Trond Trosterud

Normative language work in the age of machine learning

Abstract

Neural nets have, during the last few years, given us both an improved Google Translate, better search algorithms, better speech technology and doubtless many other things. The approach dominates current language technology to the extent that no other approach is visible. Being data driven, the hidden assumption behind this approach when used in proofing tools is that the language is used correctly in the text material, in other words, *usage equals the norm*. Although this approach is able to provide useful help for the largest languages, it leads to some serious problems. For indigenous and often also for other minority languages, the assumption does not hold. The written norm is weakly established and cannot be reliably found in usage. For normative bodies responsible for defining the written norm of a given language, usage-based proofing tools will not be able to implement the explicit norm they have defined. The present article discusses the current trend within proofing tools and looks at some alternatives.

1. Introduction

When politicians ask, language technologists answer that all they need is more data, i.e. they need a Language Bank. When constructing language tools, their preferred method is the one that **trains** the computer. The use of AI within the field of planning and implementing written norms thus increasingly equates to adding more text to the tool and hoping for the best.

This works for language societies where there is much text available, the language does not have dynamic compounding and correct forms clearly outnumber incorrect ones. However, for most languages, these assumptions do not hold.

In order to understand the role of text and explicit norms in language planning we must understand the current trends of language technology, which no doubt include the trend of machine learning from Big Data. Language technology applications are, to an increasing degree, constructed with the help of large data collections by large companies whose main focus is outside language technology. These companies will never have national language planning high on their agenda. Their optimal scenario seems to be data-driven language technology with as few philologists as possible, which is easy to roll out for new languages and with minimal additional costs for each new language. The focus is on the customer, who did not buy proofing tools but got them “for free” when buying something else, and not on the language community as such.

2. Proofing and dynamic compounding

Dynamic compounding is found in Europe in the area between English, Slavic, and Romance, i.e., it covers the Germanic, Finnish and Saami language area. In these languages, compounds like *reindeer husbandry agreement negotiations* are written as one word, with non-trivial distribution of internal morphology (the Norwegian suffix *-s-* is historically a genitive suffix), as shown in (1):

(1) *reindriftsavtaleforhandlingar* (*Norwegian*)
rein-drift-s-avtale-forhandling-ar
reindeer-operation-COMPSUFF-agreement-negotiation-PL.INDEF

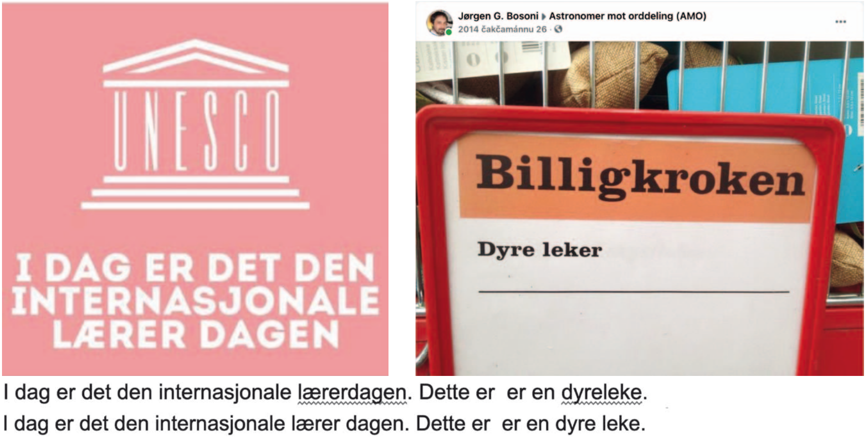
poronhoitosopimusneuvottelut (*Finnish*)
poro-n-hoito-sopimus-neuvottelu-t
reindeer-GEN-operation-agreement-negotiation-PL

reindeer husbandry agreement negotiations

The compounds in (1) are lexicalised, but also ad hoc neologisms like Finnish *yhdyssanakeskustelufoorumi* (“compound word discussion forum”) are perfectly fine.

Now, the question is how this may be handled in a spellchecker. There used to be three ways of making a spellchecker: the wordform list approach, the stem + affixes approach and the grammatical approach. The wordform list approach is good for languages with no or almost no morphology, like most Polynesian languages or even English. The stem + affixes approach is a good fit for languages with regular suffixation, such as Turkish or the Uralic language Komi. In the grammatical approach, stems and affixes are paired with lexeme and grammatical properties and subsequently combined with a model dealing with morphophonological processes. This spellchecker is good for languages with complex morphology, like the Saami languages or Finnish.

The two first methods dominated until the 1990s, and still do in many contexts. What they have in common is that they do not handle dynamic compounding. As a result of this, erroneously split compounds became common with the introduction of computers and spellcheckers during the late 1980s. The two examples in Figure 1 are taken from a Facebook group devoted to making fun of such errors. The first example, celebrating international teachers’ day, shows that (people advertising for) teachers also make these mistakes. The second example shows that the basket containing cheap commodities, *Billigkroken*, does not contain “animal toys" (*dyreleker*), as intended, but instead contains "expensive toys" (dyre leker). This error type may certainly be due to influence from English, but what is relevant to the topic of this article is that spellcheckers without dynamic compounding mark dynamic compounds as wrong and instead suggest the erroneous split forms. With no access to a spellchecker from the late 1980s, the "corrections" are taken from Google Docs.[1]


UNESCO
I DAG ER DET DEN INTERNASJONALE LÆRER DAGEN
Jørgen G. Bosoni ▸ Astronomer mot orddeling (AMO)
2014 čakčamánnu 26
Billigkroken
Dyre leker

I dag er det den internasjonale lærerdagen. Dette er er en dyreleke.
I dag er det den internasjonale lærer dagen. Dette er er en dyre leke.

Fig. 1: Norwegian compound errors posted in the Facebook group „Astronomer mot orddeling" (Astronomers against split compounds)

The grammatical method became available in the 1990s, for example in Lingsofts spellcheckers for the Nordic languages, and was integrated in Microsoft Word. In this model, there were explicit rules for compounding, and the spellcheckers were thus able to accept nonlexicalized compounds. The problem of dynamic compounding was then solved. Unfortunately, the solution introduced problems with overgeneration, leading to false negatives (unrecognised typos), like the Norwegian common typo in (2), where the correct form would be the adverb *nettopp* "recently, exactly, perfectly", but the typo is disguised by the spellchecker as an absurd compound.

(2) *netopp
ne-topp
old.moon-peak
"the peak of (the lunar phase) old moon"

[1] In fairness it must be added that Google Docs fared better than the spellcheckers of the 1980s in that it was able to recognise the plural form *dyreleker* but it still failed on the singular *dyreleke*.

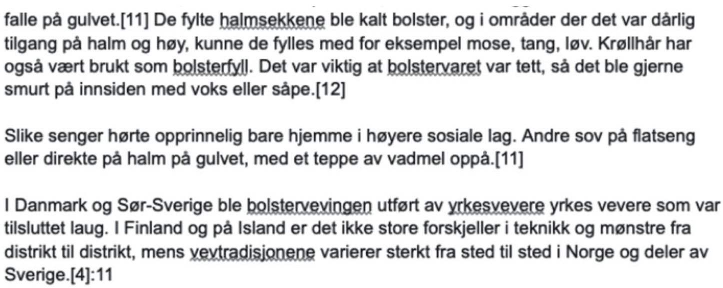
Allowing non-existing compounds of this type into the suggestion mechanism would, of course, add to the problem, since arbitrary compounding of short words in most cases would appear nonsensical and even mislead users into wrong writing habits. The obvious answer to this would be to block dynamic compounding with short words, e.g. 1-3 letter words, but keep it for longer words, like the rare but attested ones in (3):

(3) brettseglingsferie “surfing vacation”
kunnskapstype “knowledge type”
plosivgeminat “plosive geminate”

An even more drastic step would be to block dynamic compounding from the suggestion mechanism altogether.

Instead of efforts aiming at solving these problems, we now unfortunately see a return to spellcheckers based upon attested wordforms only, with Google as its main proponent.

One may think the the solution for word- and text-based approaches is “more text”, and yes, more text does help. The following two figures show text from Wikipedia in Norwegian Bokmål, corrected first by Google Docs and then by giella-nob, a spellchecker based on a finite-state transducer for Norwegian Bokmål.[2] The text contains no typos.



falle på gulvet.[11] De fylte halmsekkene ble kalt bolster, og i områder der det var dårlig tilgang på halm og høy, kunne de fylles med for eksempel mose, tang, løv. Krøllhår har også vært brukt som bolsterfyll. Det var viktig at bolstervaret var tett, så det ble gjerne smurt på innsiden med voks eller såpe.[12]

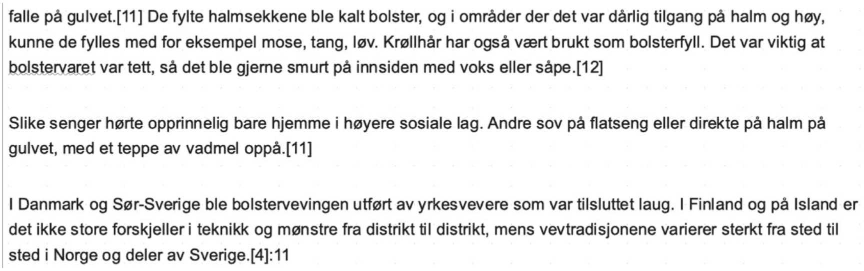
Slike senger hørte opprinnelig bare hjemme i høyere sosiale lag. Andre sov på flatseng eller direkte på halm på gulvet, med et teppe av vadmel oppå.[11]

I Danmark og Sør-Sverige ble bolstervevingen utført av yrkesvevere yrkes vevere som var tilsluttet laug. I Finland og på Island er det ikke store forskjeller i teknikk og mønstre fra distrikt til distrikt, mens vevtradisjonene varierer sterkt fra sted til sted i Norge og deler av Sverige.[4]:11

Fig. 2: Norwegian Bokmål Wikipedia text, corrected by Google Docs

Most of the alleged typos are rare words, linked to traditional handicrafts in preindustrial times. None of them is found in the 750 million word corpus NoWaC “Norwegian Web as a Corpus” created by the University of Oslo. The spellchecker based on the finite-state transducer allows for dynamic compounding. The false positive *bolstervaret* is due to the noun *var* being blocked from dynamic compounding given that it contains only 3 letters.

[2] https://giellalt.github.io/lang-nob/.



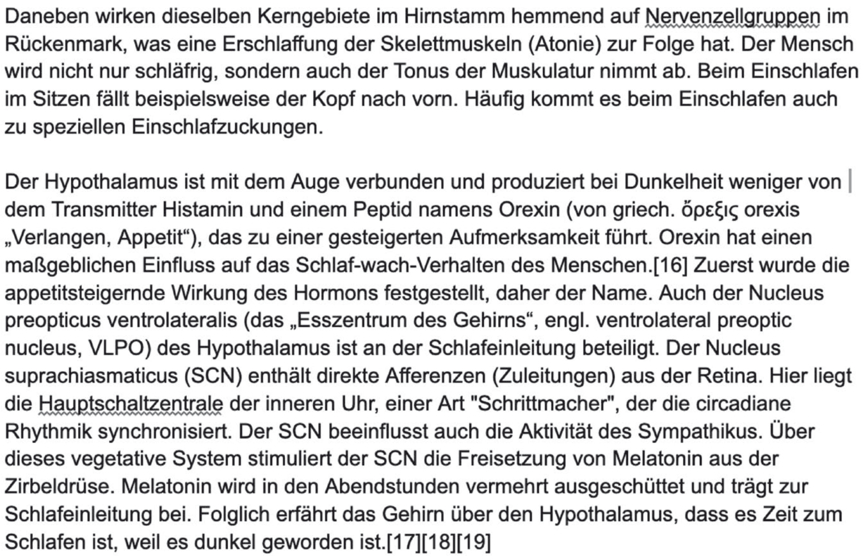
falle på gulvet.[11] De fylte halmsekkene ble kalt bolster, og i områder der det var dårlig tilgang på halm og høy, kunne de fylles med for eksempel mose, tang, løv. Krøllhår har også vært brukt som bolsterfyll. Det var viktig at bolstervaret var tett, så det ble gjerne smurt på innsiden med voks eller såpe.[12]

Slike senger hørte opprinnelig bare hjemme i høyere sosiale lag. Andre sov på flatseng eller direkte på halm på gulvet, med et teppe av vadmel oppå.[11]

I Danmark og Sør-Sverige ble bolstervevingen utført av yrkesvevere som var tilsluttet laug. I Finland og på Island er det ikke store forskjeller i teknikk og mønstre fra distrikt til distrikt, mens vevtradisjonene varierer sterkt fra sted til sted i Norge og deler av Sverige.[4]:11

Fig. 3: Norwegian Bokmål Wikipedia text, corrected by giella-nob

For a national language like Norwegian Bokmål, Google is thus not able to collect enough text to produce a reliable spellchecker. More available text does help, though. Figure 4 gives an example of German scientific text, containing no typos but technical terms, loanwords and even some English and Greek. The latter would, of course, have been out of reach for all but text-based approaches. There are two false positives, though: *Nervenzellgruppen* and *Hauptschaltzentrale*. The two words stand out as being the only 3-part dynamic compounds in the text. Even the resources available for German, the largest language in Europe, is thus not enough to cover words like these.



Daneben wirken dieselben Kerngebiete im Hirnstamm hemmend auf Nervenzellgruppen im Rückenmark, was eine Erschlaffung der Skelettmuskeln (Atonie) zur Folge hat. Der Mensch wird nicht nur schläfrig, sondern auch der Tonus der Muskulatur nimmt ab. Beim Einschlafen im Sitzen fällt beispielsweise der Kopf nach vorn. Häufig kommt es beim Einschlafen auch zu speziellen Einschlafzuckungen.

Der Hypothalamus ist mit dem Auge verbunden und produziert bei Dunkelheit weniger von dem Transmitter Histamin und einem Peptid namens Orexin (von griech. ὄρεξις orexis „Verlangen, Appetit“), das zu einer gesteigerten Aufmerksamkeit führt. Orexin hat einen maßgeblichen Einfluss auf das Schlaf-wach-Verhalten des Menschen.[16] Zuerst wurde die appetitsteigernde Wirkung des Hormons festgestellt, daher der Name. Auch der Nucleus preopticus ventrolateralis (das „Esszentrum des Gehirns“, engl. ventrolateral preoptic nucleus, VLPO) des Hypothalamus ist an der Schlafeinleitung beteiligt. Der Nucleus suprachiasmaticus (SCN) enthält direkte Afferenzen (Zuleitungen) aus der Retina. Hier liegt die Hauptschaltzentrale der inneren Uhr, einer Art "Schrittmacher", der die circadiane Rhythmik synchronisiert. Der SCN beeinflusst auch die Aktivität des Sympathikus. Über dieses vegetative System stimuliert der SCN die Freisetzung von Melatonin aus der Zirbeldrüse. Melatonin wird in den Abendstunden vermehrt ausgeschüttet und trägt zur Schlafeinleitung bei. Folglich erfährt das Gehirn über den Hypothalamus, dass es Zeit zum Schlafen ist, weil es dunkel geworden ist.[17][18][19]

Fig. 4: German Wikipedia text corrected by Google Docs

3. The text corpus and the explicit norm

Looking at the problems with the text-based approach in more general terms, the false positives shown here may be seen as an out-of-vocabulary problem. This problem is obviously worse for languages with dynamic compounding than for languages without. Even though it is of no help to the large group of North European languages, at least one may think that a language without compounding and with little morphology would probably get a good spellchecker with far less text than what is available for Norwegian Bokmål.

But the problem is far worse than this. The underlying assumption when basing correction on attested forms is that *the text collection equals the norm.* This implies a principled exclusion of language normative work done by normative bodies, indeed a principled exclusion of language planning as such. The role of normative language institutions is (among many other things) to give advice on how to spell words. The question is thus whether the set of available text collections could be seen as a de facto norm, replacing the explicitly stated norm. Such a move will no doubt result in proofing tools that can help writers "write like all the others", but for normative bodies the answer cannot be but negative.

Proofreaders will tell us that people do make mistakes in writing. Unfortunately, proofreaders are an endangered species. More and more texts are published without proofreading. The democratisation of publishing that came with computers and the internet clearly has its downsides: abolishing typographers has given us ugly typography and abolishing proofreaders has given us more typos. Developing proofing tools from collected texts is thus becoming increasingly problematic. Ideally, the collected texts should, of course, be error free, but this is, to an increasing extent, not the case for publicly available text. Whereas correct forms in most cases outnumber incorrect forms for majority languages (due to fairly good writing skills and huge amounts of text), minority language communities face the double challenge of poorer writing skills and far less text where the correct forms could outweigh the typos.

For minority languages like South Saami, with fewer than 500 speakers, there is another problem. Corpora available for such languages do not even number millions of words. There is also no point in waiting for larger corpora: Small language communities simply do not have enough writers to write the amount of text available for German or Norwegian. Typologically, minority languages often have quite complex morphologies, with a high ratio of words occurring only once in the corpus. For large and more stable written languages, it is to be hoped that the errors would be outnumbered by correct forms, but this is not the case for minority languages.

Furthermore, minority languages typically have young written languages and a norm with a weak status in the language societies concerned. These languages have a marginal position in education and mass media and the normative bodies behind the standards have few ways of enforcing the norm. The key to mastering a written standard is to be exposed to it via extensive reading. Minority languages are predominantly oral, and these languages are rarely used for commercial billboards, film subtitles, etc. The written norm often has a weak status and mother tongue speakers of minority languages tend to choose forms outside the standard. A large percentage of L2 writers also leads to both spelling and grammatical errors. For minority language communities, there is, thus, no way that a collection of texts can set the norm.

4. Tech giants and language communities

Even though the number of languages for which Microsoft and Google offer support is increasing, it is still small: Windows 11 has localisation and proofing for 85 languages and Google Translate is available for 108, when there are 3,514 languages for which there is a translation of at least the New Testament.

Microsoft is making it increasingly harder for third-party providers to add proofing tools to Microsoft Word. With Google, it has always been impossible. The single most important tool for a normative body to implement its norm among writers is the spellchecker. The normative body would thus want to control the content of the spellchecker and it will thus often not be satisfied with the proofing tools offered by the large companies. Moreover, the 3,400 ignored language communities will not get any proofing tools. The result is that the most central common infrastructure of any society, its language, is outside the control of the society to which it belongs.

As language societies, we should not accept being governed by large computer companies. What we need is an independent language technology. The large companies should, of course, make their language tools as they see fit, but they should not prevent language communities from making and distributing their own.

An independent language technology will construct explicit language models. It can take data into consideration but will not be data driven. When needed, the language models will be built as a set of explicit linguistic rules. Such models are transparent: it is possible to correct the models when they make mistakes or when we want them changed due to changes to the language norm. Language corpora are certainly not irrelevant, as any language planner knows. But rather as being seen as The Norm, they should be given the role as a test bench, a reality check: Where should we invest our normativity efforts? What is the balance between linguistic development and language norm? For terminology and vocabulary: *what is actually in use?*

This view has consequences for the relation between language and computers. As language societies, we cannot accept that the very thing that constitutes us as such societies, our language, is beyond our control. Thus, our language models must be made available to the language communities, via the word processing programs that the communities use. The large technology companies have taken it upon themselves to carry the infrastructure of our societies. For this contract to be upheld, they cannot treat language as if it were any commodity. It is not.

An independent language technology can be made in many ways. The main criteria are transparent code and the possibility of governing its properties, thus explicitly deciding the norm.

Our experiences at UiT in Tromsø in Norway are as follows: We work on complex languages with little text, in other words, we work on average human languages. We model the lexicon, compounding, derivation and inflection as finite state transducers. Syntactic analysis and language advice to writers involving sentence or text context is modelled as constraint grammar. This is then integrated in text processing programs (if possible), with good results.

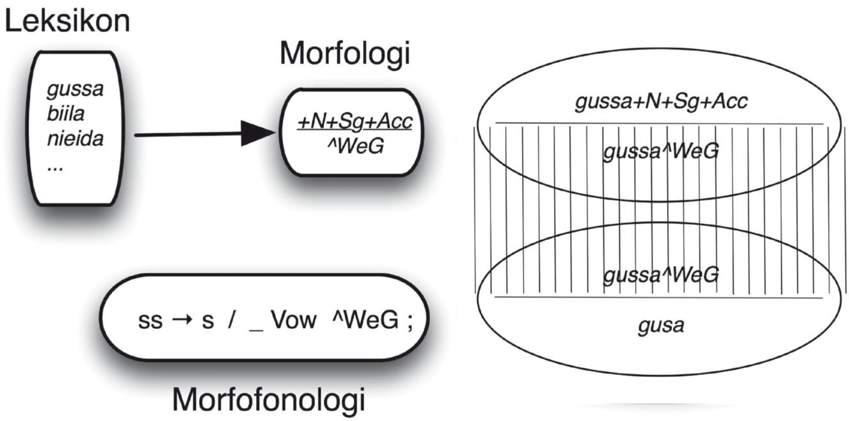
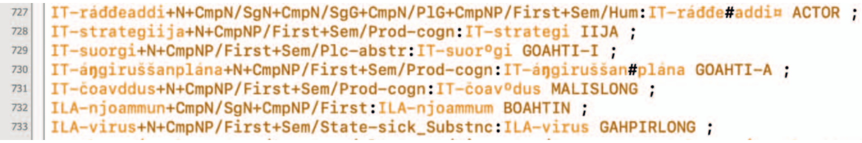


Fig. 5: Finite state transducers as language models for North Saami

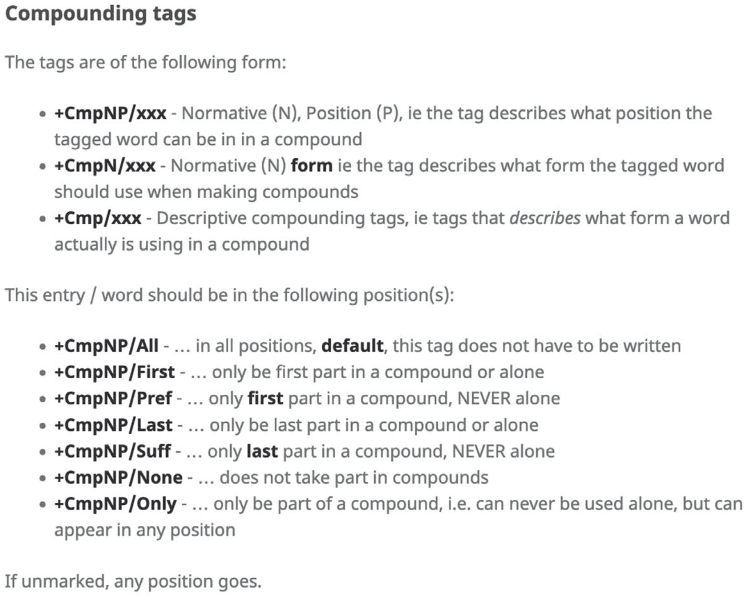
We explicitly govern dynamic compounding by adding tags to the lexicon, as in Figure 6.



```
IT-ráđđeaddi+N+CmpN/SgN+CmpN/SgG+CmpN/PlG+CmpNP/First+Sem/Hum:IT-ráđđe#addi¤ ACTOR ;
IT-strategiija+N+CmpNP/First+Sem/Prod-cogn:IT-strategi IIJA ;
IT-suorgi+N+CmpNP/First+Sem/Plc-abstr:IT-suorºgi GOAHTI-I ;
IT-áŋgiruššanplána+N+CmpNP/First+Sem/Prod-cogn:IT-áŋgiruššan#plána GOAHTI-A ;
IT-čoavddus+N+CmpNP/First+Sem/Prod-cogn:IT-čoavºdus MALISLONG ;
ILA-njoammun+CmpN/SgN+CmpNP/First:ILA-njoammum BOAHTIN ;
ILA-virus+N+CmpNP/First+Sem/State-sick_Substnc:ILA-virus GAHPIRLONG ;
```

Fig. 6: Tags governing compound behaviour, North Saami lexicon

The compound tags are defined in Figure 7.



Compounding tags

The tags are of the following form:

- **+CmpNP/xxx** - Normative (N), Position (P), ie the tag describes what position the tagged word can be in in a compound
- **+CmpN/xxx** - Normative (N) **form** ie the tag describes what form the tagged word should use when making compounds
- **+Cmp/xxx** - Descriptive compounding tags, ie tags that *describes* what form a word actually is using in a compound

This entry / word should be in the following position(s):

- **+CmpNP/All** - ... in all positions, **default**, this tag does not have to be written
- **+CmpNP/First** - ... only be first part in a compound or alone
- **+CmpNP/Pref** - ... only **first** part in a compound, NEVER alone
- **+CmpNP/Last** - ... only be last part in a compound or alone
- **+CmpNP/Suff** - ... only **last** part in a compound, NEVER alone
- **+CmpNP/None** - ... does not take part in compounds
- **+CmpNP/Only** - ... only be part of a compound, i.e. can never be used alone, but can appear in any position

If unmarked, any position goes.

Fig. 7: Compound tags (cf. https://giellalt.github.io/lang-sme/src-fst-root.lexc.html)

Others may do it differently. This is fine, as long as your language model does what you want, and you are able to put it into use in the word processor. What we do at UiT is openly available for adaption and reuse at https://giellalt.github.io/.

5. Conclusion

Normative language work must be independent from and stand above actual language use. This calls for an explicit and transparent language technology. Such a language technology is threatened from two sides: from the dominant trend within AI, favouring data-driven approaches, and from the major programming houses, preventing third-party language technology programs from being integrated in their word processor software. As shown here, an alternative path is possible: to develop transparent open source rule-based systems that can be easily integrated into the linguistic software of the big tech companies. The issue is too important to let slip.