grammar rule types in grammar checking. The *grammarchecker.cg3*,[17] which is evaluated in Section 5.3, uses *ADD*-, *COPY*-, and *ADDCOHORT*-rules. *MOVE*- and *REMCOHORT*-rules, although used in earlier versions of *grammarchecker.cg3*, are excluded as they remove entire cohorts (i.e. word forms, their lemmata and all their possible readings, including morphological, syntactic and dependency tags) from the sentence and trigger a reapplication of the set of dependency rules. This leads to analysis errors of other forms of the sentences.

Error detection rules are *ADD*-rules, as illustrated below.[18] They add one or several error tags to a given target under given context conditions. The target, i.e. the potentially erroneous form, is specified after the *TARGET*-operator. Every error detection rule is accompanied by one or several error correction rules, the type of which depends on the correction type to be performed. *COPY*-rules copy an erroneous reading with its syntactic and morphological tags and replace the lemma and/or morphological tags with the correct lemma tag combination to generate the correct form. They also add the &*SUGGEST*-tag to mark the corrected status of this line. *ADDCOHORT*-rules, on the other hand, insert word forms, their lemmata and morphological tags at a given place specified by the operators *BEFORE/AFTER*. *REMCOHORT*-rules remove a given form and its analyses from the sentence. *MOVE*-rules move a given form (and its dependents) to a different position in the sentence again specified by the operators *BEFORE/AFTER*. The latter two rule-types are not described further here, as the version of the *grammarchecker.cg3* discussed in this chapter does not apply them.

A simple pair of *ADD*- and *COPY*- or *ADDCOHORT*-rules have the following format:

[17] version r118631 (Accessed 2015-08-13)

[18] For a complete overview of operators cf. the vislcg3 pages (`http://beta.visl.sdu.dk/cg3/single/#rules` (Accessed 2017-02-06))

Rule type (operator)	**Task**	**Schematic**
ADD	Adds a tag to a line of a cohort, e.g. an error tag to "3"	**1 2 (3) &msyn-valency-error 4 5 6**
COPY	Adds a new line to e.g. cohort "3" by copying an existing one and replacing certain parts of it	**1 2 (3) &ERRORTAG 4 5 6 &SUGGEST**
ADDCOHORT	Adds a completely new cohort into a sentence before or after e.g. "3"	**1 2 3 ↓ "<ahte>" &SUGGEST 4 5 6**
REMCOHORT	Removes a complete cohort of a sentence	**1 2 ~~3~~ REMCOHORT 4 5 6**
MOVE	Moves a cohort to another position in the sentence	**1 (3) ↶MOVE-BEFORE 2 ↶~~3~~ 4 5 6**

Table 5.20: Vislcg3 rule types used in error detection

```
ADD (&valency-error tag) TARGET ("erroneous form") IF
(forms to the left or right of the erroneous form);

COPY (corrected form &SUGGEST) EXCEPT (erroneous form &valency-error tag) TARGET
(&valency-error tag) ;

ADDCOHORT ("<form>" "lemma" morphological tags &SUGGEST) BEFORE/AFTER morphological
and syntactic tags IF (context conditions);
```

The target of an error rule can contain a lemma and/or a tag sequence describing the potentially erroneous forms. The valency error detection rules described in this section are based on the previous analysis of the valencies of the six rection verbs described in Section 5.2.3.1. Although the rules are made for erroneous valencies found in *SIKOR*, they are general rules for the erroneous valencies of any governor with the same valency tags. The rule's context conditions refer to the governor's valency frames that trigger the error, and potential disambiguation errors, non-word errors and real word errors of the potential argument or governors. They also refer to satisfied valencies of potential governors and specify exceptions for certain valency-altering derivations (e.g. causatives) or inflections (e.g. perfect participle) of the governor. Additionally, they specify other constructions the targeted form can appear in without being a valency error. Consequently, valency error detection rules are very complex rules with many positive and negative conditions. Figure 5.8 shows an error detection rule for accusative THEMES of governors with illative THEMES (*<TH-Ill-Any>*). This valency is annotated to 65 verbs in *valency.cg3*.

The rule maps an error to any noun phrase head (*NP-HEAD*), i.e. to any pronoun or

```
1
2  ADD:wrong-valency-ill-acc (&msyn-valency-ill-acc) TARGET NP-HEAD IF
3
4  (0 Acc OR Gen LINK NOT 0 Ill)
5
6  (*0 (V <TH-Ill-Any>) - <aux> - REAL-W-ERROR - Der/Pass - @>N
7  BARRIER GRAMCHK-S-BOUNDARY OR GRAMCHK-VFIN-NOT-AUX OR Inf OR @CVP
8  LINK NONE c §TH OR §IN OR (Acc §CO))
9
10 (NEGATE 0 @>N OR @>A LINK *1 Gen OR Acc OR SUBJ OR @ADVL BARRIER NPNH OR @CVP)
11
12 (NEGATE 0 @>N OR @>A LINK -1 @<OBJ)
13
14 (NEGATE 0 @>N OR @>A LINK 1 @<SPRED)
15
16 (NEGATE 0 @>P LINK *1 Po OR Pr BARRIER (*) - @CNP - @>P)
17
18 (NEGATE 0 @Num< OR @-FSUBJ> OR @P<)
19
20 (NEGATE 0 §ANYROLE OR §TIME-ADJUNCT)
21
22 (NEGATE p <TH-Acc-Any> OR <CO-Acc-Ani> OR <TH-Acc-Any><TH-Inf>)
23
24 (NEGATE 0* (V <TH-Ill-Any>)
25 BARRIER GRAMCHK-S-BOUNDARY OR GRAMCHK-VFIN-NOT-AUX OR Inf OR @CVP
26 LINK 1 (".*i"r) + ?)
27
28 (NEGATE 0* (V <TH-Ill-Any>) + PrfPrc
29 BARRIER GRAMCHK-S-BOUNDARY OR GRAMCHK-VFIN-NOT-AUX OR Inf OR @CVP
30 LINK *-1 ("leat") BARRIER NOT-ADV OR COMMA)
31
32 (NEGATE 0* (V <TH-Ill-Any>) + PrfPrc
33 BARRIER GRAMCHK-S-BOUNDARY OR GRAMCHK-VFIN-NOT-AUX OR Inf OR @CVP
34 LINK *-1 SUBJ BARRIER NOT-ADV OR COMMA LINK -1 ("leat"))
35
36 (NEGATE *0 &msyn-valency-dasa-before-ahte)
37
38 (NEGATE 0 Sem/Human LINK *1 (V <TH-Ill-Any>) + Sg3 BARRIER NOT-ADV-PCLE
39 LINK *1 Inf BARRIER GRAMCHK-S-BOUNDARY);
40
```

Figure 5.8: An error detection rules targeting a form in accusative case for verbal governors with a THEME in illative case in *grammarchecker.cg3*

noun, unless it is part of a compound, cf. l.2, under the subsequently specified conditions: the target needs to be an accusative or genitive case form that is not homonymous with any illative case form (cf. l.4). The close context needs to include a verb with a THEME in illative case in its valency (cf. l.6). The verbal governor cannot have a child (*c*) that is a THEME (*§TH*), an INSTRUMENT (*§IN*) or a CO-ARGUMENT (*§CO*) in accusative case, cf. l.8. The verbal governor cannot have an auxiliary, real word error, passive, or a prenonminal modifier reading, cf. l.6. Between the accusative/genitive form and the potential governor there cannot be a sentential barrier (*GRAMCHK-S-BOUNDARY*), e.g. a subordinator or relative pronoun, a finite verb (*GRAMCHK-VFIN-NOT-AUX*), an infinitive (*Inf*), or a global conjunction (*@CVP*), cf. l.7. There are several negative restrictions to both the context and the target itself. Each of these are introduced by the *NEGATE*-operator. The context's scope is restricted by barriers, which are based on incremental testing and partly parameterized as sets as e.g. *GRAMCHK-S-BOUNDARY*. A number of correct uses of genitive/accusative forms are discarded in ll. 10–18, where the genitive/accusative form can be a pre-nominal (*@>N*) or pre-adjectival modifier (*@>A*) of a subject, adverbial, object, subject predicative, a complement of a postposition (*@>P*) or a numeral (*@Num<*) in the respective syntactic context. The rule also includes a regular expression guessing non-words that are potential arguments in illative case, terminating in an *-i* (("*i"r) + ?), cf. l.26. The target itself is further restricted with respect to its semantic role, i.e. it should have been assigned neither a semantic role (*§ANYROLE*) nor an adjunct label (*§TIME-ADJUNCT*), cf. l.20. The second set of negative conditions (cf. ll.22–44) refers to the context of the target. The governor, i.e. (*p*) parent of the targeted form, should not have a valency frame referring to an argument in accusative case (e.g. *<TH-Acc-Any>*), cf. l.22. Nor should it be followed by a guessed illative (*LINK 1 ("*i"r) + ?*), cf. l.26. The governor in question should not be a perfect participle (*PrfPrc*) preceded by a subject (*SUBJ*) or a form of *leat* 'be' either, cf. ll.28–34. The penultimate condition in l.36 refers to another possible valency error, i.e. &*msyn-valency-dasa-before-ahte*, which is given preference over this valency error. The last condition refers to a position of the governor, where it normally does not appear with an argument, cf. ll.38–39. While the targeted accusative/genitive form is a member of any human prototype category, the potential governor is a third person singular form preceding an infinitive, i.e. a potential argument.

Error detection rules are paired with *COPY*-rules that refer to error correction by suggesting a form that should replace the targeted form. The error detection rule for infinitives after verbs with an actio locative valency adds an error tag, i.e. &*valency-aktioloc-inf*, to an infinitive THEME. This is the case in ex. (56-a), where the THEME of the governor *ballat* 'fear' is realized as an infinitive, *vuosihit* 'show'. The error tag includes a suggested correction, i.e. in this case the actio locative *vuosiheamis* 'show'. The *COPY*-rule below replaces the erroneous tag sequence (*Inf*) with a correct one (*Actio Loc*) and

adds the tag *&SUGGEST* to the changed reading. This sequence is the input for the normative morphological generator *generator-gt-norm.hfstol*. Ex. (56-b) includes the sentence with the corrected form.

```
COPY (Actio Loc &SUGGEST) EXCEPT (Inf &msyn-valency-aktioloc-inf)
TARGET (Sg &msyn-valency-aktioloc-inf) ;
```

(56) a. *SlinCraze čevllohallá leahtit sápmelaš, iige bala **vuosihit dan**.
SlinCraze be.proud be.INF Sámi, not fears show.INF it.ACC
'SlinCraze is very proud to be a Sámi, he is not afraid of showing it.'

b. SlinCraze čevllohallá leahkit sápmelaš, iige bala **dan vuosiheamis**. CORR

Other rules do not replace an ungrammatical form with a grammatical form, but rather insert a new form with all its possible analyses in the sentence, i.e. a new cohort. This is done by means of *ADDCOHORT*-rules. In ex. (57-a), the infinitive *šaddat* 'become' should not be replaced with a non-finite (actio locative) form. Instead, an accusative form should be added to it. The *ADDCOHORT*-rule below inserts an accusative subject, i.e. *dan* 'this'. The rule suggested correction the suggested form and its analysis, i.e. *"<dan>" "dat" Pron Dem Sg Acc &SUGGEST*, but also the place it is inserted, i.e. directly after a verb with a subordinate clause valency (*V <TH-ahte>*). In ex. (57-a), the accusative form is inserted after the verbal governor *ballat* 'fear', cf. ex. (57-b).

```
ADDCOHORT:wrong-valency-add-acc-inf ("<dan>" "dat" Pron Dem Sg Acc &SUGGEST) AFTER
(V <TH-ahte>) IF (*1 &msyn-valency-add-acc-inf BARRIER VFIN OR GRAMCHK-S-BOUNDARY) ;
```

(57) a. *...nu ahte Ruovdemáđiidoaimmahat ballá beare divrrasin
...so that railway.authorities fear too expensive.ESS
šaddat.
become.INF
'...so that the railway authorities are afraid it will become too expensive.'

b. ...nu ahte Ruovdemáđiidoaimmahat ballá **dan** beare divrrasin **šaddat**. CORR

5.2.3.5.1 Non-word and real word error context conditions

Both spelling errors resulting in non-words and real words can complicate the identification of a potential valency error. Therefore real word error rules are placed before valency error detection rules in *grammarchecker.cg3*. Firstly, spelling errors in the context can dissimulate correct arguments of a governor. Secondly, they can lead to false identifications of a potential governor. In ex. (58-a), the correct argument of *luhttet* 'trust (Prs. 3Pl.)' in illative case *Popa:i* cannot be recognized because of a spelling error. This can cause a false positive in error detection. The guesser described in Figure 5.8, ll.24–26, is used to identify the illative THEME in illative case and prevents the valency error

detection rule from searching for an error in other places. The constructions in ex. (58-b) and ex. (58-c) are accusative + infinitive constructions satisfying one possible valency of *ballat* 'fear'. However, because of a spelling error (i.e. it should be *tuberkulosa* instead of *tuberkolose* and *rábies* instead of *rabies*), the accusative subject is not recognized and the valency error detection rule for locative valencies (*&msyn-valency-loc-ill*) will normally search for a valency error in the illative form *Norgii* 'Norway (Ill.)'. In newer versions of *grammarchecker.cg3*,[19] these forms receive an error tag analysis, e.g. *tuberkolosa Err/Orth N Sg Acc*, and can be matched with their governors. This prevents other rules from searching for valency errors.

(58) a. Dušše 1% dahje 4 jienasteaddji iskkadeamis luhttet **Popa:i**.
just 1% or 4 voters investigation trust Popa.ILL
'Just 1% or 4 voters in the investigation trust Popa.'

b. ...balle ***tuberkolose** čuožžilit gait gávjjiis ja ribais.
...fear tuberculosis emerge.INF all dust.LOC.PL and splinter.LOC.PL
'...they feared tuberculosis may develop because of all the dust and splinters.'

c. Ballet ***rabies** njoammut Norgii
fear.PRS.3PL rabies spread.INF Norway.ILL
'They fear rabies may spread to Norway'

Real word errors affect both potential arguments and potential governors within valency error detection. In ex. (59), *čohkkat* 'mountain top (Nom. Px2Sg.)' is a real word error for the infinitive *čohkkát* 'sit'. The valency of the verb *liikot* 'like', however, is satisfied by an infinitive, which is why it is essential that the real word error is corrected before applying the valency error detection rule.

(59) In liiko jaska ***čohkkat** beare guhká, dadjá son.
not like quietly mountain.top.NOM.PXSG2 too long, says s/he
'I do not like to sit quietly too long, s/he says.'

The following Constraint Grammar analysis of ex. (59) illustrates that the corrected form, i.e. *"čohkkát" V TV*, suggested by the grammar checker (*&SUGGEST*) receives the correct dependency (*#5->3*) and semantic role (*§TH*) from the real word error correction rule, cf. ll.9. The successful real word annotation prevents the valency error detection rule for verbs with an illative valency like *liikot* 'like' from being applied, and thereby avoids a false alarm.

```
1  "<In>"
2          "ii" <aux> V IV Neg Ind Sg1 @+FAUXV #2->2
3  "<liiko>"
4          "liikot" <mv> V IV Ind Prs ConNeg @-FMAINV #3->3
5  "<jaska>"
```

[19] version r155175 (Accessed 2017-07-18)

```
6         "jaska" Adv @<ADVL #4->4
7   "<čohkkat>"
8         "čohkka" Sem/Plc-elevate N Sg Nom PxSg2 @SUBJ> &real-čohkkát #5->3
9         "čohkkát" Sem/Plc-elevate @SUBJ> V TV §TH Inf <LO-Loc-Any> &SUGGEST #5->3
10  "<beare>"
11        "beare" Adv @<ADVL #6->6
12  "<guhká>"
13        "guhká" Adv Sem/Time @<ADVL #7->7
```

Potential governors can also be real word errors. As verbal governors in particular trigger valency error detection rules, their correct identification is essential in valency error detection. In the case of ex. (60-a), the verb that triggers an illative error detection, i.e. *luohttit* 'trust', is a real word error of a compound noun error. Unless the real word error is resolved, the form is likely to cause a false positive in error detection. The form should be written as one word with the subsequent noun, i.e. *luohteimprovisašuvnnain* 'yoik improvisation (Com.)'. In ex. (60-b) *ballan* 'fear (PrfPrc.)' should be *ballán* 'be scared (PrfPrc.)' or 'become afraid and run away' according to Nielsen (1926-1929, p.125), cf. ex. (60-b). The indicator is the adverb *eret* 'away'. Although the confused forms are related and of the same part of speech, their valencies differ. While *ballát* 'get scared' often appears with a DESTINATION-argument, *ballat* 'fear' does not have an illative argument in its valency. In ex. (60-c), the argument of *ballen* 'fear (Prt. 1Sg.)' is *čohkkat* 'mountain top (Px2Sg.)'. However, the noun form *čohkkat* is a real word error for the verb *čohkkát* 'sit'. *Ballen* 'fear (Prt. 1Sg.)' is also a real word error, and should be *bállen* 'be in peace (Prt. 1Sg.)' but this can only be recognized after identifying *čohkkat* 'mountain top (Px2Sg.)' as an intended infinitive. This shows that real word error rules also interact with each other.

(60) a. Konsearttas maid koarra lihkostuvai ***luohte** improvisašuvnnain.
concert.LOC also choir succeeded trust improvisation.COM
'In the concert, the choir also succeeded with the yoik improvisation'
Konsearttas maid koarra lihkostuvai **luohteimprovisašuvnnain**. CORR

b. Eanas ábegáhtut leat dál ***ballan** eret ...
most monkeys have now feared away ...
'Most monkeys then got scared and left ...'
Eanas ábegáhtut leat dál **ballán** eret ... CORR

c. ... man ollu čuoikkat diibmá ledje, ii ***ballen** baljo jaska
... how many mosquitos last.year were, not fear.PRT.1SG almost still
***čohkkat**.
mountain.top.NOM.PXSG2
'... how many mosquitos there were last year, one almost could not sit still.'
... man ollu čuoikkat diibmá ledje, ii **bállen** baljo jaska **čohkkát**. CORR

5.2.3.5.2 Semantic role context conditions

Error detection rules not only refer to other potential errors, but also test the semantic roles of the target of the rule and the context. The previous valency error detection rule for verbal governors with a THEME in illative case in Figure 5.8 tests other correct realizations of the THEME-role. As dependencies and semantic roles are mapped in a step prior to error detection, any correct realization of the THEME should be matched with the verbal governor. Therefore the child (c) of the verbal governor is tested for a THEME in l.1 of the condition below, which is an excerpt from the rule in Figure 5.8. THEME-less constructions are typical if other argument types in accusative case, i.e. an INSTRUMENT or CO-ARGUMENT, are matched with the governor. Therefore, they are also included in the negative conditions to the child of the verbal governor, cf. l.1. The target itself is also tested. If it is annotated with a semantic role or an adjunct label, it is identified as either the correct argument of any governor or an adjunct that should not be a part of the valency of any governor. The rule therefore specifies a negative condition to the target of the rule in l.3 of the condition below.

```
1 LINK NONE c §TH OR §IN OR (Acc §CO))
2
3 (NEGATE 0 §ANYROLE OR §TIME-ADJUNCT)
```

Generally, governors have more than one possible valency. The verb *ballat* 'fear' has at least 23 different valencies, cf. Table 5.13 in Section 5.2.3.1. All valencies have to be tested to match any correctly realized valency before searching for an error. Also, possible errors need to be explicitly defined. Possible erroneous forms of arguments of *ballat* 'fear' are accusatives, comitatives, nominatives, infinitives and finite verbs. However, depending on the construction, accusative forms can be grammatical or ungrammatical. In ex. (61-a), the form *huksemiid* 'constructing (Acc. Pl.)' should be a locative form (*huksemiin*). However, in ex. (61-b), the accusative *suohkana* 'community (Acc.)' is part of an accusative + locative construction where the infinitive *leat* 'be' is omitted. Furthermore, in accusative + infinitive constructions such as the one in ex. (61-c), the accusative form *dan* 'it (Acc.)' is correct. Therefore, error detection rules refer to semantic roles that are mapped to correct arguments of a potential governor and exclude them as possible targets of the error rule.

(61) a. *NBR ii bala viivvuhis **huksemiid**
NBR not fear unregulated constructing.ACC.PL
'NBR does not fear unregulated constructing'

NBR ii bala viivvuhis **huksemiin** CORR

b. ... gii ballá **Snåasa áidna suohkana** Lulli-Sámis mii
... who fears Snåasa only municipality.ACC South-Sápmi.LOC that
heive ...
fits ...

'. . . who fears Snåasa to be the only municipality in South Sámi that fits . . . '

c. **Dan** ballat erenoamážit **sáhttit** vahágahttit boazodoalus . . .
it.ACC fear specifically might.INF harm.INF reindeerherding.LOC . . .
'They fear this might harm reindeer herding in particular . . . '

5.2.3.5.3 Rule ordering

Valency error detection rules are applied in a certain order in *grammarchecker.cg3*, cf. Table 5.21. Later rules can refer to earlier rules and can refrain from being applied if an earlier rule has been applied.

The rule order depends on the complexity of the sentence. Generally, rules for more complex arguments, e.g. subclauses, hit before rules targeting nominal or adpositional phrases as many verbal governors can have valency frames with both simple and complex arguments. If there is a subclause of the wrong type, there should not be a rule looking for an error in any nouns or pronouns, which is why the subclause rule should hit first. In general, there are three main groups of rules, one detecting errors in verbal arguments, a second one detecting errors that are based on illative valencies, and a last group detecting errors that are based on locative valencies. Seven rules hit nouns and pronouns, 13 rules hit verbs, two rules hit specific adpositions (i.e. *mielde* 'with' and *birra* 'about'), one rule hits adverbs and one rule hits adjectives. A very specific rule for pronouns followed by relative pronouns hits first. The context is very immediate which is why it needs to be an early rule. It is followed by a set of rules for subclause arguments that should either be infinitival or introduced by a subordinating conjunction (*ahte* 'that' or *go* 'that, when') depending on their valency. Then, a set of rules for illative valency hits where accusative, locative or comitative forms should be illative forms. Next a set of rules adds missing infinitives, illative forms and replaces postpositions with an illative form. A set of rules for locative valencies replaces accusative and comitative forms with locative forms. This is followed by a set of rules for non-finite constructions, changing either the infinitive form or the subject of the infinitive form and adding non-finite forms. A rule for adpositional arguments changing them into locative case forms follows. The last rule is used for nominal heads (derived from verbs) marking the infinitive with an error. It hits late because verbal dependencies seem to be stronger than those of nominal heads.

Certain potential valency error types that were described earlier in this chapter are not covered by any rules. These include case errors in coordinated constructions as in ex. (62-a), where both the verbs *bargat* 'work' and *beroštit* 'miss' are matched with a comitative THEME even though *beroštit* 'miss' requires a locative THEME. The valency requirements of both *bargat* 'work' and *beroštit* 'miss' are only satisfied in the non-elliptic construction in ex. (62-b). Examples of coordinated governors with different valencies could be found in *SIKOR*. In ex. (62-c), the argument of both verbs is realized first as a locative (*mas*) and then as an illative (*masa*). While the valency error in ex. (62-a)

Rule	**Target**	**Erroneous form**	**Correct form**
valency-ahte-not-fs	V	(SUBJ OR ADVL) finite verb	*ahte* 'that' + (SUBJ OR ADVL) + finite verb
valency-ill-nom	Pron	Nom	Ill
valency-dasa-before-fs	V	*maid* + finite verb	*dasa maid* + finite verb
valency-das-before-fs	V	*maid* + finite verb	*das maid* + finite verb
valency-ahte-not-fs	V	(SUBJ OR ADVL) finite verb	*ahte* 'that' + (SUBJ OR ADVL) + finite verb
valency-inf-not-fs	V	finite verb	Inf
valency-go-not-fs	V	finite verb	*go* 'that' + finite verb
valency-ill-acc	N, Pron	Acc	Ill
valency-add-leat	Adv	Place adverb	Place adverb + *leat* 'be'
valency-ill-loc	N, Pron	Loc	Ill
valency-ill-com	N, Pron	Com	Ill
valency-dasa-before-ahte	CS	*ahte* 'that'	*dasa ahte*
valency-ill-ovddas	Po	Gen *ovddas* 'for'	Ill
valency-aktioloc-inf	V	Inf (Acc)	(Gen) Actio Loc
valency-loc-acc	N, Pron	Acc	Loc
valency-acc-inf-not-nom-inf	V	Nom + Inf	Acc + Inf
valency-add-acc-to-ess	A	Ess	Acc + Ess
valency-ahte-inf	V	Inf	*ahte* 'that' + finite verb
valency-go-inf	V	Inf	*go* + finite verb
valency-add-acc-inf	V	Inf	Acc + Inf
valency-loc-com	N, Pron	Com	Loc
valency-aktioloc-aktioess	V	ActioEss	ActioLoc
valency-add-acc-prfprc	V	PrfPrc	Acc + PrfPrc
valency-loc-not-birra	Po	Gen *birra* 'about'	Loc
valency-not-inf	V	Inf	-
valency-acc-not-nom	N	Nom	Acc

Table 5.21: Valency error detection rules in the order they can be found in *grammarchecker.cg3*

is obvious, the coordinated construction in ex. (62-d) is ambiguous. The accusative *dili* 'situation' could be the THEME of *buoridit* 'improve' only, in which case *beroštit* is THEME-less. Alternatively, *dili* 'situation' can be the THEME of both verbs, and a valency error of *beroštit*.

(62) a. *...sis lea ovdamoraš áššiin **maiguin** barget ja
...they.LOC have concern thing.LOC.PL which.COM.PL work and
beroštit.
care
'...they are worried about the things they work with and care about.'

b. ...sis lea ovdamoraš áššiin **maiguin** barget ja **main** beroštit. CORR

c. ...fuobmájin ahte ollu **mas** mun beroštan ja **masa** liikon,
...realized that many which.LOC I care and which.ILL like,
lea áitojuvvon.
is threatened
'...I realized that many things I care about and like are threatened.'

d. ...organisašuvnnat álget beroštit ja buoridit mánáid beaivválaš
...organizations start care and improve children's daily
dili johtolagas.
situation.ACC traffic.LOC
'...the organizations start to care (about) and improve the children's daily situation with the traffic.'

5.2.3.6 Summary: Global error detection

Global error detection in *grammarchecker.cg3* includes different types of valency error detection rules, both for simple and complex arguments. They target, amongst others, case forms, adpositional phrases, non-finite clauses, and different types of finite subclauses. Both semantic prototype categories and valencies are available to global error detection rules on various stages. As in local error detection rules they can be used in simple context conditions. However, in contrast to real word and adpositional errors, valency errors cannot be found without valency information. Valency tags and semantic prototype tags are the backbone of valency error detection. Valency error detection includes the following steps: adapting the disambiguator, performing incremental partial dependency annotation of verbal governors and their arguments, annotating the semantic roles of these arguments, and finally performing error detection, all of which make use of both valency and semantic prototype information. As valency error detection rules refer to real word errors and other local errors, they also indirectly use valency and semantic prototype information in disambiguation and in simple references. While valency error detection rules are general rules and can be applied for any verbal governors with the valency in question, I designed the valency error rules based on a test set of six rection verbs, for which I did a detailed valency analysis based on *SIKOR*. The analysis includes

both grammatical and ungrammatical valencies, non-word errors, real word errors and morphological forms that are excluded because of alterations in the verb's valency. The following section deals with both local and global error detection.

5.3 Evaluation

This section deals with the quantitative and qualitative evaluation of the three previously discussed error types, i.e. real word errors, local case errors, and valency errors. The quantitative evaluation includes the three measures precision, recall, and accuracy. While precision only evaluates the (alleged) errors detected by the system, recall also evaluates the errors that should have been detected by the system. Calculating recall therefore comes at a much bigger cost than calculating precision and is typically done on a much smaller corpus. When developing a commercial grammar checker, keeping the number of false alarms low is one of the main goals, "even at a noticeable loss of recall" (Arppe, 2000). Consequently, precision is the key measure in evaluation. Lastly, accuracy also evaluates the non-detection of correct forms.

Full-fledged (commercial) grammar checkers like *Grammatifix* typically have a specific threshold for the performance of the grammar checker. According to Arppe (2000, p.17), there should be "a precision of over 67 percent for each error type was chosen, i.e. two-thirds of flaggings for each error type should be justified in order for the error type to be included in the final product". Overall precision of *Grammatifix* is 70% (Birn, 2000, p.38). In Hagen and Lane's (2001) evaluation, the Norwegian Grammar (NGC) reaches a precision of 75%. Other relevant grammar checking modules with global syntactic rules like *XUXENg* for Basque local syntactic error detection in complex postpositions (i.e. case errors of nouns in postpositional phrases) reach a precision of 50.5% when evaluated on a corpus of 85 instances of 5 different postpositions. Other previously mentioned approaches do not document any results that are relevant for this evaluation.

Typically, one distinguishes between a corpus for the development of a grammar and a second corpus for evaluating the rules. Corpora for evaluation can consist of real or generated text depending on the error type. It is common to generate a corpus for errors related to a single form. For real word error detection and local case error detection, like Pedler (2007), I generated an error corpus for evaluation by replacing the correct confusion pair member with the incorrect one. For valency error detection, only real errors are used both for development and evaluation. Half of the corpus including instances of the respective verbs is used for development and the other half is used for evaluation.

The following sections deal with the evaluation of six real word error rule sets that involve semantic tags and valency tags in error detection, and adposition error rules that are disambiguated on the basis of semantic tags and valency tags. They also deal with the evaluation of valency error detection rules that are based on semantic prototype categories

and valency tags in disambiguation, dependency analysis, semantic role annotation and error detection.

5.3.1 Evaluation of real word error detection

In this section, I evaluate the real word error detection rules for the six confusion pairs discussed in Section 5.2.2.1, cf. Table 5.7, in *grammarchecker.cg3*.[20] All of the rules include semantic prototype categories, valency tags or both, and have real examples for both of the confusion pair members in *SIKOR*. While the corpus of real examples is used for constructing rules, a generated corpus switching around the confusion pair members in that corpus is used for evaluation. The test corpus for evaluation contains 20,541 confusion pair instances.

5.3.1.1 Quantitative evaluation

Table 5.22 presents the quantitative evaluation results of the real word error rules for the six confusion pairs in question, i.e. (*várra;varra*), (*ádde*;adde**), (*sáhtáš*;sáhtaš**), (*čohkke*;čohke**), (*biddjui;bidjui*), and (*čohkká*;čohkka**). Some of these (those marked with an asterisk) include rules for more than one form. The confusion pair (*ádde*;adde**), for example, covers the confusion pairs (*ádde;adde*), (*áddet;addet*), and (*ádden;adden*). Each confusion pair has rules for both members of the confusion pair, i.e. both frequent and rare forms are associated with real word error rules. Rules that target the same form and suggest the same corrected form receive the same name even if they rely on different context conditions, e.g. rules with the name *real-várra* target the form *varra* 'blood' and correct it with the form *várra* 'maybe'. In Table 5.22, the results for the rule correcting the more frequent real word error is presented first, i.e. the rule *real-várra* identifying *varra* 'blood' as a spelling error of *várra* 'maybe' precedes the rule *real-varra*. Typically, the form that has more real word error instances is also the less frequent one in *SIKOR*, cf. Table 5.7 in Section 5.2.2.1. Table 5.22 shows that more frequent real word errors are also represented by more rules than their less frequent confusion pair partners. The rule types *real-várra*, *real-ádde*, *real-várra*, *real-čohkke*, and *real-čohkká* have more rules than their respective conterparts *real-varra*, *real-adde*, *real-čohke*, *real-bidjui*, and *real-čohkka*. The only exception is the real word error type pair *real-sáhtáš-real-sáhtaš*, which are both equally represented by two rules only. The error detection rule type *real-ádde* has the most error detection rules (66). The rule set *real-ádde*, for example, covers a group of three real word error rules: *real-ádde*, *real-áddet*, *real-ádden*.

Precision for all six real word error confusion pairs is between 97% and 100% for all rule types except for *real-sáhtáš* (80%). While precision is high, recall is significantly lower in most of the cases. It is generally higher for rules treating the more frequent real error

[20] version r129849 (Accessed 2016-02-22)

Real word error rule	Target	Rules	Precision	Recall	Accuracy
Adverb vs. noun					
real-várra	*varra* 'blood'	37	99.75%	70.67%	71.06%
real-varra	*várra* 'maybe'	10	96.64%	43.73%	95.14%
Verb vs. verb					
real-ádde	*addet*, *adden*, *adde* 'give'	66	99.92%	75.13%	75.26%
real-adde	*áddet*, *ádden*, *ádde* 'understand'	15	99.92%	60.27%	61.47%
real-sáhtáš	forms of *sáhtašit* 'give a ride'	2	99.81%	98.10%	97.93%
real-sáhtaš	forms of *sáhtášit* 'can'	2	80.00%	57.14%	99.22%
real-čohkke	*čohket*, *čohken*, *čohke* 'sharpen (Prs. 3Pl., Prt. 1Sg., Prs. 1Du.)'	25	99.84%	93.11%	92.88%
real-čohke	*čohkket*, *čohkken*, *čohkke* 'collect (Inf., Prs. 1Sg., Prs. 3Sg.)'	1	100%	100%	100%
Verb vs. noun					
real-biddjui	*bidjui* 'den (Ill.)'	7	100%	96.96%	96.96%
real-bidjui	*biddjui* 'put (Pass. Prt. 3Sg.)'	1	100%	61.29%	80.00%
real-čohkká	*čohkka* 'mountain top'	17	99.56%	91.66%	91.33%
real-čohkka	*čohkká* 'sit (Prs. 3Sg.)'	3	100%	20.00%	82.61%
TOTAL		186	97.95%	72.34%	86.99%

Table 5.22: A quantitative evaluation of six real word error detection rules in *grammarchecker.cg3*

of the confusion pair, except for the rule pair *real-čohkke-real-čohke*, where recall is over 90% in both cases. *Real-čohkka* is the only rule with a recall lower than 50% (its recall is 20%). This is due to the lack of real instances of the real word error in *SIKOR*. Accuracy, as opposed to precision and recall, also includes true negatives. Accuracy is above 90% for most of the real word error rules. For *real-várra*, *real-ádde*, *real-adde*, *real-bidjui* and *real-čohkká* it lies between 61% and 83%. In the next section, I will go into detail about possible causes of high and low precision, recall, and accuracy.

5.3.1.2 Qualitative evaluation

In this section, I will evaluate the reasons for differences in precision, recall, and accuracy. I distinguish between seven general causes of unsuccessful error detection, cf. Table 5.23. These can be other errors in the sentence, e.g. non-word errors, real word errors, compound errors, valency errors and punctuation/formatting errors. They can also be shortcomings of the grammar checker, including disambiguation errors or real word error rule shortcomings.

CAUSE	Example	Corrections
ERRORS IN THE SENTENCE		
non-word errors	Dan *__galggáše__ maiddái politiijat addet. 'The police **should** also understand this'	*galggaše* 'should (Prs. 3Pl.)' should not have an accent on the <a>
real word errors	*__Mu__ ádden lieđážiid dan dihte go mu mielas son dárbbašii daid '**I** (Acc.) understand flowers because in my opinion s/he needed them'	*mun* 'I (Nom.)'
compound errors	Dan lassin lea juolluduvvon prošeaktaruhtadeapmi máŋgga konkrehta *doaibmabiddjui* dáin suohkaniin. ...	*doaibmabidjui* should be written as one word
valency errors	Čilgejeaddji sárgumat ja somás historjjálaš tevnnagat *álkidahttet* girjji __[á]ddet__.	*álkidahttit* 'simplify' does not have an infinitive valency
punctuation/-formatting errors	oppalaš tearbma **berg** mii ii govčča sámegiela __čohkká__ semantihkalaš sisdoalu.	*"berg"* ... *"čohkka"* should be in quotation marks
ERRORS IN THE SYSTEM		
disambiguation errors	*Oahppi* (V. PrsPrc.) **bidjui** čuoigat okto 10 kilomehtera amas meahcis bearráigeahču haga.	*oahppi* 'student (N. Nom. Sg.)'
error rule shortcomings	Dattetge **adde** Laiti *ahte* ođđa áigi rievdada dárbbuid sámeniibbi ektui ja Strømeng niberávdái.	the subordinating conjunction *ahte* 'that' should be referred to in the rule to recognize the real word error *adde*

Table 5.23: A qualitative evaluation: causes of unsuccessful real word error detection in *grammarchecker.cg3*

5.3.1.2.1 Non-word errors

Non-words can cause both false positives and false negatives in error detection, especially if a relevant clue is a non-word error. Real word error detection of *biddjui* 'put (Pass. Prt. 3Sg.)', which should be *bidjui* 'den (Ill.)', relies on a genitive modifier of the animal prototype category. In ex. (63-a), there is a real word error in *vilgesšáhpaniid* 'white mice (Gen. Pl.)', which is a member of the animal prototype category. The form is missing an accent on the second <a> and should be *vilgesšáhpániid*. As the relevant clue is missing the real word error in *biddjui* 'put (Pass. Prt. 3Sg.)' is not recognized. In ex. (63-b), the relevant clue for identifying *addet* 'give (Prs. 3Pl.)' as a real word error of *áddet* 'understand (Inf.)' is the auxiliary *galggáše* as it typically appears with an infinitive. However, as it is a non-word, the real word error cannot be identified.

(63) a. Vulge go áhkku ja dat eará sáhp[á]nnieiddat daid
leave Q grandmother and the other mouse.girls the
***vilgesšáhpaniid** biddjui, ...
white.mice.GEN.PL put.PASS.PRT.3SG, ...
'Do Grandma and the other mouse girls leave to the white mice's den, ...'

b. *Dan **galggáše** maiddái politiijat addet.
this should also police give.PRS.3PL
'The police should also understand this.'

5.3.1.2.2 Compound errors

Compound errors can also get in the way of successful error detection. Since *oasus spekulašuvnnat* 'stock speculations' in ex. (64-a) is not written as one word (it should be *oasusspekulašuvnnat*), the nominative reading of *spekulašuvnnat* 'speculation (Nom. Pl.; Acc. Sg. Px2Sg.)' is removed by the disambiguator in favor of the accusative case possessive reading. *Oasus* 'stock' is left as the only nominative and potential subject of the clause. The nominative plural form is a potential plural subject for *addet* 'give (Prs. 3Pl.)'. However, the nominative plural reading that serves as a clue for identifying *áddet* 'understand (Inf.)' as a real word error of *addet* 'Prs. 3Pl.' (which it agrees with in number) is removed. In ex. (64-b), the compound error concerns the form *doaibma-biddjui*, which should be written as one word and with a single consonant (*doaibmabidjui*). Consequently, the form *biddjui* 'put (Pass. Prt. 3Sg.)' is falsely identified as a real word error of *bidjui* 'den (Ill.)'.

(64) a. ***Oasus spekulašuvnnat** áddet buori vuoittu.
stock.NOM speculation.NOM.PL understand.INF;PRS.1PL good profit
'Speculating with stock gives a good profit.'

b. Dan lassin lea juolluduvvon prošeaktaruhtadeapmi máŋgga
it addition has allocate.PASS.PRFPRC project.funding many
konkrehta ***doaibma- biddjui** dáin suohkaniin.
concrete implementation.ILL these municipality.LOC.PL

‘In addition, project funding is allocated to many concrete implementations in these municipalities.’

5.3.1.2.3 Real word errors

Clues for real word error detection can also contain other real word errors that hinder successful error detection. In ex. (65), the first person singular pronoun *mun* ‘I’ is confused with *mu* ‘I (Gen.; Acc.)’. However, the first person singular subject serves as a clue for the real word error detection of *ádden* ‘understand (PrfPrc.; Prs. 1Sg.)’, which should be *adden* ‘give (Prt. 1Sg.)’. As the relevant clue is missing, the real word error cannot be identified, resulting in a false negative.

(65) *__Mu__ ádden lieđážiid dan dihte go mu
I.GEN understand.PRFPRC;PRS.1SG flower.ACC.PL because I.GEN
mielas son dárbbašii daid ...
opinion.LOC s/he needed them ...
‘I gave him/her flowers because in my opinion s/he needed them ...’

5.3.1.2.4 Valency errors

Valency errors can also prevent error detection rules from applying correctly. In ex. (66), the verb form *álkidahttet* ‘simplify (Prs. 3Pl.)’ does not have an infinitive in its valency and is consequently not annotated with the valency tag *<TH-Inf>*. However, the infinitive valency of a verb preceding *addet* ‘give (Prs. 3Pl.)’ serves as a clue to identify the real word error in *addet*. The real word error rule does not apply, and the result is a false negative.

(66) historjjálaš tevnn[e]gat *álkidahttet* girjji *__addet__.
historic drawings simplify book.ACC give.PRS.3SG
‘the historic drawings make it easier to understand the book.’

5.3.1.2.5 Formatting/punctuation errors

Formatting and punctuation errors, too, can lead to unsuccessful real word error detection. In ex. (67), both *berg* and *čohkká* ‘sit (Prs. 3Sg.)’ are cited, which is why they need to be in nominative case. However, the citation marks serve as a clue for them being nouns in nominative case. The real word error *čohkká* ‘sit (Prs. 3Sg.)’, which should be *čohkka* ‘mountain top’, is not recognized, and the result is a false negative.

(67) *...oppalaš tearbma **berg** mii ii govčča sámegiela **čohkká**
...general term berg which not cover Sámi sit.PRS.3SG
semantihkalaš sisdoalu.
semantic content
‘...the general term “berg” which does not cover the semantic content of Sámi “čohkka (i.e. mountain top)”.’

5.3.1.2.6 Disambiguation errors

Disambiguation errors are common causes of unsuccessful error detection. Disambiguation errors include both erroneously discarded readings and insufficient disambiguation, which fails to discard readings that do not match the pattern of the error detection rules. I will provide examples of both types. In a sentence containing a grammatical error, the chance that a token is wrongly disambiguated is much higher than in a grammatically correct sentence. If a grammatical error in a finite verb leads to the absence of a finite verb reading in a sentence, the disambiguator will naturally search for another finite verb reading in any other token. In ex. (68-a) the real word error *bidjui* 'den (Ill.)' should be a finite verb *biddjui* 'put (Pass. Prt. 3Sg.)'. As the finite verb *biddjui* 'put (Pass. Prt. 3Sg.)' cannot be identified as such because of the real word error, the infinitive *čalmmustuhttit* 'point out (Inf.; Prs. 3Pl.)' is wrongly disambiguated as a present tense third person plural form, i.e. finite verb. This again prevents correct real word error detection, and leads to a false negative. While disambiguation in ex. (68-a) concerns two related forms of the same lemma, disambiguation can also concern unrelated forms or even syntactic disambiguation. In ex. (68-b), *varra* 'blood' is a real word error and should be *várra* 'maybe'. It appears in a coordinated finite clause. Coordination in general can be local (coordinating e.g. noun phrases) or global (coordinating finite verb phrases). Generally, *varra* 'blood' is considered to be a noun if coordinated with another noun of the body or substance prototype category like *sáttu* 'sand'. While the local context suggests that *varra* 'blood' is correct, the global context shows that *ja* 'and' is a global coordinator between two main clauses and *varra* 'blood' is a real word error. However, as *varra* 'blood' is a noun, the disambiguator erroneously removes the global coordinator analysis of *ja* 'and', which is why the real word error is not recognized. Here, a syntactic disambiguation error causes a false negative.

(68) a. . . . erenoamáš deaddu **bidjui* **čalmmustuhttit** ahte
. . . special emphasis den.ILL point.out.INF;PRS.3PL that
Sámedikki doaimma váldoáŋgiruššan leat sámi nissonat.
Sámi.parliament activity main.focus are Sámi women
'. . . special emphasis is put on pointing out that Sámi women are the main focus of the Sámi parliament's activities.'

b. Ammahal sálti lea hálbbit go **sáttu ja** ***varra** eai dárbbaš čorget
supposedly salt is cheaper than sand and blood not need clean
muohttaga eret . . .
snow away . . .
'I suppose salt is cheaper than sand and they probably do not need to clean the snow away . . . '

There are also idiosyncratic unrelated homonymies that can lead to unsuccessful error detection, if they involve one reading that can trigger error detection, whereas the

alternative reading does not trigger it. The homonymy-pairs of the forms *manne* 'why (Adv.); pick.up.eggs (Prs. 3Sg.); go (Prt. 3Pl.)', *geassit* 'drag (V. Inf.); during the summer (Adv.)', and *amas* 'so that not (Conj.); foreign (A.)' are typical examples. In ex. (69), the third person singular of *mannet* 'pick up eggs' instead of the infinitive *mannat* 'go' is chosen assuming that *varra* 'blood' is a nominative singular subject agreeing with the finite verb in third person singular. However, this again prevents the real word error detection rule for *varra* 'blood' to apply and causes a false negative.

(69) **Varra* **manne** ieža maid Njávdámis, ...
blood collect.eggs.PRS.3SG oneself also Njávdán.LOC ...
'They probably also took a trip to Neiden themselves , ...'

Lastly, I will present an example in which insufficient disambiguation rather than erroneously discarded readings prevents successful error detection. In ex. (70) the homonymy of *eaiggádat* 'owner (Nom. Pl.); own (Prs. 2Sg.)' remains unresolved. However, only the first reading agrees with the correct form *addet* 'give (Prs. 3Pl.)' in number and person. The second reading, on the other hand, is a finite verb itself suggesting that the subsequent form should not be a finite verb. Because of insufficient disambiguation, the real word error detection rule for *áddet* 'understand (Inf.)', which should be *addet* 'give (Prs. 3Pl.)', does not apply, resulting in a false positive.

(70) ...**gárddiid** eaiggádat **áddet* fálaldaga sámi
...farm.GEN.PL owner.NOM.PL;own.PRS.2SG understand.INF offer.ACC Sámi
mánáide ...
child.ILL.PL ...
'...the farms' owners make offers to Sámi children ...'

5.3.1.2.7 Error detection shortcomings

Shortcomings of the error detection rules themselves are the largest source of unsuccessful real word error detection. Typically, context specifications are insufficient, and/or barriers are wrongly defined. In ex. (71-a), the real word error *adde* 'give (Prt. 3Pl.)' is not recognized because of insufficient specifications in the error detection rule itself. Because of an error in the barrier, the form *ássit* 'inhabitant (Nom. Pl.)' is recognized as a potential subject of *adde* 'give (Prt. 3Pl.)' although it is separated from it by a subordinating conjunction, i.e. it belongs to another clause. This causes a false negative. The barrier needs to refer to a subordinating conjunction to resolve the problem.

(71) a. Nystad ***adde** ahte gieldda ássit eai leat
Nystad give.PRT.3PL that municipality.GEN inhabitant.NOM.PL not have
diehtán dáid ođđa njuolggadusaid birra.
known these new rule.GEN.PL about
'Nystad understands that the municipality's inhabitants did not know about

these new rules.'

5.3.1.3 Conclusion regarding real word error detection

Real word error detection is placed before dependency and semantic role analysis in *grammarchecker.cg3*, based on the assumption that its rules mainly operate with local contexts. Secondly, as valency analysis has shown, real word error detection and correction must be carried out prior to establishing governor argument relations to ensure that a potential governor is not a real word error for another form. While real word error detection is generally considered local error detection, it can be more global than anticipated when local contexts of the confusion pair members are similar, cf. Table 5.22. When local contexts of the confusion pair members are similar, global contexts need to be referred to in order to detect the error. References to global constraints are also possible without establishing dependency relations and adding semantic roles. Instead of referring to a particular child or parent, one can search for particular combinations of disambiguated forms and tag sequences in an unrestricted left or right context of the sentence. However, these constraints are less robust, as only the context of the real error is singled out, whereas the dependency module within *grammarchecker.cg3* performs general governor-argument mapping of all parts of the sentence. Semantic prototypes and valency information are used in simple context conditions of the real word error rules (186) of six confusion pairs evaluated in the previous section. The mean precision of 98% is significantly higher than Arppe (2000)'s threshold of 67% for each error type. Both recall (72%) and accuracy (87%) are lower than precision, which is in line with improving precision at the possible cost of recall. Qualitative evaluation has shown that other error types are closely connected to real word error detection and can influence its outcome. Spelling errors should be annotated with error tags if their forms can be anticipated, rather than be left without analysis, in order to establish an analyzed context for the real word error. This has been done in newer versions of *grammarchecker.cg3*.[21] Disambiguation is also intertwined with real word error detection and often causes false negatives if a key form is wrongly disambiguated because of the error. Improving disambiguation and including both valencies and semantic prototypes to make it more robust is therefore necessary for future work. Furthermore, both regular and idiosyncratic homonymies can be listed systematically to avoid possible errors. While rule relaxing avoids the removal of correct homonyms, it can prevent the application of error detection rules, where error detection requires precise context information of the forms in question. Disambiguation rules should therefore undergo not only a process of relaxation, but also improvement by making more idiosyncratic rules and enhancing them with as much linguistic layers as possible. Precise context information, i.e. within complex barriers, is important for error detection as well,

[21] version r157816 (Accessed 2017-10-02)

and rules should be precise rather than too general. The above analysis and evaluation has shown that real word error detection is a heterogeneous task. It can be a global process, and in some cases, e.g. to distinguish between forms of *áddet* 'understand' and *addit* 'give', requires deep syntactic analysis including a previous dependency analysis matching verbal governors and their arguments.

5.3.2 Evaluation of local case error detection

Here, I evaluate the case error detection in adpositional phrases with five adpositions that are disambiguated by means of valency tags and semantic prototype tags in *disambiguator.cg3*[22] and error detection in *grammarchecker.cg3*.[23] These are *(n)ala, (n)alde, badjel, bokte* and *rastá*, cf. Table 5.24. I published the original study based on an earlier ver-

Adposition	**Translation**	**Homonymy**
(n)ala	'onto'	postposition, adverb
(n)alde	'on'	postposition, adverb, verb (aldat 'get closer (Prs. 1Du.; Prt. 3Pl.)')
badjel	'over'	postposition, preposition, adverb
bokte	'via'	postposition, verb (boktit 'wake (Prs. 1Du.; Prt. 3Pl.)')
rastá	'across'	postposition, preposition, adverb

Table 5.24: The homonymies of five North Sámi adpositions, cf. Wiechetek (2012, p.39)

sion of *grammarchecker.cg3*[24] in Wiechetek (2012). Since then, the system architecture of *GoDivvun* has changed and now includes more modules, cf. Figure 5.9.

The text is first tokenized and morphologically analyzed by the descriptive morphological analyzer *tokeniser-gramcheck-gt-desc.pmhfst*, which has access to the North Sámi lexicon with both error tags and semantic tags. The valency annotation grammar *valency.cg3* then adds valency tags to potential governors. It is followed by a newer constraint grammar module (than the evaluation in Wiechetek (2012)), *mwe-dis.cg3*, which can undo compound readings of multi-word expressions based on the morpho-syntactic context and valencies. The next module is *disambiguator.cg3*, which performs a morpho-syntactic analysis and disambiguation. The output of the disambiguator is analyzed by the module *grammarchecker.cg3*, which performs error detection and correction. The correct morphological forms are then generated from tag combinations suggested in *grammarchecker.cg3* by means of the normative morphological generator *generator-gt-norm.hfstol*. Also, the

[22] version r157345 (Accessed 2017-09-21)
[23] version r157681 (Accessed 2017-09-29)
[24] version r53901 (Accessed 2012-02-10)

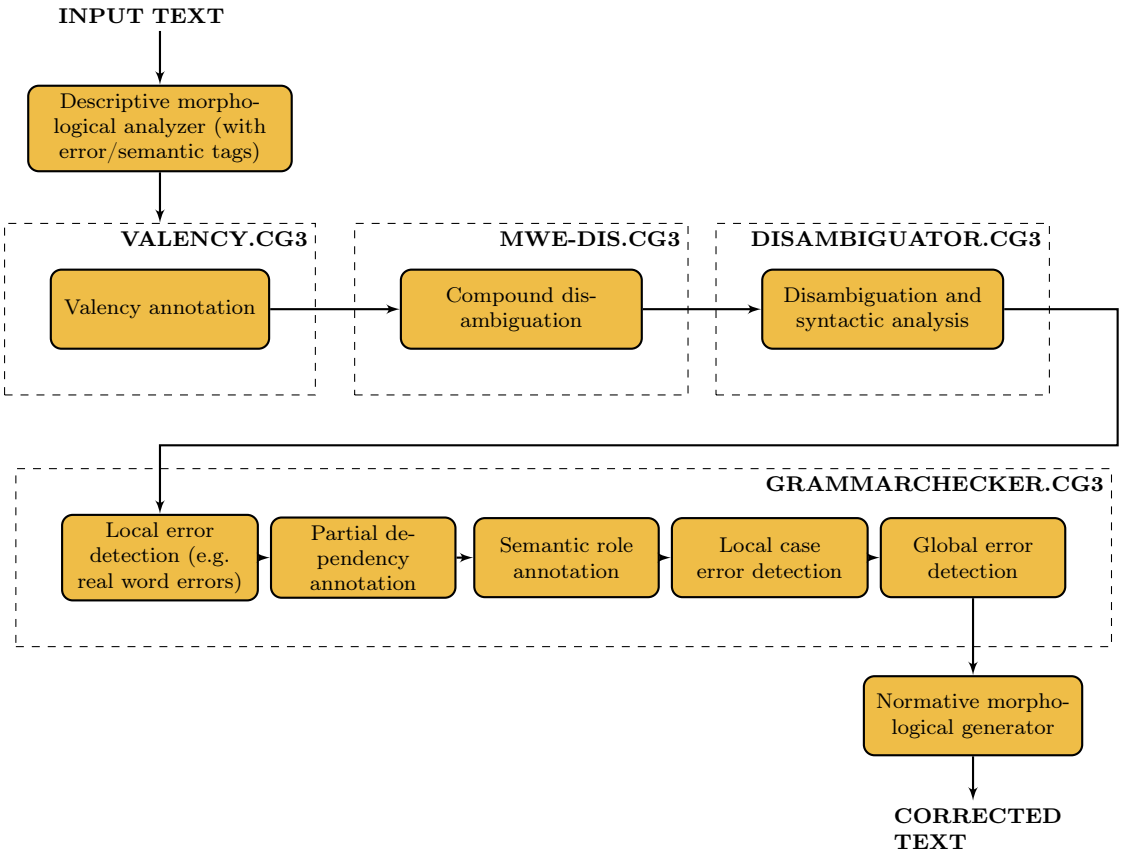

Figure 5.9: The system architecture of *GoDivvun* (version r157681)

internal structure of *grammarchecker.cg3* is more complex, and local case error detection takes place after local error detection, governor-argument dependency analysis, and semantic role mapping, but before global error detection.

Adpositional case error detection rules rely heavily on the disambiguation of the potential adpositions. The adpositional disambiguation rules are idiosyncratic, i.e. each adposition has its own set of rules. Disambiguation rules and error detection rules have not experienced any major modifications. However, a number of other features in the system architecture of *GoDivvun* have changed. It includes more and more detailed semantic tags (e.g. *Sem/Plc-water* and *Sem/Plc-line* instead of *Sem/Plc*), error annotation of real word errors and non-words (including error tags like e.g. *Err/Orth*, *Err/Orth-nom-acc*, and *Err/Orth-nom-gen*), valency annotation, and other error rules.

5.3.2.1 Quantitative evaluation

The error detection process is tested on sentences from *SIKOR*[25] containing the respective adpositions. This evaluation corpus of local case error detection rules contains 32,460 tokens and consists of 200 sentences for each adposition, 100 original sentences and 100 sentences (most of them the same) in which the genitive case of the adposition's dependent is exchanged with any other case (illative, locative, nominative). Apart from a few adaptations and additions due to inconsistences, the same corpus used in Wiechetek (2012) is used for this evaluation. Table 5.25 shows the results of the evaluation with *grammarchecker.cg3*[26] in comparison to the previous results presented in Wiechetek (2012). Precision is 98.88%, higher than in the previous analysis (where it was 97.21%). Recall is lower than precision (80.97%) and slightly lower than in Wiechetek (2012) (where it was 82.93%). Accuracy (90.13%), on the other hand, has stayed almost the same as in the earlier evaluation (where it was 90.55%). The amount of false positives is highest for *badjel* 'over' (four instances), which, together with *rastá* 'across', is the adposition with the highest homonymy. The amount of false negatives is highest for sentences containing *badjel* 'over' (65), *rastá* 'across' (59) and *ala* 'onto' (45). Due to changes in the morphological analyzers, a number of non-words and forms containing case errors receive the respective annotation. The form *eatnan* 'property', for example, receives not only its regular nominative analysis, but also a genitive/accusative reading accompanied by the error tag *Err/Orth-nom-acc*. In ex. (72), the nominative reading is further removed by means of the disambiguator *disambiguator.cg3*. The error is not marked by the error detection rule, but by means of the morphological analyzer. This is counted as a true positive for error detection.

(72) . . . go in dohkket šat johtolaga rastá mu priváhta
. . . because not accept anymore traffic.ACC across my private
eatnan . . .
property.NOM;.ACC.ERR/ORTH-NOM-ACC . . .
'. . . because I do not accept any more traffic across my private property . . . '

[25] `http://giellatekno.uit.no/doc/lang/corp/corpus-sme.html` (Accessed 2012-02-10), 18,142,181 tokens, mostly newspaper text

[26] version r157681 (Accessed 2017-09-29)

grammarchecker.cg3	**version r53901 (2017)**	**version r157681 (2012)**
False positives	10	23
True positives	804	802
False negatives	189	165
True negatives	1,013	1,000
Precision	98.88%	97.21%
Recall	80.97%	82.93%
Accuracy	90.13%	90.55%

Table 5.25: A quantitative evaluation of local case error rules in *GoDivvun* (2017) and (2012)

5.3.2.2 Qualitative evaluation

Here, I will evaluate the performance of local case error detection quantitatively, and point out different causes of false positives and false negatives. The main causes of false positives and false negatives are disambiguation errors (67.3%) and error detection rule shortcomings (31.2%), cf. Table 5.26. 40.7% of the disambiguation errors are related to the adposition itself and 26.6% are related to the adposition's dependent. As local case error detection in adpositional phrases relies heavily on successful disambiguation of the adpositions, it is not surprising that disambiguation errors are the main causes of false alarms.

Type	**Absolute**	**%**
FALSE DISAMBIGUATION		
Missing disambiguation	45	22.6%
Confusion: postposition - preposition	20	10.1%
Confusion: adverb - adposition	15	7.5%
Confusion: adposition - verb	1	0.5%
Erroneous disambiguation of the adposition's dependent	53	26.6%
Total (disambiguation)	134	67.3%
FALSE ERROR DETECTION		
Unsuccessful error detection	62	31.2%
OTHER ERRORS		
Non-word	1	0.5%
Real word	2	1.0%
Total	199	-

Table 5.26: A qualitative evaluation: causes of unsuccessful local case error detection

In 22.6% of the cases, missing disambiguation rules or insufficient specificity within the rules causes the lacking disambiguation of the potential adposition. In ex. (73), the preposition *rastá* 'across' is not disambiguated. This leads to a false negative in the detection of the case error in *biilaluodda* 'road (Nom.)'. As in the case of erroneous

disambiguation, the cause is a missing semantic tag in the disambiguation rules. Adding the semantic tag *Sem/Route* to the disambiguation rules for *rastá* 'across' resolves the problem.

(73) *...ja vázzilii rastá **biilaluodda**.
...and walked across road.NOM
'...and s/he walked across the road.'

Disambiguation errors of the potential adposition are predominantly confusions of an adverbial and adpositional reading, leading to false negatives if the adpositional reading is not recognized, and to false positives if an adverb is falsely disambiguated as an adposition.

In ex. (74-a), the case error in *duottarvárit* 'tundra mountain (Nom. Pl.)' is not recognized because *badjel* 'over' is wrongly disambiguated as an adverb, whereas it is an adposition here. The disambiguation causes a false negative in error detection here. In ex. (74-b), on the other hand, *gákti* 'costume' is wrongly associated with a case error because *nalde* 'on' is disambiguated as an adposition, whereas it is an adverb here. In this case, the disambiguation causes a false positive in error detection.

(74) a. *Mánáidgárdeoahpaheaddji Laila Aleksandersen mánáiguin ovdalaš
kindergarten.teacher Laila Aleksandersen child.COM.PL just
go **duottarvárit** garra bieggan galgat badjel.
before tundra.mountain.NOM.PL strong wind should over
'Kindergarten teacher Laila Aleksandersen and the children just before we should go over the tundra mountains in strong wind.'

b. Sus geas lea **gákti** nalde lea Emilia Henrietta ...
she.LOC who has costume.NOM on is Emilia Henrietta ...
'The one wearing the costume is Emilia Henrietta ...'

The most difficult task is to disambiguate prepositional and postpositional readings, in particular if the potential adposition is both preceded and followed by a noun. In ex. (75-a), *rastá* 'across' is a preposition. However, it is disambiguated as a postposition. Therefore, the case error in *johka* 'river (Nom.)' (it should be genitive case) is not recognized, resulting in a false negative in error detection. Disambiguation is unsuccessful because *johka* 'river' has a newer (i.e. newer than the tools used in *GoDivvun* (2012)) semantic tag, i.e. *Sem/Plc-water* instead of the general tag *Sem/Plc*. In ex. (75-b), on the other hand, *rastá* 'across' is analyzed as a preposition instead of a postposition because it is followed by a noun in genitive case, i.e. *bohccobierggu* 'reindeer meat'. Therefore the case error in *rájit* 'borders (Nom. Pl.)' is not recognized, resulting in a false negative in error detection. Semantic tags are particularly important for the disambiguation of pre- and postpositions. Certain semantic prototypes need to be given predominance over others as one cannot rely on the correctness of the morphological case of the noun phrase. Nouns with place prototype membership like *rádji* (*Sem/Plc-line*) in ex. (75-b) need to

be given priority over nouns of the food prototype category like *bohccobiergu* 'reindeer meat (Gen.)' when disambiguating the pre- and postposition reading of *rastá* 'across'.

(75) a. *Jiehtanas lei nu suhttan ahte ii lean dilli gállit rastá **johka**
giant has so get.angry that not had time wade across river.NOM
'The giant got so angry that s/he did not have the time to wade across the river'

b. *Norgga bohccobierggu importalávdegoddi lea veardideam[e]n luoitit
Norwegian reindeer.meat import.committee is considering let
fas **rájit** rastá bohccobierggu ...
again border.NOM.PL across reindeer.meat.GEN ...
'The Norwegian reindeer meat import committee is considering letting reindeer meat get across the borders again ...'

Disambiguation errors in the adposition's dependent can also cause false alarms. In ex. (76-a), *sildi* 'bridge (Nom.)' is falsely identified as a deverbal noun (derived from *sildit* 'separate') in genitive case instead of a noun in nominative case. In ex. (76-b), *dat* 'it' is falsely disambiguated as a particle instead of a demonstrative pronoun and the case error is therefore not recognized.

(76) a. ***Sildi** bokte lei luodda buot jiekŋan, ...
bridge.GEN;separate.GEN via was road all ice.ESS, ...
'Near the bridge, the road was all covered in ice, ...'

b. ***Dat** bokte lea ollu álkit muittašit gos lea leamaš ...
it.NOM;PCLE via is much easier remember where has been ...
'By means of this it is much easier to remember where it has been ...'

The second most common cause for false positives and false negatives are error detection rule shortcomings. This is due in part to complex noun phrases that contain a genitive that is falsely considered to be the dependent of the adposition. This is the case in ex. (77-a), where *juolgelahpi* 'foot' is pre-modified by a genitive *olbmá* 'man (Gen.)'. This could be resolved by means of dependency rules that match the noun phrase heads with their dependents and thereby make them unavailable for dependency annotation as dependents of the adposition. The same is true in ex. (77-b), where the adposition's dependent *ipmárdusrájit* 'understanding border (Nom. Pl.)' is preceded by the genitive *min* 'our' and therefore the case error in *ipmárdusrájit* 'understanding border (Nom. Pl.)' is not recognized. In ex. (77-c), on the other hand, a case error is attributed to *miljovnnaid* 'million (Gen. Pl.; Acc. Pl.)', which is a false positive. However, the numeral is not a part of the noun phrase dependent of *rastá* 'across' here, but rather an object of the verb *juolludit* 'grant'. Here, error detection would benefit from a dependency analysis of the sentence based on valency tags. If *miljovnnaid* 'million (Acc. Pl.)' can be associated with the verb *juolludit* 'grant' by means of dependency annotation, the local error detection rule can discard it as a potential target.

(77) a. *de ravggai dat soaldd[á]t badjel olbmá **juolgeláhpi** ...
then fell this soldier over man.GEN leg.NOM ...
'then the soldier tripped over the man's foot ...'

b. *muhto dát orro mannamen badjel **min ipmárdusrájit**.
but this seems go over our understanding.border.NOM.PL
'but this seems to go above our levels of understanding.'

c. Eiseválddit leat juolludan **miljovnnaid** rájáid rastá doaibmi
authorities have granted million.ACC.PL border.GEN.PL across functional
oktasaš prošeavttaide, ...
common project.ILL.PL, ...
'the authorities have granted millions to common projects across the borders, that work well ...'

A small amount of unsuccessful error detection is due to non-word and real word errors in the adposition's dependent. In ex. (78-a), *ovddageažit* is a non-word and should be *ovdageaži* 'front (Gen.)'. The case error is not recognized. In ex. (78-b), *gahppalat* is a real word error, i.e. a form of the verb *gahppat* 'jump', and should be *gáhppálaga* '(Gen.)'.

(78) a. *...ja fuomáša ahte Petter lea deadd[i]lan su biilla **ovddageažit**
...and realizes that Petter has pressed him car.GEN ?
ala.
onto
'...and realizes that Petter has pressed him up against the front of the car.'

b. *...dušše okta olmmái rievddai skiippa **gahppalat** nalde
...only one person drifted ship jump.PRS.1PL;PRS.2SG on
muhtun sullui.
some island.ILL
'...only one person drifted on top of a fragment of a ship towards some island.'

5.3.2.3 Conclusion regarding local case error detection

Local case error detection rules are placed before global error detection based on the assumption that local case error detection rules mainly operate with local contexts. While local case error detection rules were originally placed before partial dependency annotation and semantic role annotation in *grammarchecker.cg3*, they are now placed after them. However, error detection rules only make use of partial depenency annotation specific to adpositional phrases. Local case error detection in the adpositional phrases of the five adpositions *(n)ala, (n)alde, badjel, bokte* and *rastá* with *grammarchecker.cg3*[27] gives good results. While precision is 98.88%, recall is 80.97% and accuracy is 90.13%. The qualitative annotation showed that unsuccessful error detection is mostly due to disambiguation errors (67.3%). The key to successful error detection is a correct disambiguation of the

[27]version r157681 (Accessed 2017-09-29)

respective adposition by means of an analysis of the context. This can mostly be done locally by means of semantic tags as most adpositions have preferences as to which semantic prototype category they prefer as their dependent. Since adpositions belong to a closed category and their number is restricted, it is realistic to make idiosyncratic disambiguation rules for each one of them and achieve similar results. If there are several potential dependents of the potential adposition preceding or following it, the semantic prototype category of the noun does not always provide sufficient information, in particular if both are of the same prototype category.

While case error detection in adpositional phrases is mostly considered local error detection, it can require a global analysis of the sentence, as the potential dependent of the adposition can also be an argument of, for example, a verbal governor. The qualitative evaluation showed that a global analysis of the sentence including a dependency annotation of verbal governors and their arguments can help to exclude certain nouns as potential dependents of an adposition. The qualitative evaluation also revealed the importance of identifying complex noun phrases in adpositional phrases by means of dependency relations.

5.3.3 Evaluation of valency error detection

In this section, I will evaluate the valency error rules for the valency errors of four of the six rection verbs that were analyzed in Section 5.2.3.1, in *grammarchecker.cg3*.[28] These are the illative rection verbs *liikot* 'like', *luohttit* 'trust', and the locative rection verbs *beroštit* 'care' and *ballat* 'fear'. The error detection grammar includes 25 valency error detection rules, cf. Table 5.21 in Section 5.2.3.5.3. The rules target (pro)nominal, verbal, adjectival and adverbial forms that have governors with a number of different valency tags cf. Table 5.27.

Valency tag	**Type**
<TH-ahte>	finite subclauses introduced by *ahte* 'that'
<TH-Ill-Any>	illative arguments
<TH-Loc-Any>	locative arguments
<TH-AktioLoc>	non-finite actio locative arguments
<TH-Inf>	infinitival arguments
<TH-go>	subclauses introduced by *go* 'that, when'
<TH-Acc-Any><TH-Inf>	non-finite clauses with accusative subject and infinitive verb form
<TH-Acc-Any><TH-AktioEss>	non-finite clauses with accusative subject and actio essive (progressive) verb form

Table 5.27: An evaluation of valency error detection in *grammarchecker.cg3*: relevant valencies

[28] version r118631 (Accessed 2015-08-13)

	***liikot*-corpus**	***luohttit*-corpus**	***beroštit*-corpus**	***ballat*-corpus**	**Average**
True positives	78	121	33	400	158.0
False positives	17	19	18	124	44.5
True negatives	1,614	885	1,404	2,228	1,532.8
False negatives	36	17	50	63	41.5
Precision	82.21%	86.43%	64.71%	76.34%	77.42%
Recall	68.42%	87.68%	39.76%	86.39%	70.56%
Accuracy	96.96%	96.55%	95.48%	93.35%	95.59%

Table 5.28: A quantitative evaluation of valency error detection in *grammarchecker.cg3* version r118631

While the rules target any arguments of governors with the valency tags in question, I chose to evaluate only instances of these four verbal governors and their arguments. The reason for that is the cost and complexity of the error analysis itself. This study not only included a descriptive valency analysis, but also required a valency error definition for each governor in question, which had to be done before the error detection evaluation to maintain consistency and impartiality. For evaluation, I used a corpus different from the one used for rule development, containing verbs with the valency tags that potentially trigger those rules. The corpus for evaluation therefore consists of sentences extracted from *SIKOR* including at least one instance of the rection verbs in question. Half of the sentences were used for rule development, the other half were used for evaluation. Due to its small size compared to the other corpora, I used the complete corpus for evaluating *luohttit* 'trust'. There are four test corpora that altogether consist of 104,703 tokens or 7,291 sentences: 24,493 tokens or 1,876 sentences for *liikot* 'like' (*liikot*-corpus), 18,117 tokens or 1,150 sentences for *luohttit* 'trust' (*luohttit*-corpus), 22,770 tokens or 1,498 sentences for *beroštit* 'care' (*beroštit*-corpus), and 39,323 tokens or 2,767 sentences for *ballat* 'ballat' (*ballat*-corpus).

5.3.3.1 Quantitative evaluation

Valency error detection is considered to be successful (true positive) if the verb's incorrectly realized valency is marked in the sentence. Valency error diagnosis is considered to be successful if the error message consists of an adequate analysis of the error. Table 5.28 shows the distribution of true and false positives/negatives, and the results for precision, recall and accuracy for each of the evaluated corpora. This includes a testing of all error detection rules that find valency errors. Due to their valency tags, the results for *liikot-corpus* and *luohttit-corpus* include valency rules for illative valencies that do not hit in *ballat-corpus* and *beroštit-corpus*, while it is the opposite for valency rules for locative valencies.

	***liikot*-corpus**	***luohttit*-corpus**	***beroštit*-corpus**	***ballat*-corpus**	**Average**
True positives	64	110	27	371	143.0
False negatives	36	17	50	63	41.5
False positives	32	28	24	160	61.0
True negatives	1,614	885	1,404	2,228	1,532.8
Precision	66.67%	79.71%	52.94%	69.87%	67.30%
Recall	64.00%	86.61%	35.06%	85.48%	67.79%
Accuracy	96.10%	95.47%	95.09%	92.10%	94.69%

Table 5.29: Evaluation of error diagnosis in *grammarchecker.cg3* version r118631

Mean precision is 77%, mean recall is 70% and mean accuracy is 96%. Precision is highest for the verb *liikot* 'like', followed by *luohttit* 'trust' and *ballat* 'fear'. It is significantly lower for *beroštit* 'care' due to a change in valency assessment. In the grammar checker version evaluated here, *valency.cg3* annotates the tag *<TH-FS-Qpron>* (i.e. a finite subclause introduced by a question pronoun) to *beroštit* 'care'. However, this valency is not considered grammatical by *H*. Consequently, there is a drop in precision in the evaluation of *beroštit* 'care'. This case will be discussed further in the qualitative evaluation. Recall is 20% lower than precision, both for *liikot* 'like' and *beroštit* 'care', while it is the same as precision for *luohttit* 'trust' and even slightly higher for *ballat* 'fear'. Accuracy, however, is above 90% in all cases, and as high as 97% for *liikot* 'like'. Overall, the results are better for the verbs with illative valency than the verbs with locative valency.

The results for error diagnosis are presented in Table 5.29. True positives include only rule hits that also correctly diagnose the error. False positives, on the other hand, are those that are not recognized as valency errors, or those that are recognized as errors, but incorrectly diagnose the error. Precision, recall and accuracy for error diagnosis are therefore naturally lower for error diagnosis than for error detection. Precision is highest for *luohttit* 'trust' (79%) and lowest for *beroštit* 'care' (53%) due to the previously mentioned reasons. Precision is higher than recall for *liikot* 'like' and *beroštit* 'care', but lower for *luohttit* 'trust' and *ballat* 'fear'.

5.3.3.2 Qualitative evaluation

The qualitative evaluation analyzes the causes of unsuccessful error detection and diagnosis, pointing out weaknesses and potential for improvement. Tables 5.30 and 5.31 illustrate different causes of unsuccessful error detection and diagnosis. These can be non-word errors, real word errors, compound errors, punctuation errors, disambiguation errors, dependency errors, valency annotation errors, and error rule shortcomings. The first five causes are errors in the sentence structure. The other errors are caused by shortcomings of a module in *GoDivvun*. Disambiguation and valency annotation errors

are shortcomings in *disambiguator.cg3* and *valency.cg3*. Dependency and error detection errors, on the other hand, are shortcomings of *grammarchecker.cg3*.

Table 5.32 shows the quantitative side of the qualitative analysis. In unsuccessful error detection, 35.6% of the instances are due to missing or mismatched dependencies, and 21.9% are caused by error rule shortcomings. The latter are often due to long distances between the potential error and the governor. Disambiguation errors (19.3%) also make up a significant percentage. Diagnose errors are mostly due to error rule shortcomings (40%). The second largest group are spelling errors (23.3%). Again, disambiguation errors (11.7%) play an important role in error diagnosis as well. I will discuss the different causes in greater detail below.

5.3.3.2.1 Non-word errors

Non-words can affect valency error detection, and also alter the valency structure of a sentence. In the latter case they are not counted as false positives/negatives. In ex. (79), the non-word *dohpii* is a misspelled version of *dohppii* 'hold (Prt. 3Sg.)', cf. l.8 of the figure below, which is the governor of *eabbáriid* 'bucket (Acc. Pl.)'.

(79) Nieida liikui hirbmadit ja ***dohpii** eabbáriid ja vulggii
girl like.PRT.3SG extremely and ? bucket.ACC.PL and leave.PRT.3SG
'The girl it liked very much and grabbed the buckets and left'

However, because of the spelling error resulting in a non-word, it is not identified and the accusative *eabbáriid* 'bucket (Acc. Pl.)' is associated with the illative rection verb *liikui* 'like (Prt. 3Sg.)', cf. ll.10–11. The error detection rule therefore adds the error tag *&msyn-valency-ill-acc* to *eabbáriid* 'bucket (Acc. Pl.)'. This is counted as a true positive as the visible potential argument of *liikot* 'like' does not have the correct form.

```
1  "<liikui>"
2          "liikot" <mv> V <TH-Ill-Any> <TH-0> IV Ind Prt Sg3 @+FMAINV #2->2
3  "<hirbmadit>"
4          "hirbmadit" Adv @<ADVL #3->3
5  "<ja>"
6          "ja" CC @CNP #4->4
7  "<dohpii>"
8          "dohpii" ? #5->5
9  "<eabbáriid>"
10         "eappir" Sem/Ctain N Pl Acc @<OBJ &msyn-valency-ill-acc #6->6
11         "eappir" Sem/Ctain N Pl Gen &msyn-valency-ill-acc #6->6
12 "<ja>"
13         "ja" CC @CVP #7->7
```

In ex. (80), the locative argument of *balle* 'fear (Prs. 1Du.)', *skuvlaolbmain*, is a non-word, cf. l.4 of the figure below. This leads to a missing locative argument in the sentence.

CAUSE	Example	Correction
FALSE ERROR DETECTION		
non-word errors	ballet ođđa EO- njuolggadusat ja **gáibadusat** sáhttet dagahit 'fear new EU- rules and **requirements** could cause'	*gáibadusat* should be *gáibádusat* (Nom.), which is a potential target for the error detection rule suggesting accusative case
real word errors	In liiko jaska **čohkkat** beare guhká 'I do not like **to sit** quietly too long'	*čohkkat* 'mountain top (Nom. Sg. Px2Sg.)' should be *čohkkát* 'sit (Inf.)' so that it can be recognized as the infinitive argument of *liikot* 'like'
compound errors	ballá son ráđđehusa **guolástan politihkka** 's/he is afraid that **the government's fishing policies**'	*guolástan politihkka* should be *guolástanpolitihkka* so that it can be recognized as the accusative argument of *ballat* 'fear', which should be in locative case
punctuation/ formatting errors	eai dárbbaš ballat čázi // balduid olles leavttuin boahtit 'they do not need to fear that the water / ice flakes come down very fast'	/ should be / so that dependency rules can match *ballat* 'fear' with its accusative and infinitive arguments
disambiguation errors	ballá Nordlys jođiheaddji ... vuojahat **šaddat** (Inf.) orrut goarusin 'the Nordlys leader fears the vehicle **will** stay empty'	*šaddat* 'become' should be disambiguated as a first person plural instead of an infinitive so that a rule can recognize it as a finite clause argument of *ballat* 'fear'
dependency error	**Murmánskkas** gal balle *láhppot.* '**In Murmansk** they really feared getting lost.'	*Murmánskkas* 'Murmansk (Loc.)' is marked as a dependent of *ballat* 'fear', which is why the infinitive *láhppot* 'lose' is not recognized as a form that should have an actio locative form
valency tag errors	**beroštišgoahtit** mo gádjut sámi báikkiid árbevirolaš ealáhusaid. '**start to care** how we can save traditional industries in Sámi areas'	*beroštit* 'care' receives the valency *<TH-FS-Qst>*, which is why the subclause introduced by *mo* 'how' is not marked as an error
error rule shortcomings	Dás beroštedje politihkkárat uhccán **duorastaga** gielddastivrra čoahkkimis. 'The politicians cared little about that in the municipality meeting **on Thursday**.'	*duorastaga* 'Thursday (Acc.)' is marked as an argument of *beroštit* 'care' that should be in locative case because the correct argument *dás* 'it (Loc.)' is not recognized

Table 5.30: A qualitative evaluation: causes of unsuccessful valency error detection

CAUSE	Example	Correction
FALSE DIAGNOSIS		
non-word errors	ballá rievssatbivddu ráddjema **dramahtalažžan** 'fears ptarmigan hunting restrictions **dramatically**'	*dramahtalažžan* should be *dramáhtalažžan* 'dramatically (Ess.)', and is not recognized as an argument of *ballat* 'fear'; instead of suggesting the addition of the infinitive *leat* 'be', the accusative *ráddjema* 'restriction (Acc.)' is corrected to locative case
real word errors	Dál liikot iežat **rápmot** 'Now you like **to boast** yourself'	instead of diagnosing the real word error *rápmot* 'boast (Prt. 2Sg.)' (it should be *rábmot* 'boast (Inf.)'), a valency error detection rule assumes it is an erroneous finite subclause argument of *liikot* 'like'
compound errors	in sáhte beroštit eará go **ealáhus beroštumiin** 'I can not care about anything else but **the worries of the industries**'	instead of diagnosing a compound error in *ealáhus beroštumiin* (it should be *ealáhusberoštumiin*), *eará* 'other (Acc.)' is marked as the erroneous accusative argument of *beroštit* 'care', which should be in locative case
formatting errors	Ii galgga ballat konf[l]ivttain **Galgá** duostat buktit ovdan 'One should not be afraid of conflicts **One should** dare to present'	*Galgá* 'shall (Prs. 3Sg.)' should be preceded by punctuation marking the end of the previous sentences. Since this error is not marked, a valency error rule falsely assumes it is a subclause argument of *ballat* 'fear' lacking a subordinating conjunction (*ahte* 'that')
disambiguation errors	Ballen **maid** láhppit dan go duogábealde leat 'I was **also** afraid to lose'	*maid* is disambiguated as an accusative pronoun instead of an adverb ('also') and therefore marked as the erroneous accusative argument of *ballat* 'fear'. Instead *láhppit* 'lose' should be recognized as the erroneous infinitive argument of *ballat* 'fear'
error rule shortcomings	Santa **Barbaras** liikuime buoremusat 'We liked it best in Santa **Barbara (Loc.)**'	Instead of adding *leat* 'be' to the locative argument of *liikot* 'like', the error rule suggests replacing it with an illative form

Table 5.31: A qualitative evaluation: causes of false diagnosis

CAUSE	liikot	luohttit	ballat	beroštit	%
FALSE ERROR DETECTION					
Non-words	2	-	4	2	2.3%
Real word errors	1	2	-	1	1.1%
Compound errors	-	-	1	-	0.3%
Punctuation errors	-	2	-	-	0.6%
Formatting errors	-	-	2	-	0.6%
Disambiguation errors	16	8	30	13	19.3%
Dependency errors	16	9	94	5	35.6%
Valency tag errors	-	-	3	26	8.3%
Error rule errors	25	15	51	20	21.9%
FALSE DIAGNOSIS					
Non-word errors	3	3	7	1	23.3%
Real word errors	3	2	2	-	11.7%
Compound errors	-	1	2	3	10.0%
Formatting errors	-	-	2	-	3.3%
Disambiguation errors	4		2	1	11.7%
Dependency errors	-	-	1	-	1.7%
Error rule shortcomings	6	3	15	-	40.0%

Table 5.32: The causes of unsuccessful valency error detection and diagnosis in numbers

(80) Ja dalle balle mánat ***skuvlaolbmain**, eai duostan ...
and then feared children school.people.LOC.PL, not dare ...
'And then the children feared the school staff, they did not dare ...'

Therefore, the error detection rule associates the subsequent finite subclause with *balle* 'fear (Prs. 1Du.)', annotates the error &*msyn-valency-ahte-not-fs* to the finite verb *mánat* 'go (Prs. 2Sg.)', cf. l.8, and inserts a cohort for the subjunction *ahte* 'that', cf. ll.5–6. However, since the remaining elements of the sentence visible to the grammar checker leave an erroneous valency structure, this is considered a true positive.

```
1  "<balle>"
2          "ballat" <mv> V TV Ind Prs Du1 @+FMAINV #3->3
3  "<ahte>"
4          "ahte" CS &SUGGEST #4->4 ADDCOHORT-BEFORE:7764:wrong-valency-ahte-not-fs
5  "<mánat>"
6          "mánnat" <mv> V TV Ind Prs Sg2 @+FMAINV &msyn-valency-ahte-not-fs #5->5
7  "<skuvlaolbmain>"
8          "skuvlaolbmain" ? #6->6
```

In ex. (81), the failure to annotate the subclause argument of *ballat* 'fear' is considered a false negative. The finite verb *sáhttet* 'can (Prs. 3Pl.)' of the finite subclause argument of *ballet* 'fear (Prs. 3Pl.)' should receive the valency error tag &*msyn-ahte-not-fs*, marking the missing subjunction *ahte* 'that'.

(81) *ballet ođđa EO- njuolggadusat ja **gáibadusat** sáhttet
fear new EU- rule.NOM.PL and requirement.NOM.PL can.PRS.3PL
dagahit ...
cause.INF ...
'fear that new EU- rules and requirements could cause ...'

The non-word error *gáibadusat* 'requirement (Nom. Pl.)', which is part of the coordinated subject of the finite clause argument of *ballet* 'fear (Prs. 3Pl.)', receives the correct analysis from the descriptive morphological analyzer. It is analyzed as a nominative plural with an orthographical error (*Err/Orth*), cf. l.17. However, the error is not annotated.

```
1  "<ballet>"
2          "ballat" <mv> V <EX-Nom-Ani> <heaggabealláI> <jámas> <RS-dihte-Any>
3          <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
4          <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
5          <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prs Pl3 @+FMAINV #3->3
6  "<ođđa>"
7          "ođas" A Attr @>N #4->4
8  "<EO->"
9          "EO" N ACR RCmpnd @>N #5->5
10         "EO" N ACR Err/Orth RCmpnd @>N #5->5
11 "<njuolggadusat>"
12         "njuolggadus" N Sem/Rule Pl Nom #6->6
13 "<ja>"
14         "ja" CC @CVP #7->7
15 ;        "ja" CC @CNP
16 "<gáibadusat>"
17         "gáibádus" Err/Orth N Sem/Dummytag Pl Nom @SUBJ> #8->8
18 "<sáhttet>"
19         "sáhttit" <aux> V IV Ind Prs Pl3 @+FAUXV #9->9
```

5.3.3.2.2 Real word errors

Real word errors can also alternate the argument structure, e.g. by confusing a relevant element, i.e. a governor or an argument, for another part of speech, another verb with different valency frames, etc. Like non-words, some (but not all) real word errors are recognized by the descriptive morphological analyzer and tagged by means of the error tag *Err/Orth*, cf. l.9. of the figure below. In newer versions of *GoDivvun*, error tags include information about the error type, e.g. accent errors are annotated with the tag *Err/Orth-a-á*.

```
1  "<In>"
2          "ii" <aux> V IV Neg Ind Sg1 @+FAUXV #2->2
3  "<liiko>"
4          "liikot" <mv> V <EX-Nom-Ani> <TH-Inf> <TH-jus> <TH-go> <TH-ahte> <TH-Ill-Any>
5          <TH-0> IV Ind Prs ConNeg @-FMAINV #3->3
6  "<jaska>"
7          "jaska" Adv @<ADVL #4->4
8  "<čohkkat>"
9          "čohkkát" <mv> V IV §TH Inf Err/Orth @-FMAINV #5->3
```

```
;        "čohkka" N Sem/Plc-elevate Sg Nom PxSg2
"<beare>"
        "beare" Adv @<ADVL #6->6
"<guhká>"
        "guhká" Adv Sem/Time @<ADVL #7->7
```

In ex. (82), *čohkkat* 'mountain top (Nom. Sg. Px2Sg.)' is a real word error for the infinitive verb *čohkkát* 'sit'. As the real word error receives a morphological analysis, cf. l.9 in the figure below, it can be associated with its governor *liikot* 'like' and receive a THEME-label instead of being identified as the homonymous noun in nominative case, cf. l.10. Morphological error annotation leads here to the successful recognition of a true negative.

(82) In liiko jaska ***čohkkat*** beare guhká ...
not like quietly mountain.top.NOM.PXSG2 too long ...
'I do not like to sit quietly too long ...'

In ex. (83), the real word error *gáhtet* 'regret (Prs. 3Pl.)', which is confused with *gáhttet* 'take care (Inf.)', leads to a valency error in the argument of *liikot* 'like'. As the argument of *liikot* 'like' is a finite subclause because of the real word error, it receives the error tag *&msyn-inf-not-fs*, suggesting the addition of the subordinating conjunction *ahte* 'that', cf. ll.9–10 in the figure below. This is considered a true positive in error detection, and a false positive in error diagnosis.

(83) Son liiko ***gáhtet*** iežas.
s/he likes regret.PRS.3PL herself/himself
'S/he likes to take care of herself/himself.

```
1  "<Son>"
2          "son" Pron Sem/Hum Pers Sg3 Nom @SUBJ> #1->1
3  "<liiko>"
4          "liikot" <mv> V <EX-Nom-Ani> <TH-Inf> <TH-jus> <TH-go> <TH-ahte> <TH-Ill-Any>
5          <TH-0> IV Ind Prs Sg3 @+FMAINV #2->2
6  "<ahte>"
7          "ahte" CS &SUGGEST #3->3 ADDCOHORT-BEFORE:8203:wrong-valency-ahte-not-fs
8  "<gáhtet>"
9          "gáhtat" <mv> V <TH-Acc-Any> TV Ind Prs Pl3 @+FMAINV
10         &msyn-valency-inf-not-fs #4->4
11         "gáhtat" <mv> V <TH-Acc-Any> TV Ind Prs Pl3 @+FMAINV
12         &msyn-valency-ahte-not-fs #4->4
13 "<iežas>"
14         "ieš" §TH Pron Refl Acc PxSg3 @<OBJ #5->4
```

5.3.3.2.3 Compound errors

In ex. (84-a) the compound error *guolástan politihkka*, which should be *guolástanpolitihkka* 'fishing policies', is responsible for a false valency error diagnosis.

(84) a. *ballá son ráđđehusa **guolástan politihkka**, gos eriid sáhttá gávppašit, bágge ...
fears the government's fishing policies, where quotas can trade, forces ...
's/he fears the government's fishing policies, where one can trade quotas, forces ...'

b. ballá son ráđđehusa **guolástanpolitihka**, gos eriid sáhttá gávppašit, bágge ... CORR

While the valency error is a finite clause argument of *ballat* 'fear', which should be introduced by *ahte* 'that', the genitive/accusative modifier *ráđđehusa* 'government (Gen.)' receives the tag *&msyn-valency-loc-acc*, cf. ll.9–10 in the figure below. Instead, the finite verb *bágge* 'forces' should receive the error tag *&msyn-ahte-not-fs*. The first part of the compound *guolástan* 'fish (PrfPrc.)' is analyzed as a past participle verbal reading that typically functions as a barrier to many error detection rules, which is why *guolástan politihkka* cannot be identified as a compound subject of the finite verb *bágge* 'force (Prs. 3Sg.)'.

```
1  "<ballá>"
2          "ballat" <mv> V <EX-Nom-Ani> <heaggabealláі> <jámas> <RS-dihte-Any>
3          <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
4          <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
5          <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prs Sg3 @+FMAINV #2->2
6  "<son>"
7          "son" Pron Sem/Hum Pers Sg3 Nom @<SUBJ #3->3
8  "<ráđđehusa>"
9          "ráđđehus" N Sem/Org Sg Gen @-FSUBJ> &msyn-valency-loc-acc #4->4
10         "ráđđehus" N Sem/Org Err/Orth Sg Gen @-FSUBJ> &msyn-valency-loc-acc #4->4
11 "<guolástan>"
12         "guolástit" V IV PrfPrc @>N #5->5
13         "guolástit" V IV Actio Gen @>N #5->5
14         "guolástit" V IV Actio Nom @>N #5->5
15 "<politihkka>"
16         "politihkka" N Sem/Domain Sg Nom @<SPRED #6->6
17 "<,>"
18         "," CLB #7->7
19 ...
20 "<bágge>"
21         "bágget" <mv> V TV Ind Prs Sg3 @+FMAINV #13->13
```

5.3.3.2.4 Punctuation and formatting errors

Punctuation and formatting errors can also influence the performance of valency error detection, especially as error detection rules often depend on punctuation to identify clause and sentence boundaries. In ex. (85) the missing period or line break before *Galgá* 'shall (Prs. 3Sg.)' leads to the assumption that there is a subsequent subclause that is a potential argument of the verb *ballat* 'fear', which is why the error detection rule discovers a missing subordinating coordinator and annotates the error tag *&msyn-ahte-not-fs* to *Galgá* 'shall (Prs. 3Sg.)'.

(85) *Ii galgga ballat konf[l]ivttain **Galgá** duostat buktit ovdan ...
not should fear conflicts should dare bring forward ...
'One should not be afraid of conflicts One should dare to present ...'

5.3.3.2.5 Disambiguation errors

Disambiguation errors make up a high percentage of the causes of both unsuccessful error detection and error diagnosis. In ex. (86), the form *ballu* 'fear' is a nominative singular noun, cf. l.6 in the figure below. However, it is wrongly disambiguated as an imperative form of the verb *ballat* 'fear', ll.2–5. While the noun *ballu* 'fear' can have an infinitive argument, it is corrected to a non-finite actio locative form if it is considered a THEME-argument of the verb *ballat* 'fear'. Since *disambiguator.cg3* removes the correct nominal reading of *ballu*, *hávváduhttit* 'hurt' is considered the infinitive THEME-argument of the governor *ballat* 'fear' and receives the error label *&msyn-aktioloc-inf*, cf. ll.8–10, which is a false positive.

(86) **Ballu** hávváduhttit ja gielistit ii leat dárbbašlaš ...
fear.IMPRT.DU1;SG.NOM hurt.INF and lie.INF not be necessary ...
'The fear of hurting and lying is not necessary ...'

```
1  "<Ballu>"
2          "ballat" <mv> V <EX-Nom-Ani> <heaggabeallái> <jámas> <RS-dihte-Any>
3          <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
4          <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
5          <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Imprt Du1 @+FMAINV #1->1
6  ;        "ballu" N <TH-Ill-Any> <TH-Loc-Any> <TH-ahte> Sem/Perc-emo Sg Nom @SUBJ>
7  "<hávváduhttit>"
8          "hávváduhttit" <mv> V TV Inf @-FMAINV &msyn-valency-aktioloc-inf #2->2
9          "hávváduhttit" <mv> V TV @-FMAINV Actio Loc &SUGGEST #2->2
10 "<ja>"
11         "ja" CC @CVP #3->3
12 ;        "ja" CC @CNP
13 "<gielistit>"
14         "gielistit" <vdic> <mv> V <TH-ahte> TV Inf @-FMAINV #4->4
15 "<ahte>"
16         "ahte" CS &SUGGEST #5->5 ADDCOHORT-BEFORE:8203:wrong-valency-ahte-not-fs
17 "<ii>"
18         "ii" <aux> V IV Neg Ind Sg3 @+FAUXV &msyn-valency-ahte-not-fs #6->6
```

In ex. (87), *máid/maid* 'which, also', which can be an adverb or a pronoun in accusative case, is erroneously disambiguated as an accusative relative pronoun introducing a relative clause. The form can be difficult to disambiguate especially in an error context. As an accusative form, it is considered an argument of *ballat* 'fear' and receives the error tag *&msyn-valency-loc-acc*. Relative pronouns are typically clause barriers, which is why the subsequent infinitive *láhppit* 'lose' is not recognized as an argument of *ballat* 'fear', which should be a non-finite actio locative form. Consequently, the disambiguation error leads to a diagnosis error.

(87) *Ballen **maid** láhppit dan go ...
fear.PRT.3SG also lose.INF it.ACC because ...
'I was also afraid to lose it because ...'

```
1  "<Ballen>"
2          "ballat" <mv> V <EX-Nom-Ani> <heaggabealláі> <jámas> <RS-dihte-Any>
3          <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
4          <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
5          <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prt Sg1 @+FMAINV #2->2
6  "<maid>"
7          "mii" Pron Indef Sg Acc @<OBJ &msyn-valency-loc-acc #3->3
8          "mii" Pron Rel Sg Acc @OBJ> &msyn-valency-loc-acc #3->3
9          "mii" Pron Indef Sg @<OBJ Loc &SUGGEST #3->3
10         "mii" Pron Rel Sg @OBJ> Loc &SUGGEST #3->3
11 ;        "maid" Interj
12 ;        "maid" Adv
13 "<láhppit>"
14         "láhppit" <mv> V TV Ind Prs Pl1 @FS-<ADVL #4->4
15 ;        "láhppit" V TV Inf @-FMAINV
16 "<dan>"
17         "dat" Pron Dem Sg Acc @<OBJ
```

5.3.3.2.6 Valency annotation errors

Error detection rules are applied in the context of certain valency tags, which is why a correct tag sequence is essential for successful error detection. If a valency tag is annotated or erroneously not annotated to a particular governor this can lead to a non-application or over-application of an error rule. In the case of *beroštit* 'care' the evaluation above showed that the redundant valency tag (*<TH-FS-Qst>*) leads to massive non-application of an error detection rule where it should hit (i.e. 26 instances total), because the construction is assumed to be correct.

Ex. (88) is one of the cases in which the erroneous annotation of *<TH-FS-Qpron>* to *beroštit* 'care' results in a false negative in valency error detection. The subclause argument of *beroštit* 'care' introduced by *mo* 'how' should be preceded by *das* 'it (Loc.)'. However, the *<TH-FS-Qpron>* valency tag, cf. l.3 in the figure below, explicitly states that a subclause introduced by a question pronoun is a possible valency frame. This error in valency annotation results in a non-annotation of the valency error tag *&msyn-add-das*.

In other words, it is a false negative.

(88) *...ja beroštišgoahtit **mo** gádjut sámi báikkiid árbevirolaš ealáhusaid.
...and start.to.care how save Sámi places traditional industries
'...and start to care about how we can save traditional Sámi places traditional industries.'

```
"<beroštišgoahtit>"
        "beroštit" V* IV Der/goahti <mv> V <AG-Nom-Any> <TH-AktioLoc> <TH-Inf>
        <TH-Loc-Any> <TH-FS-Qpron> <TH-ahte> <TH-0> Ind Prs Pl1 @+FMAINV #9->9
"<mo>"
        "mo" Adv @ADVL> #10->10
"<gádjut>"
        "gádjut" <mv> V <TH-Acc-Any><RS-Loc-Any> <TH-Acc-Any>
        TV §TH Inf @-FMAINV #11->9
"<sámi>"
        "sápmi" Err/Orth N Sem/Hum_Lang Sg Gen @>N #12->12
        "sápmi" N Sem/Hum_Lang Sg Gen @>N #12->12
"<báikkiid>"
        "báiki" N Sem/Plc Err/Orth Pl Gen @>A #13->13
        "báiki" N Sem/Plc Pl Gen @>A #13->13
;        "báiki" N Sem/Plc Err/Orth Pl Acc
;        "báiki" N Sem/Plc Pl Acc
```

In the accusative + infinitive construction in ex. (89), *meahcásteami* 'hunting (Acc.)' is marked as an object (*@<OBJ*) of *gáržžiduvvot* 'restrict (Pass.)' instead of its subject because *gáržžiduvvot* has the valency tag *<PA-Acc-Any>* (i.e. a PATIENT in accusative case), cf. l.10 of the figure below. However, the valency should only be annotated to the active form of *gáržžidit* 'restrict'. The valency rule annotating *<PA-Acc-Any>* needs to include a negative constraint for passive derivations. Because of this redundant valency, *meahcásteami* 'hunting (Acc.)' is annotated as a PATIENT and dependent of the verb *gáržžiduvvot* 'restrict (Pass.)'. However, together with the infinitive it should be annotated as an argument of *ballat* 'fear'. As the correct argument is not recognized, the infinitive receives an error tag (*&msyn-valency-aktioloc-inf*), which is a false positive.

(89) Ballá meahcásteami **gáržžiduvvot**
fears hunting.ACC restrict.PASS.INF
'S/he fears that hunting will be restricted'

```
"<Ballá>"
        "ballat" <mv> V <EX-Nom-Ani> <heaggabealláis> <jámas> <RS-dihte-Any>
        <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
        <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
        <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prs Sg3 @+FMAINV #1->1
"<meahcásteami>"
        "meahcásteapmi" §PA N Sem/Act Sg Acc @<OBJ #2->3
        "meahcásteapmi" §PA N Sem/Act Sg Gen @ADVL> #2->3
"<gáržžiduvvot>"
        "gáržžidit" V* TV* Der/PassL <mv> <PA-Acc-*Ani><BE-Ill-Ani> <PA-Acc-Any>
```

```
11          V IV Inf @-FMAINV &msyn-valency-aktioloc-inf #3->3
12          "gáržžidit" V* TV* Der/PassL <mv> <PA-Acc-*Ani><BE-Ill-Ani> <PA-Acc-Any>
13          V IV @-FMAINV Actio Loc &SUGGEST #3->3
```

5.3.3.2.7 Dependency annotation errors

Unassociated or falsely associated arguments can also lead to unsuccessful error detection and error diagnosis. If correct arguments are not associated with their governors this is typically due to disambiguation errors, long distances between the governor and its argument, and/or complex clauses between the governor and its argument. It can also be due to the order in which the dependency rules are applied. Erroneous dependency annotations can either be arguments that are not associated with their correct governors or governors that are associated with incorrect arguments. False associations can happen for the same reasons as missing associations. If correct arguments are not associated, the valency error detection searches for an error in the sentence. If governors are associated with parts of the sentences that are not their arguments, valency errors in their actual arguments may not be found.

In ex. (90), there is an unrecognized valency error, i.e. the infinitive *láhppot* should be *láhppomis* 'lose (Actio. Loc.)'. The locative sentence adverbial *Murmánskkas* 'Murmansk (Loc.)' is wrongly associated with the verbal governor *balle* 'fear (Prt. 3Pl.)' and receives a THEME-label, cf. l.2 in the figure below. However, *láhppot* 'get lost' is its actual THEME and should have an actio locative (*láhppomis*) rather than an infinitive form. The dependency error leads to a false negative in error detection.

(90) *__Murmánskkas__ gal balle láhppot.
Murmansk.LOC really fear get.lost.INF
'In Murmansk they really feared getting lost.'

```
1   "<Murmánskkas>"
2           "Murmánska" §TH N Prop Sem/Plc Sg Loc @ADVL> #1->3
3   "<gal>"
4           "gal" Adv @ADVL> #2->2
5   "<balle>"
6           "ballat" <mv> V <EX-Nom-Ani> <heaggabeallái> <jámas> <RS-dihte-Any>
7           <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
8           <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
9           <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prt Pl3 @+FMAINV#3->3
10  "<láhppot>"
11          "láhppot" V IV Inf
```

Ex. (91), on the other hand, is an example of a false positive. The infinitive *áigut* 'want' (*#8->2*) and the accusative *NSR* (*#7->2*) are falsely associated with the governor *adden* 'give (Prt. 1Sg.)', instead of *ballat* 'fear', cf. ll.16–18 in the figure below. Like *ballat* 'fear', the verb *addit* 'give' has an accusative + infinitive valency (*<TH-Acc-Any><TH-Inf>*). The form *adden* 'give (Prt. 1Sg.)' is a real word error for *ádden* 'understand (Prs.

1Sg.)'. However, the valency is not corrected by the error correction rule. Therefore, the accusative + infinitive argument of *ballat* 'fear' is not associated with *ballat*, and the error tag *&msyn-valency-ahte-inf* is added to the infinitive *áigut* 'want'.

(91) Adden bures jus olbmot ballet NSR **áigut** ásahit sámi stáhta
give well if people fear NSR want.INF establish.INF Sámi state
'I understand well if people fear NSR wants to establish a Sámi state'

```
"<Adden>"
        "addit" <mv> <TH-Acc-Any><TH-Inf> V TV Ind Prt Sg1 @+FMAINV &real-áddet #2->2
        "áddet" <mv> <TH-Acc-Any><TH-Inf> V TV @+FMAINV Inf &SUGGEST #2->2
"<bures>"
        "bures" Adv @<ADVL #3->3
"<jus>"
        "jus" CS @CVP #4->4
"<olbmot>"
        "olmmoš" N Sem/Hum Pl Nom @SUBJ> #5->5
"<ballet>"
        "ballat" <mv> V <EX-Nom-Ani> <heaggabealláі> <jámas> <RS-dihte-Any>
        <RS-Acc-Reason> <TH-AktioLoc> <TH-Acc-Any><TH-Inf> <TH-Acc-Any><TH-PrfPrc>
        <TH-Acc-Any><TH-AktioEss> <TH-Com-Impers> <TH-Acc-Impers> <TH-Loc-Any>
        <TH-jus> <TH-go> <TH-FS-Qst> <TH-ahte> <TH-0> TV Ind Prs Pl3 @FS-<ADVL #6->6
"<NSR>"
        "NSR" §TH N ACR Sg Acc #7->2
"<áigut>"
        "áigut" <mv> §TH V <Inf> TV Inf @-FMAINV &msyn-valency-ahte-inf #8->2
        "áigut" <mv> §TH V <Inf> TV @-FMAINV Pl3 &SUGGEST #8->2
"<ásahit>"
        "ásahit" <mv> V TV Inf @-FMAINV #9->9
"<sámi>"
        "sápmi" Err/Orth N Sem/Hum_Lang Sg Gen @>N #10->10
        "sápmi" N Sem/Hum_Lang Sg Gen @>N #10->10
"<stáhta>"
        "stáhta" §PR N Sem/Org Sg Acc @<OBJ #11->9
```

5.3.3.2.8 Error detection errors

Lastly, error detection rules themselves can have different types of shortcomings leading to unsuccessful error detection or error correction. These include missing positive or negative conditions, missing barriers, errors in the definition of the scope of a particular condition, etc.

In ex. (92), the error detection rule faces a distance problem, leading to a false negative in valency error detection. Since *guorbademiid* 'devastation (Acc. Pl.)', l.4 in the figure below, is separated from its governor *beroštit* 'care' by a relative clause (*maid jeagelbordin guorbada* 'that gathering lichen causes'), the case error (*&msyn-valency-loc-acc*) is not recognized and the error tag is missing. Relative pronouns are often used as barriers in error detection rules as they mark the beginning of a new clause, and it can be difficult to define the end of a relative clause.

(92) *Muhto daid **guorbademiid** maid jeagelbordin guorbada eai oro
but the devastation.ACC.PL that lichen.gathering devastates not seem
okta ge berošteam[e]n
anyone either care.ACTIO.ESS
'But no one seems to care about the devastation that gathering lichen causes'

```
"<daid>"
        "dat" Pron Dem Pl Com Attr @>N #2->2
;        "dat" Pron Dem Pl Acc
"<guorbademiid>"
        "guorbadit" V* TV* Der/NomAct N Pl Acc @<OBJ #3->3
"<maid>"
        "mii" Pron Rel Pl Gen @>N #4->4
;        "mii" Pron Rel Pl Acc @OBJ>
"<jeagelbordin>"
        "jeagelbordin" N Sem/Act Sg Nom @SUBJ> #5->5
;        "jeagelbordin" N Sem/Act Sg Gen
"<guorbada>"
        "guorbadit" <mv> V TV Ind Prs Sg3 @+FMAINV #6->6
"<eai>"
        "ii" <aux> V IV Neg Ind Pl3 @+FAUXV #7->7
"<oro>"
        "orrut" <mv> V IV Ind Prs ConNeg @-FMAINV #8->8
"<okta>"
        "okta" Num Sg Nom @<SUBJ #9->9
"<ge>"
        "ge" Pcle @PCLE #10->10
"<berošteamin>"
        "beroštit" V* IV* Der/NomAct N Sem/Act Sg Loc South Err/Orth @<ADVL #11->11
        "beroštit" <mv> V <AG-Nom-Any> <TH-AktioLoc> <TH-Inf> <TH-Loc-Any>
        <TH-FS-Qpron> <TH-ahte> <TH-0> IV Actio Ess Err/Orth @-FMAINV #11->11
```

The following example does not involve the valency of any of the four previously discussed governors, but illustrates another type of error detection rule problem. In ex. (93), the valency error tag &*msyn-valency-dasa-before-ahte* is added to *ahte* 'that' because the error rule did not take into account the idiomatic use of *ahte* 'that' in constructions such as *eambbo ahte eambbo* 'more and more'. In the latter case, *ahte* 'that' does not introduce a subclause, which is a potential subclause argument of *álgit* 'begin'. Here the idiomatic construction *eambbo ahte eambbo* 'more and more' is not perceived as such by the error rule and *dasa* 'it (Ill.)' is erroneously added before *ahte* 'that' because it is assumed to be a subordinating subclause.

(93) Don álggát eambbo **ahte** eambbo beroštit dakkár **áššiin**
you start more and more care.INF such thing.LOC.PL
'You start to care more and more about these kinds of things'

```
"<álggát>"
        "álgit" <aux> V <TH-Ill-Any> IV Ind Prs Sg2 @+FAUXV #2->2
"<eambbo>"
        "eambbo" Adv Comp <ctjHead> @ADVL> #3->3
```

```
"<dasa>"
        "dat" Pron Dem Sg Ill &SUGGEST #4->4
"<ahte>"
        "ahte" CC @CNP &msyn-valency-dasa-before-ahte #5->5
"<eambbo>"
        "eambbo" Adv Comp @ADVL> #6->6
"<beroštit>"
        "beroštit" <mv> V <AG-Nom-Any> <TH-AktioLoc> <TH-Inf> <TH-Loc-Any>
        <TH-FS-Qpron> <TH-ahte> <TH-0> IV Inf @-FMAINV #7->7
"<dakkár>"
        "dakkár" Pron Dem Attr @>N #8->8
"<áššiin>"
        "ášši" N G3 Sem/Semcon Sg Com @<ADVL #9->9
        "ášši" N G3 Sem/Semcon Err/Orth Sg Com @<ADVL #9->9
;        "ášši" N G3 Sem/Semcon Pl Loc
```

5.3.3.3 Conclusion regarding valency error detection

Valency error detection requires a detailed analysis of each governor's valency and an explicit definition of its grammatical and ungrammatical valencies. This work has been done from scratch, i.e. without a valency dictionary to rely on, which is why only six verbal governors were analyzed and evaluated in regards to the detection of errors in their valency structure.

The detailed analysis of six multi-valency verbal governors and the evaluation of valency error detection rules targeting four of these governors has shown how variable valency and errors related to the governor-argument relation are. The verb *ballat* 'fear', for example, is represented by 36 different valency frames in *SIKOR*. As valency errors are only implicitly tagged in the lexicon, the absence of a valency tag means that this valency will be counted as an error if targeted by a rule. Therefore a complete analysis of a verb's valency potential is required before it is available for error detection. However, listing ungrammatical valencies in *valency.cg3* and marking them as such can be a potential improvement for the process of valency error detection and make it more robust. Tagging morphological errors explicitly in the morphological analyzer has been shown to improve non-word and real word error analysis and provide a full sentential context for the valency error. Since explicit tagging of non-word errors and real word errors improves the analysis, it suggests that explicit valency error tagging can also improve the analysis. Even without a full valency analysis of a verb, one could assume certain types of valency errors.

For grammaticality decisions I followed the language norm in the cases where there is one. Otherwise, I followed the linguistic intuitions of the informants *H* and *N*. Some of their grammaticality decisions may be controversial, and due to high frequencies of certain constructions in *SIKOR*, some of the valency error detection rules evaluated in the

previous section are not in use in newer versions of *grammarchecker.cg3*,[29] and valencies described as ungrammatical in this chapter, e.g. both illative and locative valencies of *dolkat* 'get fed up', have been added to newer versions of *valency.cg3*.[30]

Valency error detection requires a global analysis of the sentential context and needs to take into account several linguistic layers. Apart from a morphological analysis, morpho-syntactic analysis and disambiguation, this also includes a morphological analysis of non-words and compounds, an adaptation of disambiguation rules for potential errors, a dependency analysis of relevant relations for the error, a semantic role analysis and lastly a set of error detection rules.

Valency tags and semantic prototype tags are the backbone of a global analysis and are used in all modules of *GoDivvun*, i.e. valency annotation, disambiguation, dependency analysis, semantic role analysis and error detection. A general principle is that the denser the linguistic analysis of the context of the error is, the easier it is to correctly identify the error. There are several ways of making a sentence analysis denser for grammar checking. Non-words can to some extent be enhanced with error tags and be included in the morphological analyzer. Error detection rules can also guess possible case forms of non-words with characteristic endings by means of regular expressions.

Disambiguation can be improved by including more linguistic information, i.e. valencies and semantic prototype tags, in specific disambiguation rules. In the previous sections, I described a four-step system to adapt the regular disambiguator to grammatically erroneous context. Performing real word error detection prior to valency error detection further enhances the sentential context with linguistic information, and can avoid false positives in valency error detection by correctly identifying confused arguments or governors and correcting them. In a next step, governor-argument relations in correct valency constellations are established by means of dependency annotation. Certain parts of the sentence can thereby be discarded as potential errors. Dependency annotation of arguments and governors relies heavily on valency tags and semantic prototype tags. As distance plays an important role in dependency annotation, rules work incrementally and test the immediate context first. They are followed by rules that test larger distances and more complex contexts if an immediate argument could not be found. Semantic roles are necessary for argument indexing and the distinction of different arguments of a governor in their possible morpho-syntactic realizations. The more semantic roles can be matched, the denser the context is and the fewer possible error candidates there are. Semantic role annotation also includes adjunct annotation. Semantic role annotation reduces the possible targets for error detection by more than half. Lastly, global error detection rules can refer to previously established relations, error annotations, and identifications of correct contexts in their context conditions.

[29] version r157816 (Accessed 2017-10-02)
[30] version r155649 (Accessed 2017-08-11)

Extensive valency descriptions are necessary for an identification of correct valency constellations. Identifying e.g. correct THEME-less constructions that include optional MANNER-arguments is crucial for a distinction from ill-formed THEMES. Therefore, I apply a wide definition of valency including many optional arguments of a governor, even those that are considered adjuncts in other descriptions, in the valency of a governor.

The evaluation of valency error detection is a small scale evaluation of the valency error detection with respect to four rection verbs. It shows that good results can be achieved based on a detailed valency analysis and government argument annotation. Mean precision is 77%, mean recall is 71% and mean accuracy is 96%. However, for a full coverage of the valency error detection rules, an analysis of correct and incorrect valencies for each governor is necessary. In the future, a full use of this type of annotation and error detection will depend on the existence of a syntactic norm in North Sámi.

5.4 Conclusion

In this chapter, I presented the structure of *GoDivvun*, which consists of various modules. The first module of the most recent version described here is a tokenizer/descriptive morphological analyzer, *tokeniser-gramcheck-gt-desc.pmhfst*, that is based on a lexicon containing lemmata, part of speech information, morphological tags, error tags and semantic prototype tags. The analyzer provides all possible homonymous analyses of a particular form. It also analyzes potential two-word compounds that have a lexicalized compound analysis as one word in the lexicon and adds an error tag to the combinations that are lexicalized. The morphological analysis is followed by an analysis of the valency annotation grammar *valency.cg3*, adding multiple valency tags to the respective governors, which provide the basis for dependency and semantic role analysis. A constraint grammar module, *mwe-dis.cg3*, can then undo the compound analysis based on basic undisambiguated morphological and valency context conditions. The subsequent module, *disambiguator.cg3*, performs morpho-syntactic annotation and disambiguation. Disambiguation is necessary as a grammatical error can only be found in a certain grammatical context. The error detection module *grammarchecker.cg3* performs local error detection followed by a dependency analysis and a semantic role annotation of governors and their arguments, which are largely based on valency tags and semantic prototype tags. This is followed by local case error detection and global error detection. Lastly, a normative analyzer, *generator-gt-norm.hfstol*, and a reformatter, *divvun-suggest*, generate the correct forms from tag combinations suggested by the error detection module.

In this chapter, I also analyzed and evaluated local and global grammatical errors that benefit from valency and semantic prototype information on different stages in the grammar checking process. For local error error detection, I analyzed six real word error confusion pairs and the idiosyncratic relations between the confused items. Additionally,

I analyzed systematic local case errors in the adpositional phrases of five adpositions, i.e. confusions of genitive/accusative case forms with any other case form. As representatives of global errors, I analyzed and evaluated the valency errors of four verbal governors, which can be simple case errors or involve arguments realized as non-finite clauses, finite subclauses, and adpositional phrases. While real word error detection requires valency information and semantic prototype information in the context conditions of the error detection rule, local case errors in adpositional phrases require this information predominantly in the disambiguation process for the governor of the respective case, e.g. adverb-adpositional disambiguation. Valency error detection, the most complex of all processes, on the other hand, requires valency information on all stages, i.e. in disambiguation, previous real word error detection, dependency annotation, semantic role annotation, and error detection itself. Valency error detection in particular requires valency information regarding a potential governor and morpho-syntactic and semantic information about a potential argument. While valency errors can also be errors in the derivational form of the governor, I mainly focused on the form of the argument.

The analysis has also shown that a dense linguistic context is necessary for robust global, and also local, syntactic error detection. While global errors can only be found if the global structure of a sentence is analyzed, local errors also face a number of challenges that can only be overcome by a global analysis. In a context where syntax is only partly reliable and predictable, homonymies, non-words and real word errors can exponentially increase the number of possible analyses of the error context and the relevant clue for finding the error. In order to reduce the number of analyses and identify the clue, the error context needs to be as linguistically dense as possible. A linguistically dense context analysis requires a rich lexicon enhanced by error analyses, valencies and semantic information. This information is then included in disambiguation, syntactic analysis and error detection rules.

Adapting the disambiguator to syntactically unreliable input is essential. The qualitative analysis showed that disambiguation errors play a significant role in unsuccessful error detection. They are responsible for almost 20% of the instances of unsuccessful valency error detection and for almost 70% of the instances of unsuccessful local case error detection. While the general procedure in rule-based grammar checking approaches consists of simple adaptations like rule-relaxing and adding homonymy-specific rules for certain ambiguities, I applied a more elaborate approach consisting of four steps that facilitate the work load of the error detection module. My aim in taking these measures was to avoid the removal of correct analyses and try to achieve an accurate disambiguation of the context of an error. Discarding default rules and relaxing systematic homonymy rules prevents the removal of correct analyses. Specific rules for idiosyncratic homonymies and systematic rules enhanced by valency and semantic tags are necessary for precise and sufficient disambiguation of the error context. Precise disambiguation is particularly rel-

evant in case error detection in post- or prepositional phrases as the postposition itself serves as the main indicator of the case of its dependent. A precise disambiguation of the adposition, which typically is homonymous with an adverb and sometimes with forms of other parts of speech, is therefore crucial for case error detection. As the relevant clue for the disambiguation of an adposition is typically the case of its dependent, which in the case of error detection is not reliable, disambiguation can only be performed by referring to semantic and/or idiosyncratic information. While semantic prototype information is the backbone of local case error detection, valency information is the backbone of global case error detection and valency error detection in general.

Although disambiguation that is enhanced with semantic and valency information is relevant for valency error detection, valency error detection cannot do without a deeper analysis of the government-argument structures in a sentence. The analysis of these structures is based on the valency tags added to each potential governor. Valency tag errors are responsible for a significant 40% of the cases of unsuccessful error detection of the valency structures of *beroštit* 'care'. Valency tags are used in simple context conditions negating a certain valency context or requiring it in local error detection rules. Real word error rules often draw from semantic and valency information if the syntactic contexts they can appear in are very similar. In global error detection rules, valency tags are predominantly used to associate governors with their arguments, which can then be referred to in the complex context conditions of the global error detection rules. Governor-argument structures are established by means of partial dependency analysis. Dependency annotation rules work incrementally, and closer contexts are tested for potential arguments before further contexts are tested. Each successfully established dependency relation between a governor and its arguments further facilitates the application of other dependency rules as an argument can only be matched to one governor. Semantic roles are then mapped to successfully matched arguments and ensure the distinction of different types of arguments. By means of this procedure one is able to reduce the potential nominal targets of the valency error detection rules by nearly half. As successfully matched arguments are considered grammatical structures, valency error detection rules can discard any form in the sentence that has received a semantic role as a possible target. The qualitative evaluation showed that dependency and semantic role annotation can also be relevant for local case error detection in adpositional phrases, especially when the potential adposition is both preceded and followed by a noun phrase of similar semantic prototype categories. Both local and global error detection rules have been shown to rely on semantic and valency information in different stages of the error detection process.

The evaluation of the rules for six real word error confusion pairs resulted in a mean precision of 98%, a recall of 72% and an accuracy of 87%. The evaluation of adposition error detection is based on five adpositions and shows high precision, 99%. Recall is 81% and accuracy is 90%. The qualitative evaluation shows further that significantly more

than half (67%) of the false negatives and positives are due to insufficient or erroneous disambiguation. Mean precision for the valency error detection regarding the four verbal governors is 77%, mean recall is 71% and mean accuracy is 96%. The small-scale evaluation of the error detection of these different error types shows that a dense linguistic annotation is worthwhile. Understanding a grammatically unreliable context and identifying a grammatical error in this context requires rich lexical information. A rich lexicon with semantic prototype tags and valency tags, a morphological analysis of common errors, an analysis of homonymies and confusion pairs creates a linguistically dense context facilitating local error detection and making global error detection possible.

To my knowledge, existing global error detection modules go little further than agreement error detection, and full-scale valency error detection has not been realized and evaluated by any documented grammar checker. Newer versions of *GoDivvun* have benefited from an error analysis and include an even richer analysis including a compound analysis and detailed error tags for real word and non-word error detection. Future versions can certainly be improved by expanding these small-scale studies on certain error types to a large-scale analysis of real word error confusion pairs, disambiguation of adpositions, and valency error analysis of verbal governors.

Part III

End

Chapter 6

Conclusion

This chapter concludes my investigation of valencies and syntactically relevant semantic categories in North Sámi and their integration in automatic linguistic analysis and error detection. The test case is *GoDivvun*, the grammar checker for North Sámi, with a focus on its syntactic modules. This syntactic analysis is based on machine-readable grammars with explicit rules that choose and reject a certain morpho-syntactic output of the sentence and its components. These rules require an annotation of word forms on different linguistic levels, including information about the lemma, morphological information, etc. Like a human, a machine-readable grammar analyzes a sentence by putting together information from different linguistic levels and based on this, selects or discards certain interpretations of a sentence.

Giella-sme, the infrastructure for North Sámi analysis, originally included a morpho-syntactic analysis based on finite-state transducers and Constraint Grammars. The new task of grammar checking requires a deeper syntactic and semantic analysis and focuses on grammatically ill-formed input. Both extensive homonoymy of well-formed input and possible grammatical errors in running text complicate a reliable sentence analysis based on the existing tools as the grammatical clues cannot be trusted. Inspired by the process of human parsing of a sentence, challenges in disambiguation and error detection were resolved by the addition of valencies to potential governors and semantic prototype categories to potential arguments. Valencies and semantic prototype categories were used to identify government-argument structures, which are central to a global syntactic analysis of a sentence. Additionally, the analysis was enhanced by a semantic layer one can refer to when morpho-syntactic information alone is unrealiable. The practical part of this work included the annotation of the *Giella-sme* lexica by means of semantic prototype categories (cf. Chapter 4), and the development of three Constraint Grammars: *valency.cg3* (cf. Chapter 3), *disambiguator.cg3*, and *grammarchecker.cg3* (cf. Chapter 5). These grammars make up the linguistic core of *GoDivvun*. The theoretical part of this work included a discussion of the theory and methodology (Chapter 2), and focuses on a description and evaluation of the previously mentioned grammars (Chapters 3 & 5) and

the annotation of the lexica (Chapter 4).

Chapter 2 gave a general overview of valency in linguistic research, previous research on valency in North Sámi, and its role in language technology. I described different levels of valency, in particular syntactic valency, semantic roles and semantic selection restrictions, which formed the theoretical basis for the valency tags in *valency.cg3*. In addition, I discussed different types of potential governors, and the distinctions between arguments and other parts of a sentence that are not considered part of a governor's valency. I described different approaches to distinguishing argument types and defining sets of semantic roles as the valency tags in *valency.cg3* required the definition of a set of semantic roles for North Sámi. With regard to implementation, I presented the framework, i.e. Constraint Grammar, of the grammars in *GoDivvun*. Lastly, I discussed the use of introspection and the corpus (*SIKOR*) when annotating valencies and making grammaticality decisions, and presented different measures used in the evaluation of the natural language processing tools.

In Chapter 3, I described the valency annotation grammar *valency.cg3* and gave an overview of the valencies of the 500 most frequent verbs North Sámi. I based my approach for North Sámi on Bick's (2000) valency tags for Portuguese. As in Bick's (2000) approach, valency tags were directly included in the morpho-syntactic analysis/disambiguation (*disambiguator.cg3*) and here in particular in grammar checking (*grammarchecker.cg3*) and could therefore be directly tested by the tool and adapted to its needs. There are two main differences to Bick's (2000) tags. Valency tags for North Sámi refer not only to syntactic valencies and semantic selection restrictions, but also to the argument types (i.e. semantic roles). Secondly, valency tags specify the whole argument constellation of a governor rather than referring to a single argument. In this context, I discussed the linguistic information referred to in the valency tags, i.e. semantic roles, selection restrictions, and morpho-syntactic distinctions. Based on a number of syntactic criteria, I distinguished between auxiliaries and main verbs in the valency annotation and also took into account a number of multi-word governors. As valencies are not referred to in a lexicon, but annotated by means of a grammar, there is room for dynamic processes, and context restrictions can be specified. These restrictions typically involve a number of valency-changing diathesis alternations or certain inflections of the governor. The resulting valency annotation grammar *valency.cg3* added one or multiple valency tags to 1,700 verbal, nominal and adjectival governors, including multi-word verbs, and thereby covered 73% of the verb cohorts in *SIKOR*. With only 7% of verb type coverage in *SIKOR* this is quite effective.

Chapter 4 dealt with syntactically relevant semantic prototype categories and the annotation of the North Sámi lexicon in *Giella-sme*, and an evaluation of its coverage. The annotation of valency tags and semantic prototype tags is the prerequisite for a number of language technological tasks, including dependency analysis, semantic role annotation,

morpho-syntactic disambiguation of, for example, accusative objects and genitive modifiers, and global error detection. The set of semantic prototype tags for North Sámi was organized hierarchically with the aim of drawing a complete semantic map of the world. I discussed a number of lexicon-related issues such as category membership of compounds and multiple membership in the case of polysemy, homonymy, etc., and handling of semantic prototype tags in dynamic compounding. While most compounds are predictable (i.e. head-final) or lexicalized in *Giella-sme*, a number of productive last elements that produce semantically unpredictable compounds showed that part of the semantic tagging will need to be resolved in a rule-based manner in future analyzers. In this work, 71% of the entries in the North Sámi noun lexicon were annotated with at least one valency tag, covering almost 90% of the nouns in *SIKOR*. Four test cases of adpositional phrases and verbal governor-argument constellations further showed the syntactic relevance of semantic prototype categories and their use in morpho-syntactic disambiguation (e.g. between adverbs and adpositions and between genitive modifiers and accusative objects), in error detection (e.g. lexical adposition errors), and in the lexical selection of e.g. polysemous verbs with different translation equivalents.

Chapter 5 dealt with the actual use of semantic prototype tags and valency tags in local and global grammatical error detection. As Tesnière (1959) and Helbig and Schenkel (1973) stressed early on, a formalization of valency information is necessary for the distinction between grammatical and ungrammatical constructions in second language learning. A grammar for the detection of grammatical errors essentially requires the same access to linguistic information (including valency information and semantic prototype information) as a language learner to analyze a sentence despite an error, and consequently identify the error. Semantic prototype tags and valency tags are included in all the grammars of the North Sámi grammar checker *GoDivvun*. These are *disambiguator.cg3*, which performs morpho-syntactic annotation and disambiguation, and *grammarchecker.cg3*, which performs dependency analysis, semantic role annotation and error detection. The compound disambiguation grammar *mwe-dis.cg3* also includes valency tags and semantic prototype tags. However, as it is a newer module, it was not discussed further in this work.

While certain real word errors required valency information and semantic prototype information in the context conditions of their error detection rules, local case errors required this information predominantly in the disambiguation process for the governor of the respective case, i.e. adverb-adpositional disambiguation. Valency error detection, on the other hand, required valency information on all stages, i.e. disambiguation, previous real word error detection, dependency annotation and semantic role annotation, and in their error detection rules. I analyzed six North Sámi rection verbs with regard to their grammatical and ungrammatical valencies and evaluated the valency rules regarding four of them.

The evaluation of six confusion pairs for real word errors resulted in a mean precision of

98%, a mean recall of 72% and a mean accuracy of 87%. The evaluation of adposition error detection was based on five adpositions and disambiguation rules that made extensive use of semantic tags. Precision was as high as 99%, recall was 81% and accuracy was 90%. For valency error detection of the small test set of four verbs, precision was 77%, recall was 71% and accuracy was 96%.

The grammar checkers described in the relevant literature mostly test local errors and very specific global errors, e.g. agreement errors. However, to my knowledge, no approach has attempted to fix a full range of valency errors in running text. In the course of this work, I successfully managed to detect valency errors based on a deep syntactic and semantic analysis. The main measures included making disambiguation more robust, dense and specific, including partial dependency analysis based on valency tags, adding semantic roles, and lastly searching for the error among unmatched potentially erroneous forms. The evaluation showed that a dense linguistic context including valencies and semantic prototypes is necessary for both global and local error detection. Despite the call for an integration of valency information and semantic information into linguistic analysis, to my knowledge, very few error detection approaches make use of valencies, and even fewer refer to semantic roles and dependencies.

This work has shown that while valency annotation is the backbone of global error detection, semantic prototype tagging is the backbone of local error detection. A rich lexical annotation including semantic prototypes, valencies and an annotation of typical real word errors and non-words provides a dense analysis of the context of homonymous forms, syntactically ambiguous forms and grammatical errors. It can be argued that it is close to impossible to pick out an analysis from the infinite possibilities without some healthy prejudices as to what it probably means. While this work has successfully resolved a number of challenges within the linguistic analysis of North Sámi, it also leaves some tasks for future work. Valency error detection definitely requires a closer analysis and annotation of ungrammatical valencies, possibly by means of explicit valency error tags. Also, non-traditional valencies (including "adjuncts") need to be investigated to avoid false positives in valency error detection of e.g. grammatical THEME-less constructions. In addition, completing and fine-tuning governor-argument dependency analysis and semantic role analysis is beneficial for a syntactic sentence analysis in general and both local and global error detection. A formalization of valencies and semantics is also necessary to ensure a correct realization of the arguments of a particular governor and the selection of the correct governor in machine translation.

An extensive valency description and a semantic annotation of the lexicon is clearly not only a cornerstone in frequently cited syntactic theories like Tesnière's (1959), but also the key to a deep syntactic analysis in advanced natural language processing.

Bibliography

Ahlberg, Malin, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal and Jonatan Uppström (2013), Korp and Karp – a bestiary of language resources: the research infrastructure of Språkbanken, *in* 'Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)', Vol. 16 of *NEALT Proceedings Series*, pp. 429–433.

Aldezabal, Izaskun (2004), Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz, PhD thesis, Euskal Filologia Saila. Zientzia Fakultatea. Leioa. Euskal Herriko Unibertsitatea.

Antonsen, Lene (2013), 'Čállinmeattáhusaid guorran [Tracking misspellings]', *Sámi dieđalaš áigečála* **2/2013**, 7–32.

Antonsen, Lene (2014), Evaluation of an fst-based spellchecker for North Saami. Presentation at SLTC, Uppsala, 13.11.2014.

Antonsen, Lene and Laura Janda (2015), 'Oamastanráhkadusat davvisámi girjjálašvuođas [Possessive constructions in North Saami prose]', *Dieđut* **2/2015**, 9–43.

Antonsen, Lene, Laura Janda and Berit Anne Bals Baal (2012), 'Njealji davvisámi adposišuvnna geavahus [The use of four North Sami adpositions]', *Sámi dieđalaš áigečála* **2/2012**, 7–38.

Antonsen, Lene, Linda Wiechetek and Trond Trosterud (2010), Reusing grammatical resources for new languages, *in* 'Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)', The Association for Computational Linguistics, Stroudsburg, pp. 2782–2789.

Antonsen, Lene and Trond Trosterud (2017), 'Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk', *Norsk Lingvistisk Tidsskrift* **35**, 153–185.

Aranzabe, María Jesús (2008), Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala, PhD thesis, Euskal Filologia Saila. Euskal Herriko Unibertsitatea.

Aristotle (1932), *Poetics*, Vol. 23 of *Aristotle in 23 Volumes*, Cambridge, MA, Harvard University Press, London. Translation by W.H. Fyfe.

Arppe, Antti (2000), Developing a grammar checker for Swedish, *in* T.Nordgård, ed., 'Proceedings of the 12th Nordic Conference of Computational Linguistics (NoDaLiDa 1999)', Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, pp. 13–27.

Atwell, Eric S. (1987), How to detect grammatical errors in a text without parsing it, *in* 'Proceedings of the 3rd Conference of the EACL', University of Copenhagen, Copenhagen, Denmark, pp. 38–45.

Baayen, Harald (1993), On frequency, transparency and productivity, *in* G.Booij and J.van Marle, eds, 'Yearbook of Morphology 1992', Springer, Dordrecht, pp. 181–208.

Badia, Toni, Àngel Gil, Martí Quixal and Oriol Valentín (2004), NLP-enhanced error checking for Catalan unrestricted text, *in* J.Carson-Berndsen, ed., 'Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)', European Language Resources Association, pp. 1919–1922.

Baker, Collin F., Charles J. Fillmore and John B. Lowe (1998), The Berkeley FrameNet project, *in* 'Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998), Volume 1', Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 86–90.

Banu, R.S.D. Wahida and R. Sathish Kumar (2004), Using selectional restrictions for real word error correction, *in* S.Manandhar, J.Austin, U.Desai, Y.Oyanagi and A.Talukder, eds, 'Proceedings of the Second Asian Applied Computing Conference (AACC 2004)', Vol. 3285 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 130–136.

Bartens, Raija (1972), *Inarilapin, merilapin ja luulajanlapin kaasussyntaksi*, Vol. 148 of *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, Suomalais-ugrilainen Seura, Helsinki.

Bartens, Raija (1978), *Synteettiset ja analyyttiset rakenteet lapin paikanilmauksissa*, Vol. 166 of *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, Suomalais-ugrilainen Seura, Helsinki.

Beesley, Kenneth R. and Lauri Karttunen (2003), *Finite State Morphology*, CSLI Studies in Computational Linguistics, CSLI Publications, Stanford.

Benešová, Václava, Markéta Lopatková and Klára Hrstková (2008), Enhancing Czech valency lexicon with semantic information from FrameNet: The case of communication verbs, *in* J.Webster, N.Ide and A. C.Fang, eds, 'Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)', City University of Hong Kong, Hong Kong, China, pp. 18–25.

Bergsland, Knut (1961), *Samisk grammatikk med øvelsesstykker*, second edn, Kirke- og undervisningsdepartementet, Oslo.

Beronka, Johan (1937), *Lappische Kasusstudien: zur Geschichte des Komitativ-Instruktivs und des Genitivs im Lappischen: 1*, Brøgger, Oslo.

Bick, Eckhard (2000), *The Parsing System 'Palavras': Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press, Aarhus.

Bick, Eckhard (2006*a*), A constraint grammar based spellchecker for Danish with a special focus on dyslexics, *in* M.Suominen, A.Arppe, A.Airola, O.Heinämäki, M.Miestamo,

U.Määttä, J.Niemi, K. K.Pitkänen and K.Sinnemäki, eds, 'A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday', Vol. 19/2006 of *Special Supplement to SKY Jounal of Linguistics*, The Linguistic Association of Finland, Turku, pp. 387–396.

Bick, Eckhard (2006*b*), Noun sense tagging: Semantic prototype annotation of a Portuguese treebank, *in* J.Hajic and J.Nivre, eds, 'Proceedings of the 5th Workshop on Treebanks and Linguistic Theories', Prague, pp. 127–138.

Bick, Eckhard (2007*a*), Automatic semantic role annotation for Portuguese, *in* 'Proceedings of the 5th Workshop on Information and Human Language Technology (TIL 2007) / Anais do XXVII Congresso da SBC', Rio de Janeiro, pp. 1713–1716.

Bick, Eckhard (2007*b*), Dan2eng: Wide-Coverage Danish-English Machine Translation, *in* B.Maegaard, ed., 'Proceedings of 11th Machine Translation Summit (MT Summit XI)', Copenhagen, Denmark, pp. 37–43.

Bick, Eckhard (2007*c*), 'Semantic roles for automatic corpus annotation'.
URL: *beta.visl.sdu.dk/∼eckhard/pdf/semantic_ roles_ manual.pdf (Accessed 2017-11-16)*

Bick, Eckhard (2009*a*), Deepdict – a graphical corpus-based dictionary of word relations, *in* K.Jokinen and E.Bick, eds, 'Proceedings of the 17th Nordic Conference of Computational Linguistics (NoDaLiDa 2009)', Vol. 4 of *NEALT Proceedings Series*, pp. 268–271.

Bick, Eckhard (2009*b*), 'Semantic prototype tags for nouns'.
URL: *http://visl.sdu.dk/semantic_ prototypes_ overview.pdf (Accessed 2017-11-28)*

Bick, Eckhard (2010), A dependency-based approach to anaphora annotation, *in* 'Extended Activities Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language Apr. 27-30', Porto Alegre, Brazil.
URL: *http://visl.sdu.dk/∼eckhard/pdf/PROPOR2010_ anaphora.pdf (Accessed 2017-11-29)*

Bick, Eckhard (2011), A FrameNet for Danish, *in* 'Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa 2011)', NEALT Proceedings Series, Tartu University Library, Tartu, pp. 34–41.

Bick, Eckhard (2012), Towards a semantic annotation of English television news – building and evaluating a Constraint Grammar FrameNet, *in* 'Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation', Faculty of Computer Science, Universitas Indonesia, pp. 60–69.

Bick, Eckhard (2013), ML-tuned Constraint Grammars, *in* S.-C.Tseng, ed., 'Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation', Department of English, National Chengchi University, Taipei, Taiwan, pp. 440–449.

Bick, Eckhard (2015), DanProof: Pedagogical spell and grammar checking for Danish, *in* G.Angelova, K.Bontcheva and R.Mitkov, eds, 'Proceedings of the 10th International Conference Recent Advances in Natural Language Processing (RANLP 2015)', INCOMA Ltd., Hissar, Bulgaria, pp. 55–62.

Bick, Eckhard (2017*a*), From treebank to Propbank: A semantic-role and VerbNet corpus for Danish, *in* 'Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa 2017)', Association for Computational Linguistics, Gothenburg, Sweden, pp. 202–210.

Bick, Eckhard (2017*b*), Propbank annotation of Danish noun frames, *in* 'Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)', Association for Computational Linguistics.

Bick, Eckhard and Pilar Valverde (2009), Automatic semantic role annotation for Spanish, *in* 'Proceedings of the 17th Nordic Conference of Computational Linguistics (NoDaLiDa 2009)', Vol. 4 of *NEALT Proceedings Series*, Tartu University Library, Tartu, pp. 215–218.

Bick, Eckhard and Tino Didriksen (2015), CG-3 – beyond classical Constraint Grammar, *in* B.Megyesi, ed., 'Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)', Linköping University Electronic Press, Linköpings universitet, pp. 31–39.

Birn, Juhani (2000), Detecting grammar errors with Lingsoft's Swedish grammar checker, *in* T.Nordgård, ed., 'Proceedings of the 12th Nordic Conference of Computational Linguistics (NoDaLiDa 1999)', Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, pp. 28–40.

Borges, Jorge Luis (1960), *El hacedor*, Emecé Editores, Buenos Aires.

Borges, Jorge Luis (1999), *Collected fictions*, Penguin Books, New York. Translation by Andrew Hurley.

Borin, Lars and Markus Forsberg (2009), All in the family: A comparison of SALDO and WordNet, *in* B.Sandford Pedersen, A.Braasch, S.Nimb and R.Vatvedt Fjeld, eds, 'Proceedings of the NoDaLiDa 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies', Vol. 7 of *NEALT Proceedings Series*.

Borin, Lars, Markus Forsberg and Johan Roxendal (2012), Korp – the corpus infrastructure of språkbanken, *in* N.Calzolari, K.Choukri, T.Declerck, M. U.Doğan, B.Maegaard, J.Mariani, J.Odijk and S.Piperidis, eds, 'Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)', European Language Resources Association (ELRA).

Borin, Lars, Markus Forsberg and Lennart Lönngren (2013), 'SALDO: a touch of yin to WordNet's yang', *Language Resources and Evaluation* **47**(4), 1191–1211.

Borin, Lars, Markus Forsberg and Lennart Lönngren (2008), *SALDO 1.0 (Svenskt associationslexikon version 2)*, Språkbanken, Göteborg universitet.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer and Anne Schumacher (2016), Sparv: Språkbanken's corpus annotation pipeline infrastructure, *in* 'Proceedings of the 6th Swedish Language Technology Conference (SLTC 2016)', Umeå University, 17-18 November, 2016.

Carlberger, Johan, Rickard Domeij, Viggo Kann and Ola Knutsson (2004), The development and performance of a grammar checker for Swedish: A language engineering perspective, Technical report, Kungl Tekniska Högskolan, Stockholm, Sweden.
URL: *http://www.csc.kth.se/tcs/projects/granska/rapporter/granskareport.pdf (Accessed 2017-11-29)*

Catt, Mark (1988), Intelligent diagnosis of ungrammaticality in computer-assisted language instruction, Master's thesis, Department of Computer Science, University of Toronto.

Chomsky, Noam (1981), *Lectures on Government and Binding: The Pisa Lectures*, Foris Publications Holland, Dordrecht.

Čállinrávagirji (2003), Sámedikki giellaossodat/ Sámedikki oahpahusossodat, Guovdageaidnu.
URL: *http://bit.ly/2yRk15C (Accessed 2017-11-29)*

Didriksen, Tino (2010), *Constraint Grammar Manual: 3rd version of the CG formalism variant*, GrammarSoft ApS, Denmark.
URL: *http://visl.sdu.dk/cg3/vislcg3.pdf (Accessed 2017-11-29)*

Díaz de Ilarraza, Arantza, Koldo Gojenola and Maite Oronoz (2010), Evaluating the impact of morphosyntactic ambiguity in grammatical error detection, *in* H.Loftsson, E.Rögnvaldsson and S.Helgadóttir, eds, 'Proceedings of the 7th International Conference on NLP (IceTAL 2010)', Advances in Natural Language Processing, Springer, pp. 155–160.

Estarrona, Ainara, Izaskun Aldezabal, Arantza Díaz de Ilarraza A. and María Jesús Aranzabe (2016), 'Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/Verbnet model', *Digital Scholarship in the Humanities* **31**(3), 470–492.

Faulhaber, Susen (2011), *Verb Valency Patterns: A Challenge for Semantics-Based Accounts*, Vol. 71 of *Topics in English linguistics*, De Gruyter Mouton, Berlin, New York.

Fillmore, Charles J. (1968), The case for case, *in* E.Bach and R. T.Harms, eds, 'Universals in Linguistic Theory', Holt, Rinehart and Winston, New York, pp. 1–88.

Fillmore, Charles J. (1971), Some problems of case grammar, *in* R. J.O'Brien, ed., 'Proceedings of the 22nd Annual Round Table Meeting on Linguistics and Language Studies', Vol. 24 of *Monograph Series on Languages and Linguistics*, Georgetown University Press, New York, pp. 35–56.

Fillmore, Charles J., Christopher R. Johnson and Miriam R. L. Petruck (2003), 'Background to FrameNet', *International Journal of Lexicography* **16**(3), 235–250.

Fischer, Klaus (1997), *German-English verb valency: A contrastive analysis*, Tübinger Beiträge zur Linguistik, Narr, Tübingen.

Fliedner, Gerhard (2001), Überprüfung und Korrektur von Nominalkongruenz im Deutschen, Diplomarbeit, Universität des Saarlandes, Saarbrücken.

Friis, Jens Andreas (1856), *Lappisk Grammatik: udarbeidet efter den finmarkiske Hoveddialekt eller Sproget, saaledes som det almindeligst tales i norsk Finmarken*, J. W. Cappelen, Christiania.

Gildea, Daniel and Daniel Jurafsky (2002), 'Automatic labeling of semantic roles', *Computational Linguistics* **28**(3), 245–288.

Glare, P.G.W., ed. (1983), *Oxford Latin Dictionary*, Clarendon Press.

Groot, Albert Willem de (1949), *Structurele Syntaxis*, Servire, The Hague.

Gruber, Jeffrey S. (1965), Studies in Lexical Relations, PhD thesis, MIT, Cambridge, Massachusetts. Reprinted in 1976.

Hagen, Kristin, Janne Bondi Johannessen and Pia Lane (2001), Some problems related to the development of a grammar checker, *in* A.Sågvall Hein, ed., 'Proceedings of the 13th Nordic Conference in Computational Linguistics (NoDaLiDa 2001)', Department of Linguistics, Uppsala University.

Hagen, Kristin, Pia Lane and Trond Trosterud (2001), 'En grammatikkontroll for bokmål', *Språknytt* **3/2001**, 6–9;47.

Hardwick, Sam, Miikka Silfverberg and Krister Lindén (2015), Extracting semantic frames using hfst-pmatch, *in* B.Magyesi, ed., 'Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)', Vilnius, Lithuania, pp. 305–308.

Hashemi, Sylvana Sofkova (2003), Automatic Detection of Grammar Errors in Primary School Children's Texts. A Finite State Approach, PhD thesis, Department of Linguistics, Göteborg University.

Haugen, Tor Arne (2013), 'Adjectival valency as valency constructions: Evidence from Norwegian', *Constructions and Frames* **5**(1), 35–68.

Helander, Nils Øyvind (2001), *Ii das šat murrii iige báktái: Davvisámegiela illatiivva geavaheapmi*, Vol. 1/2001 of *Dieđut*, Sámi Instituhtta, Guovdageaidnu.

Helbig, Gerhard (1992), *Probleme der Valenz- und Kasustheorie*, Vol. 51 of *Konzepte der Sprach- und Literaturwissenschaft*, Max Niemeyer Verlag, Tübingen.

Helbig, Gerhard and Wolfgang Schenkel (1973), *Wörterbuch zur Valenz und Distribution deutscher Verben*, De Gruyter, Berlin, Boston.

Izquierdo Beviá, Rubén, Armando Suárez Cueto and German Rigau Claramunt (2007), Exploring the automatic selection of basic level concepts, *in* G.Angelova, K.Bontcheva, R.Mitkov, N.Nicolov and N.Nikolov, eds, 'Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP 2007)', Borovetz, Bulgaria, pp. 298–302.

Johannessen, Janne Bondi, Kristin Hagen and Pia Lane (2002), The performance of a grammar checker with deviant language input, *in* S.-C.Tseng, ed., 'Proceedings of the 19th International Conference on Computational Linguistics', Taipei, Taiwan, pp. 1223–1227.

Karlsson, Fred (1990), Constraint Grammar as a Framework for Parsing Running Text, *in* H.Karlgren, ed., 'Proceedings of the 13th Conference on Computational Linguistics (COLING 1990)', Vol. 3, Association for Computational Linguistics, Helsinki, Finland, pp. 168–173.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä and Arto Anttila (1995), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin.

Keidan, Artemij (2011), 'The kāraka-vibhakti device as a heuristic tool for the compositional history of pāṇini's aṣṭādhyāyī', *Rivista Degli Studi Orientali* **84**, 273–288.

Kettnerová, Václava and Markéta Lopatková (2013), The representation of Czech light verb constructions in a valency lexicon, *in* 'Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)', Matfyzpress, Praha, Czech Republic, pp. 147–156.

Kettnerová, Václava and Markéta Lopatková (2015), At the lexicon-grammar interface: The case of complex predicates in the functional generative description, *in* 'Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)', Uppsala University, Uppsala, Sweden, pp. 191–200.

Kettnerová, Václava, Markéta Lopatková and Eduard Bejček (2012), 'Mapping semantic information from FrameNet onto VALLEX', *The Prague Bulletin of Mathematical Linguistics* **97**, 23–41.

Kipper Schuler, Karen (2005), VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon, PhD thesis, University of Pennsylvania, Philadelphia, PA.

Kittilä, Seppo and Jussi Ylikoski (forthcoming), Some like it transitive: Remarks on verbs of liking and the like in the Saami languages, *in* I.Seržant, A.Witzlack-Makarevich and K.Mann, eds, 'The diachronic typology of differential argument marking', Studies in Diversity Linguistics, Language Science Press, Berlin.

Kittilä, Seppo, Katja Västi and Jussi Ylikoski (2011), Introduction to case, animacy and semantic roles, *in* S.Kittilä, K.Västi and J.Ylikoski, eds, 'Case, Animacy and Semantic Roles', Vol. 99 of *Typological Studies in Language*, John Benjamins, Amsterdam, pp. 1–26.

Lagercrantz, Eliel (1929), *Sprachlehre des Nordlappischen nach den seelappischen Mundarten*, Vol. 3 of *Bulletin*, Oslo etnografiske museum, Oslo.

Lakoff, George (1987), *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*, University of Chicago Press, Chicago.

Lakoff, George and Mark Johnson (1980), *Metaphors We Live By*, University Of Chicago Press, Chicago.

Leem, Knud (1748), *En Lappisk Grammatica: Efter den Dialect, som bruges af Field-Lapperne udi Porsanger-Fiorden*, Gottman Friderich Kisel, Kiøbenhavn.

Levin, Beth (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago.

Liin, Krista (2008), Reeglipõhine komavigade tuvastaja eestikeelsetele tekstidele, Master's thesis, Tartu Ülikool, Tartu.
URL: *http://lepo.it.da.ut.ee/∼kaili/juhendamised/kliinmag.pdf (Accessed 2017-11-29)*

Lopatková, Markéta and Jarmila Panevová (2005), Recent developments in the theory of valency in the light of the Prague Dependency Treebank, *in* M.Šimková, ed., 'Insight into Slovak and Czech Corpus Linguistics', Veda, Bratislava, Slovakia, pp. 83–92.

Lopatková, Markéta, Ondřej Bojar, Jiří Semecký, Václava Benešová and Zdeněk Žabokrtský (2005), Valency lexicon of Czech verbs VALLEX: Recent experiments with frame disambiguation, *in* V.Matoušek, P.Mautner and T.Pavelka, eds, 'Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)', Vol. 3658 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 99–106.

Lopatková, Markéta, Zdeněk Žabokrtský and Karolína Skwarska (2006), Valency lexicon of Czech verbs: Alternation-based model, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)', ELRA, Genova, Italy, pp. 1728–1733.

Magga, Ole Henrik (1980), *Giellaoahppa: Jietna-, hápme- ja cealkkaoahppa oanehaččat čilgejuvvon*, Vol. 3/1980 of *Dieđut*, Sámi instituhtta, Guovdageaidnu.

Magga, Ole Henrik (1982), *Modalverb og infinitiv innen verbalet: prosjektrapport*, Vol. 1/1982 of *Dieđut*, Sámi instituhtta, Guovdageaidnu.

Magga, Ole Henrik (1986), Studier i samisk infinitivsyntaks: Del 1: Infinitivsetning. Akkusativ og infinitiv, PhD thesis, Universitetet i Oslo.

Magga, Ole Henrik (2002), 'Muhtun čuolmmat sámi cealkkaoahpas', *Sámi dieđalaš áigečála* **1/2002**, 59–69.

Mikalsen, Anne Dagmar Biti (1993), Rekšuvdnavearbbat, Master's thesis, Tromssa universitehtta/Sámi allaskuvla, Tromsø.

Miller, George A. (1990), 'Nouns in WordNet: A lexical inheritance system', *International Journal of Lexicography* **3**(4), 245–264. Revised August 1993.
URL: *http://pami.uwaterloo.ca/∼khoury/ece457f07/Miller1993.pdf (Accessed 2017-11-30)*

Miller, George A. (1995), 'WordNet: A lexical database for English', *Communications of the ACM* **38**(11), 39–41.

Moshagen, Sjur, Tommi Pirinen and Trond Trosterud (2013), Building an open-source development infrastructure for language technology projects, *in* 'Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)', Vol. 16 of *NEALT Proceedings Series*, pp. 343–352.

Nickel, Klaus Peter (1994), *Samisk grammatikk*, second edn, Davvi Girji, Kárášjohka.

Nickel, Klaus Peter and Pekka Sammallahti (2011), *Nordsamisk grammatikk*, Davvi Girji, Kárášjohka.

Nielsen, Konrad (1926-1929), *Lærebok i lappisk. I. Grammatikk*, second edn, A.W. Brøggers boktrykkeris forlag, Oslo. Reprinted in 1979.

Nielsen, Konrad (1932-1960*a*), *Lappisk (samisk) ordbok. Lapp Dictionary. Vol. I A-F*, second edn, A.W. Brøggers boktrykkeris forlag, Oslo. Reprinted in 1979.

Nielsen, Konrad (1932-1960*b*), *Lappisk (samisk) ordbok. Lapp Dictionary. Vol. II G-M*, second edn, A.W. Brøggers boktrykkeris forlag, Oslo. Reprinted in 1979.

Nielsen, Konrad (1932-1960*c*), *Lappisk (samisk) ordbok. Lapp Dictionary. Vol. III N-Æ*, second edn, A.W. Brøggers boktrykkeris forlag, Oslo. Reprinted in 1979.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi (2007), 'MaltParser: A language-independent system for data-driven dependency parsing', *Natural Language Engineering* **13**(2), 95–135.

Oronoz, Maite (2009), Euskarazko errore sintaktikoak detektatzeko eta zuzentzeko baliabideen garapena: datak, postposizio-lokuzioak eta komunztadura, PhD thesis, Lengoaia eta Sistema Informatikoak Saila. Donostia. Euskal Herriko Unibertsitatea.

Oronoz, Maite, Arantza Díaz de Ilarraza and Koldo Gojenola (2010), Design and evaluation of an agreement error detection system: Testing the effect of ambiguity, parser and corpus type, *in* H.Loftsson, E.Rögnvaldsson and S.Helgadóttir, eds, 'Proceedings of the 7th international conference on Advances in natural language processing (IceTAL 2010)', Vol. 6233 of *Lecture Notes in Computer Science*, Springer, pp. 281–292.

Panevová, Jarmila (1974), 'On verbal frames in functional generative description. part 1', *The Prague Bulletin of Mathematical Linguistics* **22**, 3–40.

Panevová, Jarmila (1994), Valency frames and the meaning of the sentence, *in* P. A.Luelsdorff, ed., 'The Prague School of Structural and Functional Linguistics', John Benjamins Publishing Company, Amsterdam, pp. 223–243.

Pasierbsky, Fritz (2003), Toward a classification of complements, *in* V.Agel, L. M.Eichinger, H. W.Eroms, H. J.Heringer and P.Hellwig, eds, 'Dependenz und Valenz/Dependency and Valency. Part 1', Walter de Gruyter, pp. 803–813.

Pedler, Jennifer (2007), Computer Correction of Real-word Spelling Errors in Dyslexic Text, PhD thesis, London University, Birkbeck.

Petrović, Sonja (2009), Collecting and processing error samples for a Constraint Grammar-based language helper for Esperanto, Master's thesis, Stockholms Universitet.

Pinker, Steven (1989), *Learnability and Cognition: The Acquisition of Argument Structure*, MIT Press, Cambridge, MA.

Pitkänen, Harri (2006), Hunspell-fi in kesäkoodi 2006: Final report, Technical report.
URL: *http://www.puimula.org/htp/archive/kesakoodi2006-report.pdf (Accessed 2017-11-30)*

Pope, Kirsten and Máret Sárá (2004), *Eatnigiella. Giellaoahpu váldogirji*, Davvi Girji, Kárášjohka.

Rask, Rasmus K. (1832), *Rœsonneret lappisk Sproglære efter den Sprogart, som bruges af Fjældlapperne i Porsangerfjorden i Finmarken: en Omarbejdelse af Prof. Knud Leems Lappiske Grammatica*, Schubothe, Kiøbenhavn.

Riektačállinrávvagat (2015), Sámedikki giellaossodat/Sámedikki oahpahusossodat, Guovdageaidnu.
URL: *https://www.sametinget.no/content/download/870/13825 (Accessed 2017-11-3)*

Rosch, Eleanor (1973), 'Natural categories', *Cognitive Psychology* **4**(3), 328–350.

Ruong, Israel (1970), *Min sámegiella. Lärobok i samiska*, Utbildningsförlaget, Stockholm.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson and Jan Scheffczyk (2010), *FrameNet II: Extended Theory and Practice*, International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Sammallahti, Pekka (2005), *Láidehus sámegiela cealkkaoahpa dutkamii*, Davvi Girji, Kárášjohka.

Sammallahti, Pekka (2007), *Gielladutkama terminologiija*, Davvi Girji, Kárášjohka.

Sammallahti, Pekka and Klaus Peter Nickel (2006), *Sámi-duiskka sátnegirji=Saamisch-deutsches Wörterbuch*, Davvi Girji, Kárášjohka.

Sara, Laila Susanne (2002), Suorggádusaid geavaheapmi: Guovdageainnu nuoraid giella-geavaheapmi, MPhil thesis, University of Tromsø.

Sgall, Petr (1980), 'Case and meaning', *Journal of Pragmatics* **4**(6), 525–536.

SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection (2015-03-01), **URL:** *http://gtweb.uit.no/korp (Accessed 2015-03-01).*

SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection (2016-12-08), **URL:** *http://gtweb.uit.no/korp (Accessed 2016-12-08).*

Simons, Gary F. and Charles D. Fennig, eds (2017), *Ethnologue: Languages of the World*, twentieth edn, SIL International, Texas.
URL: *http://www.ethnologue.com (Accessed 2017-11-30)*

Stockfleth, Nils Christian Vibe (1840), *Grammatik i det lappiske Sprog, saaledes som det tales i norsk-Finmarken: Første Del: Bogstav- og Formlæren*, Grøndahl, Christiania.

Svonni, Mikael (2013), *Sátnegirji: davvisámegiela-ruoŧagiela, ruoŧagiela-davvisámegiela = Ordbok: nordsamisk-svensk, svensk-nordsamisk*, ČálliidLágádus, Kárášjohka.

Svonni, Mikael (2015), *Davvisámegiella – sánit ja cealkagat: Láidehus sámi lingvistihkkii*, Ravda lágádus, Giron/Kiruna.

Tarvainen, Kalevi (2011), *Einführung in die Dependenzgrammatik*, Vol. 35 of *Reihe Germanistische Linguistik*, second edn, Max Niemeyer Verlag, Berlin, Boston.

Tesnière, Lucien (1959), *Elements of structural syntax*, John Benjamins, Amsterdam. Translation by Timothy John Osborne and Sylvain Kahane (2015).

Trosterud, Trond (2003), Morfologi og leksikalsk semantikk, *in* P.Bye, T.Trosterud and Ø.Vangsnes, eds, 'Språk og språkvitskap. Ei innføring i lingvistikk', Det Norske Samlaget, Oslo, pp. 49–122.

Trosterud, Trond (2004), Porting morphological analyses and disambiguation to new languages, *in* J.Carson-Berndsen, ed., 'Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)', European Language Resources Association, Paris, pp. 90–92.

Trosterud, Trond (2006), Grammar-based language technology for the Sámi languages, *in* 'Proceedings of the Conference on Lesser Used Languages and Computer Linguistics (LULCL 2005)', Europäische Akademie, Bozen, Italy, pp. 133–148.

Trosterud, Trond and Linda Wiechetek (2007), Disambiguering av homonymi i nord- og lulesamisk, *in* J.Ylikoski and A.Aikio, eds, 'Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahti miessemánu 21. beaivve 2007', Vol. 253 of *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, Suomalais-Ugrilainen Seura, Helsinki, pp. 375–395.

Uria, Larraitz (2009), Euskarazko erroreen eta desbideratzeen analisirako lan-ingurunea. Determinatzaile-erroreen azterketa eta prozesamendua, PhD thesis, Euskal Filologia Saila. Euskal Herriko Unibertsitatea.

Uszkoreit, Hans (1996), Grammar checking. theory, practice, and lessons learned in LATESLAV. Concluding oral presentation at the final review meeting of the Lateslav Project (PECO 2824), Prague.

Vernerová, Anna, Václava Kettnerová and Markéta Lopatková (2014), To pay or to get paid: Enriching a valency lexicon with diatheses, *in* 'Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)', European Language Resources Association, Reykjavík, Iceland, pp. 2452–2459.

Vinka, Mikael (2002), Causativization in North Sámi, PhD thesis, McGill University, Montréal, Canada.

Vuolab, Káre Márjá (1996), Akkusatiivaobjeakta, *in* V.Guttorm, ed., 'Proceedings of Čavčča 1995 sámegiela ja sámi girjjálašvuođa dutkan- ja bagadansymposia: symposiaraporta nr. III', Sámi instituhtta, Guovdageaidnu, pp. 49–53.

Wedbjer Rambell, Olga (1999), Error Typology for Automatic Proof-reading Purposes, PhD thesis, Uppsala University, Department of Linguistics.

Wiechetek, Linda (2012), Constraint Grammar based correction of grammatical errors for North Sámi, *in* G. D.Pauw, G.-M.de Schryver, M.Forcada, K.Sarasola, F.Tyers and P.Wagacha, eds, 'Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)', European Language Resources Association (ELRA), Istanbul, Turkey, pp. 35–40.

Wiechetek, Linda, Francis M. Tyers and Thomas Omma (2010), Shooting at flies in the dark: Rule-based lexical selection for a minority language pair, *in* H.Loftsson, E.Rögnvaldsson and S.Helgadóttir, eds, 'Proceedings of the 7th International Conference on NLP (IceTAL 2010)', Vol. 6233 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 418–429.

Wiechetek, Linda and Jose Mari Arriola (2011), An experiment of use and reuse of verb valency in morphosyntactic disambiguation and machine translation for Euskara and

North Sámi, *in* E.Bick, K.Hagen, K.Müürisep and T.Trosterud, eds, 'Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa 2011) Workshop Constraint Grammar Applications', Vol. 14 of *NEALT Proceedings Series*, Northern European Association for Language Technology, pp. 61–69.

Wierzbicka, Anna (1996), *Semantics: Primes and Universals*, Oxford University Press, Oxford.

Ylikoski, Jussi (2006), 'Davvisámegiela -nláhkai ∼ -nládje -suffiksála ii-finihtta vearbaráhkadusat [The North Saami non-finite in –nláhkai ∼ –nládje]', *Sámi dieđalaš áigečála* **1/2006**, 18–38.

Ylikoski, Jussi (2009), *Non-finites in North Saami*, Vol. 257 of *Suomalais-ugrilaisen seuran toimituksia = Mémoires de la Société Finno-Ougrienne*, Suomalais-Ugrilainen Seura, Helsinki.

Ylikoski, Jussi (2016), 'Future time reference in Lule Saami, with some remarks on Finnish', *Eesti ja soome-ugri keeleteaduse ajakiri = Journal of Estonian and Finno-Ugric Linguistics* **7**(2), 209–244.

Čech, Radek, Petr Pajas and Ján Mačutek (2010), 'Full valency: Verb valency without distinguishing complements and adjuncts', *Journal of Quantitative Linguistics* **17**(4), 291–302.

Žabokrtský, Zdeněk and Markéta Lopatková (2007), 'Valency information in VALLEX 2.0: Logical structure of the lexicon', *The Prague Bulletin of Mathematical Linguistics* **87**, 41–60.

Appendix A

The 500 most frequent verbs in *SIKOR*

SIKOR **(disambiguated)**	**Verb**	**Other lemmata that have homonymous forms**
1,088,872	leat	"leahki", "leapma"
273,088	ii	"allut", "amas", "amasmuvvat", "amastit", "amat"
121,456	galgat	"gal", "galgamuš"
99,295	sáhttit	"sáhttitbehtet", "sáhttot", "sáhttu", "sáhtán"
70,422	oažžut	"oaččohit", "oaččostit", "oaččostuvvat", "oažžul", "oažžun", "oččodit", "ožžodit", "oččohallat", "ožžos", "ožžoš"
62,462	lohkat	"loahkkit", "loahku", "logadit", "logahaddat", "logahahtti", "logahalahit", "logahallan", "logahallat", "logahallojupmi", "logahat", "logaheapme", "logaheapmi", "logahit", "logastit", "logus", "lohka", "lohkalit", "lohkameahttun", "lohkamuš", "lohkan", "lohkka", "lohkket", "lohkki", "lohkkot", "lohkkádit", "lohku"
52,506	fertet	-
48,625	boahtit	"boahtti", "boađihit", "bohtat", "bohtti"
45,773	šaddat	"šaddadit", "šaddan", "šaddi"
37,702	muitalit	-
35,396	dahkat	"dagahit", "dahkalit", "dahkamuš", "dahkan", "dahkki"
34,938	bargat	"bargi", "bargu"
34,499	dadjat	"dajadit"
34,311	váldit	"váldu"
34,006	addit	"addi", "addin", "addu"
33,222	beassat	"beasadit", "beassadit", "beassan", "beassi", "bessen", "besset"
31,134	mearridit	"mearrat", "mearridit"
29,720	bidjat	"bidjan", "bidjet", "bijahit", "bijat"
28,360	geavahit	"geavaheapmi"
28,317	mannat	"manadit", "manahit", "mannan", "manne", "mannet", "mannut", "mánná"
26,517	áigut	-
25,303	oaidnit	"oaidni", "oaidnu", "oainnihit", "oidnot"
24,690	háliidit	-
22,200	gullat	"gulahit", "gulladit", "gullan", "gullet", "gulli", "gullot"
22,069	čájehit	-
21,878	ovddidit	"ovddidit", "ovddit"
21,148	ráhkadit	"ráhkadeapmi", "ráhkadus"
21,024	orrut	"orodit", "orrostit", "orrot", "orru", "orrun", "orut"
20,693	diehtit	"diehtti", "dieđihit", "dihtti"
19,066	čállit	"čálihit", "čálli", "čállin", "čállosupmái", "čállu", "čálán"
18,683	atnit	-
17,227	ásahit	"ása", "ásadit", "ásahus", "ássat"
16,519	oaivvildit	-

15,906	čilget	"čielgat", "čilgedit", "čilgen", "čilgestit"
15,609	guoskat	"guoskkahit"
15,280	dohkkehit	"dohkket"
14,766	álggahit	"álgga", "álggadit", "álggahus"
14,180	dárbbašit	"dárbbaš"
14,078	deattuhit	"deaddu"
13,458	geahččat	"geahčadit", "geahčahit", "geahčastallat", "geahčastit", "geahččaladdat", "geahččalit", "geahčči", "geahčču"
13,440	álgit	"Álgu", "álgu", "álgáge"
13,233	čađahit	"čađaheapmi", "čađat", "čađđa"
12,979	nannet	-
12,879	gáibidit	-
12,299	buktit	"buvttihit"
12,173	čujuhit	"čujuhus"
12,147	lágidit	-
12,090	doallat	"doaladit", "doalahit", "doalan", "doallan", "doalli", "doallu", "dollehit", "dollet"
12,022	ohcat	"oahcut", "ohca", "ohcalit", "ohcan", "ohcci"
12,010	evttohit	"eavttuheapme", "evttohus"
11,856	bivdit	"bivdet", "bivdi", "bivdu"
11,816	sáddet	"sáddedit", "sáddehit", "sádden"
11,804	árvvoštallat	"árvvoštallan", "árvvoštalli", "árvvoštit"
11,477	gávdnat	"gávdni", "gávdnot", "gávnnadit", "gávnnahit"
11,448	válljet	"válljen"
11,298	máksit	"máksu"
11,091	berret	"bearrat"
10,718	čuovvut	"čuovvu", "čuovvulit", "čuvodit"
10,367	jáhkkit	"jáhkkemeahttun", "jáhkku"
10,362	gávdnot	"gávdnat"
10,169	almmuhit	"almmuheapmi"
10,143	juolludit	"juollut"
10,089	vuolgit	"vulgot"
10,062	searvat	"searvadit", "searvan", "searvi", "searvvadit", "searvvuš"
9,994	doaibmat	"doaibma", "doaibmi", "doaimmahat", "doaimmahit", "doibmet"
9,992	váikkuhit	"váikkuheapmi", "váikkuhus"
9,800	dagahit	"dagadit", "dahkat"
9,565	cealkit	"cealki", "cealkilit"
9,541	dieđihit	"diehtit"
9,491	geahččalit	"geahččaladdat", "geahččaleapmi", "geahččat"
9,425	oahppat	"oahpadit", "oahpahit", "oahppa", "oahppi", "oahppu"
9,249	meannudit	-
9,222	doarjut	-
9,133	sihkkarastit	"sihkkarasti", "sihkkarastin"
9,046	rievdadit	"rievdadallat", "rievdadus"
8,742	joatkit	"joatki"
8,571	mearkkašit	"mearka"
8,369	vástidit	-
8,344	vuordit	"vuorddihit", "vuordilit", "vuordin", "vurdet"
8,261	dáhpáhuvvat"	-
8,099	hukset	"huksehit", "huksen"
8,005	dovdat	"dovdan", "dovdda", "dovddadit", "dovddahit", "dovddastit"
7,757	vuoruhit	"vuorru"
7,557	jearrat	"jeara", "jearadit", "jearahit", "jearralit", "jearru"
7,508	vuoitit	"vuoiti", "vuoitu"
7,276	hálddašit	"hálddašeapmi"
7,269	namuhit	-
7,173	dáhttut	"dáhttu", "dáhttun"
7,096	lassánit	"lassáneapmi"

7,081	govvet	"govva", "govven"
6,943	heivehit	"heivedit", "heivehallat", "heivet"
6,911	vuovdit	"vuovdi", "vuovdin", "vuovdái", "vuvdot"
6,775	máhttit	"máhttu", "máhttá"
6,670	soaitit	-
6,660	duođaštit	"duohta", "duođas", "duođaštus"
6,583	oastit	"oasti", "oastin", "oasttestit", "oasttistit"
6,567	jurddašit	-
6,518	nagodit	-
6,395	juohkit	"juhkat"
6,376	lasihit	"lasidit", "lasiheapmi", "lassi"
6,222	gohčodit	"gohčču̇t"
6,171	sávvat	"sávadit", "sávvamis"
6,162	sirdit	"sirdu", "sirdán"
6,056	loahpahit	"loahpaheapmi", "loahppa", "loahppat"
5,937	vásihit	"vássit"
5,857	buoridit	"buorre", "buorri"
5,755	vuodjit	"vuoddji", "vuodjat", "vuodjin", "vuodjut" "vuojehit", "vuojihit", "vuoján"
5,695	gustot	"gusto"
5,664	guorahallat	"guorahallan", "guorrat"
5,430	ballat	"baladit", "balahit", "ballu"
5,405	beroštit	-
5,358	nammadit	-
5,352	fitnat	"fidnehit", "fidnu", "finadit", "finála", "fitnet"
5,211	eallit	"ealihit", "ealli", "eallin", "eallu", "ealán" "ealát", "eleš", "ellot"
5,161	veahkehit	"veahkehallat", "veahkihit", "veahkki"
5,117	áddet	"ádden", "áddestallat"
4,972	lávet	"lávvi"
4,854	fállat	"fáladit", "fális", "fálli", "fállot"
4,833	ávžžuhit	"ávžu", "ávžut", "ávžžuhus"
4,767	massit	-
4,737	jođihit	"johtit", "jođiheapme"
4,709	deavdit	-
4,655	árvalit	"árvalus"
4,631	ássat	"ásadit", "ásahit", "ásat", "ássi", "ássut"
4,609	oahpahit	"oahpadit", "oahpahallat", "oahpaheapme", "oahpahus", "oahppat"
4,603	doalahit	"doaladit", "doallat"
4,596	gč	-
4,443	čohkket	"čohkat", "čohkka", "čohkkestit"
4,433	čielggadit	"čielgat", "čielggadeapmi"
4,433	báhcit	-
4,346	heaittihit	"heaitit"
4,329	čuovvolit	-
4,318	láhčit	-
4,241	doaimmahit	"doaibma", "doaibmat", "doaimmahat", "doaimmaheapme"
4,211	vázzit	"váccihit", "vázzi", "vázzin", "vázzot"
4,207	goddit	"goaddat", "goddet", "goddi", "goddin", "goddot"
4,101	heivet	"heivehit"
3,993	gieđahallat	"gieđahallan"
3,986	johtit	"johtti", "jođihit", "jođán"
3,951	plánet	"plána", "plánen"
3,928	viežžat	"viežžan", "vižžet"
3,924	čuožžut	"čuoččuhit"
3,920	bovdet	"bovden"
3,908	čuojahit	"čuodjat", "čuojadit"
3,898	ovdánit	"ovdáneapmi"
3,868	geahčadit	"geahčadeapmi", "geahčahit", "geahčču̇t"

3,838	doaivut	"doaivu"
3,837	liikot	"liikostit", "liikostuvvat", "liiku"
3,835	oassálastit	"oassálasti"
3,802	fuolahit	"fuoladit", "fuolaheapme", "fuollat"
3,756	čuoččuhit	"čuožžut"
3,744	vuođđudit	"vuođđudus"
3,738	ipmirdit	-
3,734	muitit	"muiti", "muitu"
3,671	ruhtadit	"ruhtadeapmi"
3,671	deaivvadit	"deaivat", "deaivvadeapmi", "deaivvahit"
3,598	čoavdit	"čovdot"
3,595	váilut	-
3,575	mielddisbuktit	-
3,554	gávnnahit	"gávdnat", "gávnnadit"
3,549	lohpidit	-
3,539	gártat	"gárta", "gártadit", "gárten", "gártet", "gártá"
3,515	várret	-
3,505	fárret	"fárrehit", "fárren", "fárrestallat", "fárrestit"
3,482	jorgalit	-
3,471	cegget	"ceagga", "ceggestit"
3,454	doalvut	"doalvu", "doalvun"
3,439	guođđit	"guođát", "guđđat", "guđđot"
3,409	hupmat	"humadit", "humahit", "hupma"
3,392	heaitit	"heaittihit"
3,390	váldot	"váldu"
3,384	dáidit	-
3,378	mieđihit	-
3,361	rievdat	"rievdan", "rievddadit"
3,320	hábmet	"hábmen", "hápma", "hápmi"
3,313	nohkat	"noahkut", "nohkan", "nohkkat", "nohkkot"
3,288	oidnot	"oaidnit", "oaidnu", "oidnostit"
3,284	jápmit	"jábmi", "jámet"
3,267	rahpat	"rabadit", "rahppi", "rahppot"
3,265	ovddastit	"ovddas"
3,261	hállat	"háladit", "hálla", "hállan", "hállanit", "hállet", "hálli", "hállái"
3,239	olahit	-
3,234	čuohcat	"čuozašit"
3,234	dutkat	"dutkan", "dutki"
3,167	ovdánahttit	-
3,164	báhčit	"báhčči", "bážá"
3,139	gokčat	-
3,121	jávkat	"jávkan"
3,071	ráddjet	"rádjat"
3,062	viiddidit	-
3,045	hilgut	-
2,999	lávlut	"lávlat", "lávllodit", "lávlu", "lávlun"
2,995	ovttasbargat	"ovttasbargan", "ovttasbargi", "ovttasbargu"
2,993	birget	"birgehit", "birgestit"
2,958	čatnat	"čatnan"
2,955	iskat	"iskan", "iskkadit", "iskkahit"
2,952	jienastit	"jienastat", "jienasteapmi", "jienastus"
2,944	juogadit	-
2,933	geatnegahttit	-
2,931	geavvat	"geavat", "geavvadit"
2,918	suodjalit	"suodjalus"
2,917	bissehit	"bissehat"
2,904	čiekčat	"čiekčan", "čiekči", "čiekčá", "čievččadit"
2,892	borrat	"boaru", "bora", "boradit", "borahit", "boran", "borra" "borralit", "borri"

2,889	fievrridit	-
2,876	riegádit	"riegádahttit"
2,861	bistit	-
2,836	unnidit	"unni", "unnit"
2,770	čohkkát	"čohkkádit", "čohkkáhit"
2,768	geargat	"gerget"
2,760	gozihit	-
2,741	lihkostuvvat	-
2,697	gullot	"gullat", "gullu"
2,693	gulahallat	"gulahit"
2,690	besset	"beassat", "bessen"
2,672	fuomášit	-
2,645	stivret	"stivra"
2,640	háhkat	"hága", "háhkan", "háhkka"
2,639	ollet	"olle", "ollit"
2,631	boktit	-
2,597	illudit	"illodit"
2,585	diktit	"diktet", "divttášit"
2,582	bisuhit	"bissut"
2,539	ávvudit	"ávvudeapmi"
2,536	juhkat	"juhkan", "juohkit"
2,503	eaktudit	-
2,498	bissut	"bissostit", "bissu", "bisuhit"
2,472	váidit	"váidi", "váidut"
2,448	eahpidit	-
2,437	navdit	-
2,412	vuohttit	"vuhtii", "vuhttot", "vuohttut"
2,403	hehttet	-
2,396	rihkkut	"riehkkat", "rihkku", "rihkkuhit", "rihkkun"
2,395	čuohcit	-
2,358	loktet	"loakta", "loaktit", "lokta", "loktat", "lokten", "loktestit", "loktit"
2,307	buvttadit	"buvttadeapmi"
2,279	viggat	-
2,277	ávkkástallat	-
2,261	organiseret	-
2,229	čielgat	"čielggadit", "čilgedit", "čilgehit", "čilgen", "čilget"
2,229	ságastallat	"ságastallan", "ságastit", "sáhkat"
2,224	fuobmát	-
2,221	eavttuhit	"eaktu", "eavttuheapme", "evttohus"
2,220	deaivat	"deaivan", "deaivvadit"
2,198	digaštallat	"digaštallan", "digaštit"
2,181	duostat	"duostut", "dustet"
2,175	ollašuhttit	"ollašuvvat"
2,154	ráhkkanit	-
2,149	muittuhit	"muitu"
2,111	gillát	"gilli"
2,107	geassit	"geassi", "geassut", "geasis"
2,090	gohččut	"gohčodit", "gohču", "gohčču", "gohččot", "gožu"
2,068	vuhtiiváldit	-
2,053	biehttalit	-
2,035	oahpásmuvvat	-
2,034	njuovvat	"njuovadit", "njuovahat", "njuovahit", "njuovvan", "njuovvi"
2,003	gieldit	"gildit", "gildot"
1,978	vrd	-
1,950	guldalit	"guldalas"
1,949	soahpat	"soabadit", "soabahit", "sohpat"
1,941	čoahkkanit	"čoahkkanaddat"
1,939	giddet	"gidden"

1,930	rehkenastit	-
1,916	registreret	-
1,898	fidnet	"fidnehit"
1,896	merket	"mearka", "merken"
1,896	geiget	-
1,891	divvut	"dievvat", "divodit", "divohat", "divuhit", "divvulit", "divvut"
1,890	hárjehallat	"hárjehallan", "hárjehalli", "hárjehit"
1,889	luoitit	"luoitilit"
1,874	juoigat	"juoigan", "juoiggadit", "juoigi"
1,852	movttiidahttit	"movttiidahtti"
1,842	njiedjat	"njiedja", "njiedjan"
1,829	moaitit	-
1,820	oččodit	"oažžut"
1,816	smiehttat	"smiehtadit"
1,816	guođohit	"guođoheapmi"
1,800	geahpedit	"geahpehit"
1,763	gilvalit	-
1,762	seailluhit	-
1,762	gáržžidit	"gáržžideapmi"
1,754	dulkot	"dulkon"
1,752	dovddahit	"dovdat", "dovdda", "dovddadit"
1,742	dárkkistit	"dárki"
1,732	čuožžilit	-
1,730	čadnot	-
1,723	gáhttet	"gáhtat"
1,722	miehtat	"miehta", "miehtut", "miehtá", "mieđu", "mihttu"
1,709	ovttastahttit	-
1,709	billistit	-
1,707	joavdat	"joavdit"
1,686	ođasmahttit	"ođasmahtti", "ođasmahttin", "ođasmuvvat"
1,685	hárjánit	-
1,681	eaiggáduššat	"eaiggáduššan"
1,677	čoaggit	"čoaggi", "čoakkán"
1,677	ovdanbuktit	-
1,654	rahčat	"rahča"
1,649	geasuhit	"geassut"
1,647	dovddastit	"dovdat"
1,640	áitit	"áiti"
1,629	čorget	-
1,612	náitalit	"náitaladdat"
1,569	buohtastahttit	"buohtastit"
1,567	sihtat	"sihta"
1,561	buo	-
1,522	buorránit	-
1,513	máhccat	"máhcadit", "máhcahat", "máhcahit", "máhccut"
1,510	gierdat	"gierdu", "girdit"
1,508	govvidit	-
1,488	muddet	"muddehit", "muddit"
1,488	bisánit	"bisánaddat"
1,479	giitit	"giitu"
1,474	riŋget	"riŋga", "riŋgestit"
1,468	ollit	"oallut", "olle", "ollet", "olli", "ollu"
1,467	sihkkut	-
1,466	váruhit	"várohit", "váruheapme"
1,463	čalmmustahttit	"čalmmustit"
1,459	namahit	"namahus", "namat", "namma"
1,458	ilbmat	"ilbmanit"
1,445	vuolggahit	"vuolgga"

1,442	cuiggodit	"cuoigut"
1,442	bearráigeahččat	"bearráigeahččan", "bearráigeahčči", "bearráigeahčču"
1,438	ordnet	"ordnedit", "ortnet"
1,437	maŋidit	"maŋit", "maŋŋi"
1,432	rávvet	"rávvestit"
1,426	vuollánit	-
1,420	duššat	"dušše"
1,419	sisttisdoallat	-
1,414	garvit	-
1,403	gahččat	"gahčadit", "gahčahit", "gahčat"
1,396	unnut	-
1,380	imaštit	"imaš", "imaštallat"
1,380	gilvit	"gilvu"
1,369	sárdnut	"sárdnut"
1,367	dáhkidit	-
1,357	áŋgiruššat	-
1,355	vuolláičállit	"vuolláičálli", "vuolláičállin"
1,337	bearrat	"bearaš", "berret"
1,328	juksat	"juoksat", "juoksut", "juvssat"
1,327	stuorrut	-
1,325	duddjot	-
1,324	dinet	"dinen"
1,318	imaštallat	"imaštit"
1,313	galledit	"galledeapmi", "gallehit", "gallet"
1,309	šállošit	-
1,294	eastadit	-
1,286	dubmet	-
1,281	kártet	"kárta"
1,270	vuhttot	"vuohttit"
1,267	háleštit	-
1,266	meroštallat	"meroštallan"
1,258	váillahit	-
1,246	beaggit	"beaggi", "beaggin", "beakkán"
1,244	bajidit	-
1,228	mátkkoštit	-
1,226	guohtut	"guohtun"
1,225	guoddit	"guoddi", "guoddá", "guottestuvvat"
1,224	ábuhit	"ábuheapme"
1,221	seastit	-
1,219	láigohit	"láigohat"
1,217	dikšut	"dikšu", "dikšun", "divššodit", "divššohat", "divššuhit"
1,205	beaivádit	-
1,199	oaggut	"oaggu", "oaggun"
1,199	ihtit	"ihtin", "ihttot", "ihtá", "iđistit"
1,193	luohttit	"luohttemeahttun"
1,188	guoskkahit	"guoskat", "guoskkahat"
1,182	váidalit	"váidalus"
1,181	čálihit	"čállit"
1,178	gádjut	"gádjat"
1,174	láhttet	"láhtta", "láhttestit"
1,159	gárvvistit	"gárvi", "gárvvis"
1,150	dávistit	-
1,148	nákcet	"nákca"
1,134	orrot	"orrut"
1,133	vuoittáhallat	-
1,132	gaskkustit	"gaskkusteapmi"
1,127	prioriteret	-
1,125	áimmahuššat	-

1,109	geassádit	-
1,109	dollet	"doallat", "dolla", "dollehit"
1,104	doahttalit	"doahttat"
1,103	rámidit	-
1,097	ohcalit	"ohcalas", "ohcat"
1,095	siskkildit	-
1,094	vuodjat	"vuoddji", "vuodja", "vuodjan", "vuodjit", "vuodjut", "vuojadit", "vuojahat", "vuojaš"
1,088	oamastit	"oamastallat", "oamastus"
1,086	defineret	-
1,085	vuostálastit	"vuostálasti"
1,082	bálkestit	-
1,066	bágget	"bággehit"
1,059	seaguhit	"seahkut"
1,052	fátmmastit	-
1,039	ánssášit	-
1,038	ádjánit	"ádjit"
1,036	bilidit	-
1,036	bajáššaddat	"bajáššaddan"
1,035	loaktit	"loakti", "loktat", "lokten", "loktet", "loktit", "loktut"
1,029	gullet	"gullat", "guolla"
1,028	šiehtadit	"šiehtadus", "šiehttat"
1,028	diŋgot	-
1,022	kommenteret	-
1,017	čuohppat	"čuohpadit"
1,017	šiehtadallat	"šiehtadallan", "šiehtadalli"
1,010	ollašuvvat	"ollašuhttit"
1,010	njulget	"njuolgat"
1,009	válddahallat	-
1,007	čatnasit	"čanas"
1,000	hástalit	"hástalus"
997	časkit	"časkkis", "časkkát"
996	bálvalit	-
993	mearredit	-
992	buollit	"buleš", "bulle", "bulli", "buollán"
991	deaddit	"deaddilit", "deaddu"
991	bealuštit	-
983	biebmat	"biebman", "biebmu"
982	finadit	"finahit", "fitnat"
981	justit	"juste"
981	guolástit	"guolásteapmi", "guolástus"
975	návddašit	"návddašeapmi"
969	jearahallat	"jearahallan", "jearahit"
966	vánddardit	"vánddardeapmi"
964	caggat	"caggát", "cakkadit"
962	čuoigat	"čuoigan", "čuoiggadit", "čuoiggahit", "čuoiggan", "čuoigi"
959	lonuhit	"lonohallat", "lonuhus", "lotnut"
958	girdit	"gierdat", "girddihit", "girdi", "girdilit", "girdin"
957	čuovvulit	"čuovvut"
956	vuosttaldit	-
954	joksat	-
953	njeaidit	"njeaidinviđá"
946	ráđđádallat	"ráđđidit", "ráđđádallan"
944	rahppot	"rahpat"
943	guoddalit	-
940	oskut	"osku"
940	mihtidit	-
938	jávkadit	-

929	earuhit	"earru"
928	báhtarit	"báhtaraddat"
925	sirret	"sierrat", "sirren"
923	gollat	"goallut", "golahit", "gollet", "golli", "gollát"
906	áššáskuhttit	"áššáskuvvat"
902	čuorvut	-
901	sárdnidit	-
897	goarrut	"goarru", "goarrun", "gorrat"
894	fallehit	"falleheapmi", "fallet"
891	vuostáiváldit	"vuostáiváldi", "vuostáiváldin"
888	reguleret	-
880	badjelgeahččat	-
869	buhtadit	"buhtadus"
862	bálkáhit	"bálkádit"
860	mannet	"mannat", "manne"
851	guorrasit	-
851	astat	-
850	vuvdot	"vuovdit"
848	ceavzit	-
847	veardidit	"vearditmeahttun", "veardádallat"
846	veadjit	-
844	láhppot	"láhppit", "láhppu"
839	coggat	"cokkan"
837	suoládit	"suoládeapmi"
831	ákkastallat	"ákkastit"
831	sámástit	"sámistit", "sámás"
830	nuppástuhttit	"nuppástuvvat"
829	roggat	"roggan", "rokkadit"
821	vahágahttit	-
820	dassat	"dassá", "dassái", "dat"
816	viehkat	"viegadit", "viehka", "viehkki", "vihkut"
815	nuorrat	"nuorra"
814	oasálastit	-
811	čuvget	"čuovgat"
808	ođastit	"ođas", "ođasmuvvat", "ođastus"
807	dustet	"duostat"
804	molsut	"molssodit"
804	guhkidit	"guhkit", "guhkki"
801	vuolidit	"vuolit"
801	vuohttut	"vuhttot", "vuohttit"
792	dohppet	"dohppa", "dohppestallat", "dohppestit"
789	stoahkat	"stoahka", "stoahkan"
788	badjánit	-
782	vajálduhttit	"vajálduvvat"
782	hedjonit	"headjut"
780	ráhkistit	"ráhkis"
773	vuoiŋŋastit	"vuoigŋat", "vuoiŋŋasteapmi"
771	njuiket	"njuikestit"
771	lihkkat	"liehkku", "lihkadit", "lihkahit", "lihkastit", "lihkkan", "lihkkos", "lihkkostit", "lihkku", "lihkká"
766	gillet	"giellat"
761	ráđđet	-
757	suhttat	"suhtadit"
753	boradit	"bora", "borahit", "borra", "borrat"
751	divodit	"divvut"
749	einnostit	-
744	váibat	"váibadit", "váibbat"
744	soabadit	"soahpat"

744	ovdanboahtit	-
744	logahallat	"logahallan", "logahit", "lohkat"
742	guoimmuhit	-
738	spiehkastit	"spiehkastat"
738	revideret	-
736	neaktit	-
736	gávppašit	"gávppašeapmi"
733	gudnejahttit	-
728	jorrat	"joradit", "jorralit", "jorran", "jorri"

Table A.1: The 500 most frequent North Sámi verbs in *SIKOR* and other lemmata with homonymous forms

Appendix B

Semantic prototype categories in *Giella-sme*

Table B.1: Semantic prototype categories for North Sámi nouns in *nouns.lexc*

Semantic prototype category and tag	**Members**
Sem/Act (activity)	čorgen 'cleaning', bargu 'work', hommá 'occupation', prošeakta 'project', fotosyntesa 'photosynthesis'
Sem/Amount (amount)	látna 'pile', albbasmearri 'amount of lynx', biebmohivvodat 'amount of food', vihttanuppelogátoassi 'one fifteenth', čuohteproseanta 'ten percent'
Sem/Ani (animal)	beana 'dog', boazu 'reindeer', bamse 'teddy bear', guovdi 'dragon', dihkki 'lice'
Sem/AniProd (animal product)	bivastat 'sweat', duollji 'reindeer skin', dihkimonni 'lice egg', gumpposvarra 'blood for making dumplings', gužža 'pee'
Sem/Body (body part)	beallji 'ear', dákti 'bone', bealljeráigi 'ear canal', goanstajuolgi 'artificial leg', sepmon 'mustache', nearvafierbmi 'nervous system'
Sem/Body-abstr (non-physical body part)	jierbmi 'reason', siellu 'soul', jietna 'voice', oaidnu 'eyesight', oamedovdu 'conscience'
Sem/Build (building)	viessu 'house', musea 'museum', lávvu 'Sámi tent', beassi 'nest', sáttošloahtta 'sandcastle'
Sem/Build-part (part of a building)	latnja 'room', uksa 'door', balkoŋga 'balcony', basseaŋŋa 'pool', kantuvra 'office'
Sem/Cat (category)	namma 'name', subjunkšuvdna 'subjunction', suffiksa 'suffix', beassansátni 'password', eksistentiálacealkka 'existential sentence'
Sem/Clth (clothing)	báidi 'shirt', gahpir 'hat', teáhterkostyma 'theater costume', liidni 'shawl', biilaboagán 'seatbelt', libar 'diaper'
Sem/Clth-jewl (jewelry and similar)	giehtadiibmu 'watch', beaivečalbmeláset 'sunglasses', suorpmas 'ring', čeabetbáddi 'necklace', kruvdnu 'crown'

Sem/Clth-part (part of clothes)	lubma ‘pocket’, healbmi ‘bottom part of an article of clothing’, hiitta ‘upper part of trousers’, sávdnji ‘seam’, boallu ‘button’
Sem/Ctain (container)	goaffar ‘suitcase’, terrárium ‘terrarium’, skábe ‘closet’, lihtti ‘container’, bensentáŋka ‘gas tank’
Sem/Ctain-abstr (abstract container)	foanda ‘fund’, doaibmakonto ‘account’, loatnakássa ‘loan fund’, pohttu ‘pot’, báŋkokonto ‘bank account’
Sem/Curr (currency)	euro ‘euro’, US-dollár ‘US dollar’, denára ‘denar’, dánskkakruvdna ‘Danish crown’, valuhtta ‘currency’
Sem/Dance (dance)	swinga ‘swing’, rumba ‘rumba’, baleahtta ‘ballet’, čoavjedánsa ‘belly dance’, soahtedánsa ‘war dance’
Sem/Dir (direction)	GPS-kursa ‘GPS course’, börsakursa ‘stock exchange price’, gráfa ‘graph’, tendeansa ‘tendency’, seahpebordi ‘starboard’
Sem/Domain (domain)	antropologiija ‘anthropology’, punkrohkka ‘punk rock’, biologiija ‘biology’, lingvistihkka ‘linguistics’, medisiidna ‘medicine’
Sem/Drink (drink)	deadja ‘tea’, vuolla ‘beer’, h-mielki ‘UHT milk’, bruvsa ‘soda’, girkoviidni ‘communion wine’
Sem/Dummytag (default tag for uncategorized nouns)	-
Sem/Edu (educational event)	čuoigangymnása ‘skiing high school’ skiing academy, gursa ‘course’, musihkkadiibmu ‘music lesson’, oahpahus ‘lesson’, váldofága ‘master’
Sem/Event (event)	heajat ‘wedding’, čoahkkin ‘meeting’, gilvu ‘competition’, válga ‘election’, festivála ‘festival’
Sem/Feat (feature)	ahkeerohus ‘age difference’, homoseksualiteahtta ‘homosexuality’, feminitehta ‘femininity’, identitehta ‘identity’, kongrueansa ‘congruence’
Sem/Feat-phys (physical feature)	sturrodat ‘size’, ivdni ‘color’, allodat ‘height’, hápmi ‘shape’, deaddu ‘weight’, heastafápmu ‘horsepower’
Sem/Feat-measr (measurable feature)	rádius ‘radius’, diamehter ‘diameter’, voluma ‘volume’, birramihttu ‘circumference, perimeter’, frekveansa ‘frequency’
Sem/Feat-psych (psychological feature)	autoriteahtta ‘authority’, luondu ‘nature’, mánnálašvuohta ‘childishness’, kreativiteahtta ‘creativity’, čeavlivuohta ‘arrogance’
Sem/Fem (female names)	Márjá, Maria, Fátima, Iŋgá, Kátjá
Sem/Food (food)	láibi ‘bread’, vegetárabiebmu ‘vegetarian food’, jáffu ‘flour’, duhpát ‘tobacco’, sálti ‘salt’
Sem/Food-med (medicine)	p-pilla ‘birth-control pill’, ástmádálkkas ‘asthma medicine’, medisiidna ‘medicine’, penicilliidna ‘penicillin’, vaksiidna ‘vaccine’
Sem/Furn (furniture)	truvdnu ‘throne’, stuollu ‘chair’, beavdi ‘table’, áltár ‘altar’, trampoliidna ‘trampoline’
Sem/Game (game)	biŋgo ‘bingo’, tv-speallu ‘TV game’, flipper ‘flipper’, paintball ‘paintball’, šáhkka ‘chess’

Sem/Geom (geometrical object)	golbmačiehka ‘triangle’, 3-čiegahas ‘triangle’, tetraedar ‘tetrahedron’, asymtohta ‘asymptote’, násti ‘star’
Sem/Group (group)	bearaš ‘family’, eallu ‘herd’, joavku ‘group’, eamiálbmot ‘indigenous people’, delegašuvdna ‘delegation’
Sem/Ideol (ideology)	nomadisma ‘nomadism’, buddhisma ‘buddhism’, feminisma ‘feminism’, kristtalašvuohta ‘christianity’, fanatisma ‘fanaticism’
Sem/Lang (language)	lullisámegiella ‘South Sámi’, eatnigiella ‘mother tongue’, maori ‘Maori’, jiddisch ‘Yiddish’, nubbigiella ‘second language’
Sem/Mal (male name)	Áilu, Jesus, Máhtte, Åge, Adam
Sem/Mat (material)	bábir ‘paper’, stálli ‘steel’, muorra ‘wood’, náhkki ‘leather’, ullu ‘wool’
Sem/Measr (measure)	geassoovttadat ‘unit of volume’, njealjádasmettar ‘quarter meter’, diibmu ‘hour’, buolašgráda ‘minus degree’, wátta ‘watt’
Sem/Money (money)	dávvir ‘treasure, belongings’, vealgi ‘debt’, biebmohaddi ‘food price’, rehket ‘bill’, penšuvdna ‘pension’
Sem/Obj (concrete object)	diŋga ‘thing’, pokála ‘cup’, dávvir ‘thing’, duhkoras ‘toy’, maleriija ‘painting’
Sem/Obj-clo (cloth object)	rátnu ‘carpet’, leavga ‘flag’, glássaliidni ‘curtain’, silkegávdni ‘silk sheets’, servieahtta ‘napkin’
Sem/Obj-el (electrical object)	čuojanas ‘player’, lámpá ‘lamp’, TV ‘TV’, rádioapparáhta ‘radio’, uvdna ‘oven’
Sem/Obj-rope (rope-like object)	biikasreaŋga ‘barbed wire’, árpu ‘thread’, báddi ‘rope’, jođas ‘cable’, batninárpu ‘dental floss’
Sem/Obj-surfc (surface object)	távval ‘blackboard’, tevdnenbábir ‘drawing paper’, lerret ‘canvas’, speallanbreahtta ‘board (for playing board games)’, seđel ‘(money) bill’
Sem/Org (organization)	áviisa ‘newspaper’, alimusriekti ‘supreme court’, fitnodat ‘company’, musea ‘museum’, administrašuvdna ‘administration’
Sem/Part (part of something)	bealli ‘half’, oassi ‘part’, proseanta ‘percent’, reasta ‘rest’, logádas ‘tenth’
Sem/Perc-emo (emotional perception)	ballu ‘fear’, identitehtadovdu ‘feeling of identity’, empatiija ‘empathy’, moraš ‘sadness’, barganmiella ‘working motivation’
Sem/Perc-phys (physical perception)	oalgebávčča ‘shoulder pain’, bensiidnahádja ‘gass smell’, idjanagir ‘sleep (during the night)’, oađđindárbu ‘need of sleep’, nealgi ‘hunger’
Sem/Plant (plant)	jeagil ‘lichen’, šaddu ‘plant’, agurka ‘cucumber’, alitbiellorássi ‘bluebell’, mirkoguoppar ‘poisonous mushroom’
Sem Plant-part (part of a plant)	rissi ‘twig’, lasta ‘leaf’, ruohtas ‘root’, siepman ‘seed’, beahcemátta ‘pine trunk’
Sem/Plc (place)	máilbmi ‘world’, luondu ‘nature’, girdišillju ‘airport’, bargosadji ‘workplace’, árran ‘fireplace’

Sem/Plc-abstr (abstract place)	bachelordássi ‘bachelor level’, bargomárkan ‘job market’, bronsasadji ‘third place’, Troms-neahttasadji ‘Troms website’, čujuhus ‘address’
Sem/Plc-elevate (elevated place)	várri ‘mountain’, gáisi ‘peak’, čohkka ‘mountain top’, juovva ‘scree’, vulkána ‘volcano’
Sem/Plc-line (place limitations)	riikkarádjá ‘national border’, rádjá ‘border’, moallasáhcu ‘finish line’, bissánansáhcu ‘stop line’, ekváhtor ‘equator’
Sem/Plc-water (water)	johka ‘river’, jávri ‘lake’, jiekŋaáhpi ‘polar sea’, mearra ‘sea’, ája ‘well’
Sem/Pos (position)	beallevirgi ‘50% position’, presideantasadji ‘presidency’, fástabargu ‘fixed position’, mánaidgárdesadji ‘kindergarten place’, servodatrolla ‘role in society’
Sem/Prod-audio (audible product)	luohti ‘yoik’, jupma ‘roar’, Beatles-lávlla ‘Beatles song’, biibalsálbma ‘Bible psalm’, blues ‘blues’
Sem/Prod-cogn (product of a cognition)	jurdda ‘thought’, mearrádus ‘decision’, máhttu ‘knowledge’, eahpeipmárdus ‘lack of understanding’, gáibádus ‘requirement’
Sem/Prod-ling (linguistic product)	dieđáhus ‘message’, gažaldat ‘question’, šiehtadus ‘agreement’, jorgalus ‘translation’, kritihkka ‘criticism’
Sem/Prod-vis (visual product)	govva ‘picture’, ealligovva ‘film’, TV-ráidu ‘TV series’, dokumentára ‘documentary’, dáidda ‘art’
Sem/Rel (relation)	oktavuohta ‘relation’, dependeansa ‘dependency’, subordinašuvdna ‘subordination’, analogiija ‘analogy’, ekvivaleansa ‘equivalence’
Sem/Route (route-like place)	geaidnu ‘street’, bálggis ‘path’, feaskkir ‘corridor’, šaldi ‘bridge’, doalli ‘winter path’
Sem/Rule (rule)	kulturárbevierru ‘cultural tradition’, abortaláhka ‘abortion law’, EU-njuolggadus ‘EU rule’, cosinusláhka ‘law of cosines’, fair play ‘fair play’
Sem/Semcon (abstract semantic concept)	boađus ‘result’, ulbmil ‘objective’, sivva ‘reason’, hearbevárri ‘alternative’, sáhka ‘case’
Sem/Sign (sign)	ID-nummar ‘ID number’, ČSV-bustávat ‘ČSV letters’, aisttonmearka ‘quotation mark’, hieroglyfa ‘hieroglyph’, symbola ‘symbol’
Sem/Sport (sport)	beavdetennis ‘table tennis’, judo ‘judo’, muohtaskohtercrossa ‘motor cross’, jiekŋahockey ‘ice hockey’, sisbandy ‘floorball’
Sem/State (state)	hoahppu ‘hurry’, fáŋgavuohta ‘captivity’, anarkiija ‘anarchy’, biodiversitehta ‘biodiversity’, moivi ‘chaos’
Sem/State-sick (illness)	allergiija ‘allergy’, nuorvu ‘cold’, autisma ‘autism’, demetiija ‘dementia’, somnambuilsma ‘somnambulism’
Sem/Substnc (substance)	sáttu ‘sand’, áibmu ‘air’, suovva ‘smoke’, karbohydráhta ‘carbohydrate’, vitamiidna ‘vitamin’, gavja ‘dust’
Sem/Sur (surname)	Gaup, Eira, Johansson, Hill, García
Sem/Time (time)	áigi ‘time’, cuoŋománnu ‘April’, diibmobealli ‘half an hour’, disdat ‘Tuesday’, áigemearri ‘deadline’

Sem/Tool (prototypical tool)	ákšu ‘axe’, niibi ‘knife’, dollaruovdi ‘fire striker’, plastihkkaveažir ‘plastic hammer’, skruvenčoavdda ‘wrench’
Sem/Tool-catch (tool for catching)	dolgevuogga ‘artificial fly’, dorskefierbmi ‘fishing net for cod’, bivdostággu ‘fishing rod’, buolašsuohpan ‘lasso used in wintertime’, sáhpándoalli ‘mouse trap’
Sem/Tool-clean (tool for cleaning)	suohpal ‘broom’, ruonasboršta ‘vegetable brush’, bátnegusta ‘toothbrush’, buhtistanrusttet ‘cleaning equipment’
Sem/Tool-it (tool within IT)	IT-infrastruktuvra ‘IT infrastructure’, analysáhtor ‘analyzer’, ohcanfunkšuvdna ‘searching function’, dihtorprográmma ‘computer program’, neahttalohkki ‘browser’
Sem/Tool-measr (tool for measuring)	baromehter ‘barometer’, tiibmoláse ‘hourglass’, linjála ‘ruler’, váđir ‘spirit level’, breavaviehkat ‘scale’
Sem/Tool-music (musical instrument)	noaiderumbu ‘shaman drum’, gitárra ‘guitar’, fioliidna ‘violin’, musihkkainstrumeanta ‘musical instrument’, njálbmehárpa ‘jaw harp’
Sem/Tool-write (writing tool)	ivdnenpeanná ‘colored pen’, bliánta ‘pencil’, kriita ‘chalk’, málenkusta ‘paintbrush’, ivdni ‘paint’
Sem/Txt (written document)	bábir ‘paper’, girji ‘book’, reive ‘letter’, e-mail ‘e-mail’, lávlla ‘song’
Sem/Veh (vehicle)	biila ‘car’, fanas ‘boat’, gielká ‘sled’, vuoján ‘vehicle, draft reindeer’, sihkkel ‘bicycle’
Sem/Wpn (weapon)	bissu ‘rifle’, juoksa ‘bow’, miehkki ‘sword’, njuolla ‘arrow’, soahteákšu ‘war axe’
Sem/Wthr (weather condition)	balvadálki ‘cloudy weather’, biegga ‘wind’, vuodjinsiivu ‘driving conditions’, idjabeaivvádat ‘night sunlight’, arveoakti ‘rain shower’

Appendix C

Grammatical tags in *grammarchecker.cg3*

C.1 Parts of speech and their subcategories

(Subcategories of parts of speech are indented)

A = adjective
    Ord = ordinal
Adv = adverb
CC = conjunction
CS = subjunction
Interj = interjection
N = noun
    NomAg = agent noun
    Prop = proper noun
Num = numeral
    Coll = collective numeral
Pcle = particle
    Qst = question particle
Po = postposition
Pr = preposition
Pron = pronoun
    Dem = demonstrative
    Indef = indefinite
    Interr = interrogative
    Pers = personal
    Recipr = reciprocal
    Refl = reflexive
    Rel = relative
V = verb
    IV = intransitive verb
    TV = transitive verb
    <vdic> = verba dicendi
    <mv> = main verb
    <aux> = auxiliary
    <copula> = copula

ABBR = abbreviation
ACR = acronym

C.2 Morpho-syntactic properties

Case:
Acc = accusative
Com = comitative
Ess = essive
Gen = genitive
Ill = illative
Loc = locative
Nom = nominative

Number:
Du = dual
Pl = plural
Sg = singular

Compounding potential:
RCmpnd = hyphenated compound
SgNomCmp = compound with the first part in nominative case

SgGenCmp = compound with the first part in genitive case
CmpN/SgN = compound with the first part in nominative singular
CmpN/SgG = compound with the first part in genitive singular
CmpN/PlG = compound with the first part in genitive plural

Possessive inflection:
PxSg1 = first person singular possessive
PxSg2 = second person singular possessive
PxSg3 = third person singular possessive
PxDu1 = first person dual possessive
PxDu2 = second person dual possessive
PxDu3 = third person dual possessive
PxPl1 = first person plural possessive
PxPl2 = second person plural possessive
PxPl3 = third person plural possessive

Adjective inflection:
Comp = comparative form
Superl = superlative form
Attr = attributive form

Focus clitics:
Foc/ge
Foc/gen
Foc/ges
Foc/gis
Foc/naj
Foc/ba
Foc/be
Foc/hal
Foc/han
Foc/bat
Foc/son

Tense:
Prt = past tense
Prs = present tense

Mode:
Ind = indicative
Pot = potential
Cond = conditional
Imprt = imperative
ImprtII = biblical imperative

Person:
Sg1 = first person singular
Sg2 = second person singular
Sg3 = third person singular
Du1 = first person dual
Du2 = second person dual
Du3 = third person dual
Pl1 = first person plural
Pl2 = second person plural
Pl3 = third person plural

Non-finite verb forms:
Actio = actio form
ConNeg = connegative form
ConNegII = biblical connegative form
Ger = gerund
Inf = infinitive
Neg = negation
PrfPrc = past participle
PrsPrc = present participle
Sup = supinum
VGen = verb genitive
VAbess = verb abessive

Morpho-phonological properties:
G3 = geminate grade three in consonant gradation
South = southern dialect form

Spelling errors:
Err/Orth = undefined orthographical error
Err/Orth-a-á = accent error
Err/Orth-nom-acc = case error (nominative should be accusative)
Err/Orth-nom-gen = case error (nominative should be genitive)

C.3 Derivational tags

Tags with multiple functions and suffixes are only listed and not explained.

A* = derivation from an adjective
N* = derivation from a noun
IV* = derivation from an intransitive verb
TV* = derivation from a transitive verb

V* = derivation from a verb
Der/Caus = causative derivation
Der/Dimin = diminutive derivation
Der/NomAct = action noun derivation
Der/PassL = long passive derivation
Der/PassS = short passive derivation

Der/adda
Der/ahtti
Der/alla
Der/asti
Der/easti
Der/d
Der/diibmosaš
Der/duohke
Der/duohkai
Der/eaddji
Der/eamoš
Der/amoš
Der/geahtes
Der/gielat
Der/goahti
Der/h
Der/heapmi
Der/hudda
Der/huhtti
Der/huvva
Der/halla
Der/j
Der/jagáš
Der/jahkásaš
Der/l
Der/lágan
Der/lágán
Der/lágaš
Der/laš
Der/las
Der/hat
Der/meahttun
Der/muš
Der/st
Der/stuvva
Der/upmi
Der/supmi
Der/vuohta
Der/viđá
Der/viđi
Der/veara
Der/vuolle
Der/vuollai
Der/vuolde

C.4 Syntactic tags

@+FAUXV = finite auxiliary
@+FMAINV = finite main verb
@-FADVL> = non-finite adverbial to the left of its governor
@-F<ADVL = non-finite adverbial to the right of its governor
@-FAUXV = non-finite auxiliary
@-FMAINV = non-finite main verb
@-FOBJ> = non-finite object to the left of its governor
@-F<OBJ = non-finite object to the right of its governor
@-FOPRED> = non-finite object predicative to the left of its governor
@-F<OPRED = non-finite object predicative to the right of its governor
@-FSUBJ> = non-finite subject to the left of its governor
@-FSPRED> = non-finite subject predicative to the left of its governor
@-F<SPRED = non-finite subject predicative to the right of its governor
@>A = pre-adjectival modifier
@ADVL = any adverbial
@ADVL< = right-hand modifier of an adverbial
@>ADVL = left-hand modifier of an adverbial
@ADVL> = adverbial to the left of its governor
@<ADVL = adverbial to the right of its governor
@APP = apposition
@APP-ADVL< = apposition of an adverbial
@APP-N< = apposition of a noun
@APP-Num< = apposition of a numeral
@APP-Pron< = apposition to the right of a pronoun
@APP>Pron = apposition to the left of a pronoun

@CNP = noun phrase conjunction
@COMP-CS< = complement of a subjunction
@CVP = verb phrase conjunction
@HAB = habitive
@HNOUN = head noun
@INTERJ = interjection
@>N = pre-nominal modifier
@N< = post-nominal modifier
@Num< = post-numeral modifier
@>Num = pre-numeral modifier
@OBJ = object
@<OBJ = object to the right of its governor
@OBJ> = object to the left of its governor
@OPRED = object predicative
@<OPRED = object predicative to the right of its governor
@OPRED> = object predicative to the left of its governor
@P< = post-adpositional modifier
@>P = pre-adpositional modifier
@PCLE = particle
@Pron< = post-pronominal modifier
@>Pron = pre-pronominal modifier
@PPRED = any predicative of a predicative
@SPRED = subject predicative
@<SPRED = subject predicative to the right of its governor
@SPRED> = subject predicative to the left of its governor
@<PPRED = predicative of a predicative to the right of its governor
@SUBJ = subject
@<SUBJ = subject to the right of its governor
@SUBJ> = subject to the left of its governor
@VOC = vocative
@X = default tag

FAUXV = any auxiliary (finite or non-finite)
FMAINV = any main verb (finite or non-finite)
FOBJ = any non-finite object
<OBJ = any right-handed object
OBJ> = any left-handed object
OPRED = any object predicative
SPRED = any subject predicative
SUBJ = any subject
<ctjHead> = head in coordination

C.5 Semantic role tags

Arguments:
§AG = agent
§AT = attribute
§BE = beneficiary
§CO = co-argument
§DE = destination
§EX = experiencer
§ID = identity
§IN = instrument
§LO = location
§MA = manner
§OR = origin
§PA = patient
§PO = possessor
§PR = product
§PU = purpose
§PT = path
§PV = partitive
§RF = referent
§RE = recipient
§RO = role
§RS = reason
§SO = source
§TH = theme
§XT = extent

§ANYROLE = any semantic role

Adjuncts:
§MANNER-ADJUNCT = manner adjunct
§PART = part
§TIME-ADJUNCT = time adjunct

§ANYADJUNCT = any adjunct

C.6 Valency tags

<0>
<Acc><TH-Inf>
<AktioEss>
<AG-Acc-Ani><TH-Inf>
<AG-Acc-Ani>
<AG-Ill-Ani><PR-Acc-Any>
<AG-Ill-Ani>
<AG-Ill-Any>
<AG-Loc-Any>
<AG-Nom-Abs><TH-Ill-Abs>
<AG-Nom-Abs><TH-Ill-Plc>
<AG-Nom-Ani>
<AG-Nom-Any>
<AT-Abe-Any>
<AT-Ess-Any>
<AT-Loc-Mat>
<AT-Nom-Any>
<AT-Nom-Adj><EX-Ill-Ani>
<atnui>
<badjel>
<bajás>
<BE-Acc-Ani>
<BE-Acc-Ani><RO-Ess-Any>
<BE-Acc-Ani><TH-Com-*Ani>
<BE-Acc-Ani><TH-Ill-*Ani>
<BE-Acc-Ani><TH-Inf>
<BE-Acc-Ani><TH-Loc-Any>
<BE-Acc-Ani><PU-Ill-*Ani>
<BE-Acc-Any><PU-Inf>
<BE-Acc-Any><vuostá>
<BE-Acc-Any>
<BE-Acc-Any><TH-AktioLoc>
<BE-Acc-Hum><LO-Loc-Pos>
<BE-Ill-Ani>
<BE-Ill-Ani><veahkkin>
<BE-Ill-Any>
<BE-ovddas-Ani>
<BE-ovdii-Ani>
<birra>
<CO-0>
<CO-Acc-Ani>
<CO-Com-Ani>
<CO-Com-Ani><TH-Loc-Any>
<CO-Com-Hum>
<CO-haga-Any>
<CO-Ill-Any>
<CO-mielde-Ani>
<CO-vuostá-Any>
<DE-0>
<DE-Ill-Any>
<DE-Ill-*Ani>
<DE-Ill-Plc>
<DE-Ill-Plc><PU-Inf>
<DE-Ill-Time>
<DE-lusa-Ani>
<DE-sisa-Build>
<eret>
<eret><AktioLoc>
<eret><RF-Loc-Any>
<EX-Acc-Ani><TH-Inf>
<EX-Acc-Any>
<EX-Com-Any>
<EX-Ill-Ani>
<EX-Ill-Ani><TH-Nom-Adj>
<EX-Loc-Any>
<EX-Nom-Ani>
<EX-Nom-Any>
<EX-Nom-Time>
<fárrui>
<gitta>
<guossái>
<guossis>
<heaggabeallái>
<iežas>
<ID-Nom-Any>
<IN-0>
<IN-Acc-Any>
<IN-Acc-Any><MA-Ess-Any>
<IN-Acc-Any><PU-Ill-Any>
<IN-Acc-Lang>
<IN-Acc-Veh>
<IN-Acc-Veh><DE-Ill-Any>
<IN-Acc-Veh><SO-Loc-Any><DE-Ill-Any>
<IN-Com-Any>
<IN-Com-Veh>
<IN-Ill-Lang>
<Inf>
<jámas>
<johtui>
<johtui><DE-Ill-Plc>
<LO-0>
<LO-Acc-Plc>
<LO-Acc-Time>

<LO-Adv-Time>
<LO-ala-Plc>
<LO-Com-Ani>
<LO-Ill-Any>
<LO-Ill-Body>
<LO-Ill-Plc>
<LO-Ill-Time>
<LO-Loc-Any>
<LO-Loc-Any><gitta>
<LO-Loc-Any><guossis>
<LO-Loc-johtu><DE-Ill-Plc>
<LO-Loc-Plc>
<LO-Loc-Time>
<LO-luhtte-Any>
<LO-maŋŋil-Time>
<LO-Nom-Any><TH-Acc-Any>
<MA-Adv-Manner>
<MA-Com-*Plc>
<MA-Com-Any>
<MA-Ess-Adj>
<mátkái><DE-Ill-Plc>
<mielde>
<oktii>
<oktii><RF-Com-Any>
<olggos>
<OR-Loc-Any>
<OR-eret-Plc>
<OR-Loc-HumGroup>
<OR-Loc-Mat><PR-Nom-Any>
<ovttas>
<ovttas><CO-Com-Ani>
<PA-0>
<PA-Acc-Ani>
<PA-Acc-*Ani><BE-Ill-Ani>
<PA-Acc-Ani><LO-Ill-Body>
<PA-Acc-Ani><LO-Loc-Body>
<PA-Acc-Any>
<PA-Acc-Any><atnui>
<PA-Acc-Any><CO-gaskka-Any>
<PA-Acc-Any><DE-Ill-Any>
<PA-Acc-Any><gitta>
<PA-Acc-Any><eret>
<PA-Acc-Any><IN-Com-*Ani>
<PA-Acc-Any><LO-Ill-Any>
<PA-Acc-Any><LO-Loc-Any>
<PA-Acc-Any><LO-birra-Any>
<PA-Acc-Any><PA-Com-Any>
<PA-Acc-Any><PR-Ess-Any>
<PA-Acc-Any><PR-Ill-*Ani>
<PA-Acc-Any><PU-Ill-Any>
<PA-Acc-Any><PU-Inf>
<PA-Acc-Any><ráiggil>
<PA-Acc-Any><RE-Loc-Ani>
<PA-Acc-Any><RF-Ill-Any>
<PA-Acc-Any><RF-Loc-Any>
<PA-Acc-Any><RF-ektui-Any>
<PA-Acc-Any><RF-mielde-Any>
<PA-Acc-Any><RF-vuostá-Any>
<PA-Acc-Any><RF-vuođul-Any>
<PA-Acc-Any><RF-váste-Any>
<PA-Acc-Any><RO-Ess-Any>
<PA-Acc-Any><SO-Loc-Any>
<PA-Acc-Any><SO-Loc-Any><DE-Ill-Any>
<PA-Acc-Any><TH-Com-Any>
<PA-Acc-Ani><TH-Ill-Any>
<PA-Acc-Ani><TH-Inf>
<PA-Acc-Any><XT-Com-Measure>
<PA-Acc-BessetN>
<PA-Acc-Body>
<PA-Acc-boktitN>
<PA-Acc-Food>
<PA-Acc-Hum>
<PA-Acc-Hum><LO-Ill-Plc>
<PA-Acc-ieš><LO-Ill-Any>
<PA-Acc-Substnc>
<PA-Acc-Veh>
<PA-Com-Any>
<PA-gaskkas-Any>
<PA-Ill-*Ani>
<PA-Ill-Ani>
<PA-Ill-Ani><LO-Ill-Body>
<PA-Ill-Ani><TH-Acc-*Ani>
<PA-Ill-Ani><TH-Inf>
<PA-Ill-Any>
<PA-Loc-Ani><LO-Acc-Body>
<PA-Loc-Food>
<PA-Nom-Any>
<PO-Gen-Hum>
<PR-0>
<PR-Acc-Any>
<PR-Acc-Any><bajás>
<PR-Acc-Any><BE-Ill-Any>
<PR-Acc-Any><LO-Ill-*Ani>
<PR-Acc-Any><MA-Ess-Any>
<PR-Acc-Any><MA-Acc-Adj>

<PR-Acc-Any><OR-Loc-Mat>
<PR-Acc-Any><OR-Loc-Any>
<PR-Ess-Any>
<PR-Ess-Any><BE-Ill-Ani>
<PR-Ill-Any>
<PR-Nom-Any>
<PT-Gen-Plc>
<PT-Gen-Plc><DE-Ill-Any>
<PT-bokte-Plc>
<PT-meaddel-Plc>
<PT-čađa-Plc>
<PT-rastá-Plc>
<PU-AktioEss>
<PU-Ill-*Ani>
<PU-Inf>
<PV-Loc-Any>
<rabas>
<rasta>
<RE-Acc-Ani>
<RE-Acc-Ani><TH-Loc-Any>
<RE-Acc-Ani><TH-ahte>
<RE-Acc-Ani><TH-Inf>
<RE-Acc-Ani><TO-Ill-Any>
<RE-Com-Ani>
<RE-Com-ieš>
<RE-Ill-Ani>
<RE-Ill-Any><TH-Acc-Any><namman>
<RE-Ill-Ani><TH-Acc-*Ani>
<RE-Ill-Ani><TH-FS>
<RE-Ill-Ani><TH-ahte>
<RE-Ill-ieš>
<RE-Loc-Ani>
<RE-Loc-Ani><TH-ahte>
<RF-Ill-Any>
<RF-Loc-Any>
<RO-Ess-Any>
<RO-Ess-Any><PU-Ill-Act>
<RO-Ill-Any>
<RS-Acc-Reason>
<RS-Acc-*Ani>
<RS-alde-Any>
<RS-Com-Any>
<RS-Com-Clth>
<RS-Com-Impers>
<RS-dihte-Any>
<RS-geažil-Any>
<RS-go>
<RS-Ill-Any>
<RS-Loc-Any>
<sisa>
<SO-Loc-Any>
<SO-Loc-Time>
<SO-Loc-Plc>
<SO-Loc-*Ani>
<SO-Loc-*Ani><DE-Ill-*Ani>
<SO-Loc-Any><DE-Ill-Any>
<SO-Loc-Lang><DE-Ill-Lang>
<SO-Loc-Time><DE-Ill-Time>
<SO-luhtte-Ani>
<TH-0>
<TH-Acc-Ani>
<TH-Acc-Ani><DE-Ill-*Ani>
<TH-Acc-*Ani>
<TH-Acc-*Ani><BE-Ill-Ani>
<TH-Acc-*Ani><BE-Loc-Ani>
<TH-Acc-*Ani><RE-Loc-Ani>
<TH-Acc-*Plc>
<TH-Acc-Any>
<TH-Acc-Any><XT-Ill-Money>
<TH-Acc-Any><XT-Loc-Money>
<TH-Acc-Any><EX-Loc-Ani>
<TH-Acc-Any><IN-Com-Veh>
<TH-Acc-Any><ala>
<TH-Acc-Any><fárrui>
<TH-Acc-Any><gitta>
<TH-Acc-Any><CO-Com-Ani>
<TH-Acc-Any><CO-vuostá-Any>
<TH-Acc-Any><DE-Ill-Any>
<TH-Acc-Any><DE-Ill-*Ani>
<TH-Acc-Any><DE-Ill-Lang>
<TH-Acc-Any><DE-Ill-Time>
<TH-Acc-Any><EX-Ill-Any>
<TH-Acc-Any><IN-Com-*Ani>
<TH-Acc-Any><IN-Com-Money>
<TH-Acc-Any><IN-bokte-Money>
<TH-Acc-Any><LO-Ill-*Ani>
<TH-Acc-Any><LO-Ill-WPlc>
<TH-Acc-Any><LO-Loc-Any>
<TH-Acc-Any><LO-ala-Any>
<TH-Acc-Any><MA-Ess-Any>
<TH-Acc-Any><MA-Ill-áigi>
<TH-Acc-Any><MA-Ill-háldu>
<TH-Acc-Any><OR-Loc-Any>
<TH-Acc-Any><OR-Loc-Any><RE-Ill-Any>
<TH-Acc-Any><PU-Ill-Any>

<TH-Acc-Any><PU-Inf>
<TH-Acc-Any><RE-Com-Ani>
<TH-Acc-Any><RE-Ill-Any>
<TH-Acc-Any><RE-Loc-Ani>
<TH-Acc-Any><RE-Loc-Any>
<TH-Acc-Any><RF-Com-Any>
<TH-Acc-Any><RF-Loc-*Plc>
<TH-Acc-Any><RF-vuostá-Any>
<TH-Acc-Any><RF-vuostái-Any>
<TH-Acc-Any><RO-Ess-Any>
<TH-Acc-Any><RO-Ess-Adj>
<TH-Acc-Any><RO-Ill-Any>
<TH-Acc-Any><RS-Loc-Any>
<TH-Acc-Any><RS-ovddas-Any>
<TH-Acc-Any><SO-Loc-Any>
<TH-Acc-Any><SO-Loc-Any><DE-Ill-Any>
<TH-Acc-Any><SO-Loc-Lang>
<TH-Acc-Any><SO-Loc-Lang><DE-Ill-Lang>
<TH-Acc-Any><TH-AktioEss>
<TH-Acc-Any><TH-PrfPrc>
<TH-Acc-Any><TH-Com-Any>
<TH-Acc-Any><TH-Inf>
<TH-Acc-Any><TH-Loc-Any>
<TH-Acc-Any><TO-Ill-Any>
<TH-Acc-Any><TO-Loc-Any>
<TH-Acc-Any><XT-Com-Measure>
<TH-Acc-Any><XT-Ill-Freq>
<TH-Acc-Any><árvvus>
<TH-Acc-Any><badjelii>
<TH-Acc-Any><bajás>
<TH-Acc-Any><doibmii>
<TH-Acc-Any><eret>
<TH-Acc-Any><fápmui>
<TH-Acc-Any><johtui>
<TH-Acc-Any><mátkái>
<TH-Acc-Any><mielde>
<TH-Acc-Any><oktii>
<TH-Acc-Any><ovdan>
<TH-Acc-Any><sisa>
<TH-Acc-Body>
<TH-Acc-Clth><ala>
<TH-Acc-Obj>
<TH-Acc-Dance>
<TH-Acc-Edu>
<TH-Acc-Elect><ala>
<TH-Acc-Hum><eret>
<TH-Acc-Impers>
<TH-Acc-Money>
<TH-Acc-Obj><CO-Com-Ani>
<TH-Acc-Obj><DE-DePp-Any>
<TH-Acc-Obj><XT-Acc-Measure>
<TH-Acc-Txt>
<TH-Acc-Txt><LO-Ill-Txt>
<TH-Acc-vuođđu><LO-Loc-Any>
<TH-ahte>
<TH-ahte><RE-Ill-Any>
<TH-ahte><ovdan>
<TH-ala-*Plc>
<TH-alde-Any>
<TH-AktioEss>
<TH-AktioCom>
<TH-AktioLoc>
<TH-AktioLoc><RF-Loc-Any>
<TH-badjel-Ani>
<TH-badjel-Any>
<TH-beale-Any>
<TH-birra-Any>
<TH-birra-Any><CO-Com-Ani>
<TH-birra-Any><RE-Com-Ani>
<TH-birra-Any><RE-Acc-Any>
<TH-Com-*Ani>
<TH-Com-Any>
<TH-Com-Impers>
<TH-Ess-Ani>
<TH-Ess-Wthr>
<TH-FS>
<TH-FS-Qst>
<TH-FS-Qpron>
<TH-gaskkas-Any>
<TH-haga-Any>
<TH-go>
<TH-hárrái-Any>
<TH-Ill-Any>
<TH-Ill-Obj>
<TH-Ill-*Plc>
<TH-Inf>
<TH-Inf><RE-Ill-Any>
<TH-jus>
<TH-Loc-Any>
<TH-Loc-Ani><RS-Acc-*Ani>
<TH-Loc-Concept>
<TH-Loc-Event>
<TH-Loc-Plc>
<TH-lusa-Any>

<TH-maŋis-Ani>
<TH-maŋŋái-*Plc>
<TH-maŋŋái-Any>
<TH-Nom-*Ani><MA-Adv-Manner>
<TH-Nom-*Ani><PR-Ess-Any>
<TH-Nom-Any>
<TH-Nom-Any><AG-Ill-Any>
<TH-Nom-Any><RO-Ess-Any><EX-Ill-Any>
<TH-Nom-Any><PO-Ill-Any>
<TH-Nom-Any><XT-Acc-Measure>
<TH-Nom-Any><XT-Acc-Money>
<TH-Nom-Obj><RE-Ill-Ani>
<TH-Nom-Time>
<TH-ovddas-Any>
<TH-ovddas-Any><RE-Ill-Any>
<TH-PrfPrc>
<TH-vearu>
<TH-vuostá-Ani>
<TO-0>
<TO-Acc-*Ani>
<TO-Acc-*Ani><RE-Ill-Ani>
<TO-Acc-*Ani><RE-Loc-Ani>
<TO-Acc-Any>
<TO-Acc-Any><RO-Ess-Any>
<TO-ahte>
<TO-ahte><RE-Ill-Any>
<TO-badjel-Any>
<TO-beale-Any>
<TO-birra-Any>
<TO-go>
<TO-Inf>
<TO-Loc-Any>
<TO-vuostái-Any>
<vuhtii>
<vuostá>
<verb-part>
<XT-Acc-Measure>
<XT-Acc-Money><RE-Ill-Any>
<XT-Acc-Money><TH-ovddas-Any>
<XT-Acc-Money><TH-ovddas-Any><RE-Ill-Any>
<XT-Acc-Time>
<XT-Com-Measure>
<XT-Com-Money>
<XT-Com-Time>
<XT-Gen-Measr>
<XT-Ill-Money>
<XT-Ill-Freq>
<XT-Loc-Money>