The decade construction rivalry in Russian: Using a corpus to study historical linguistics

Tore Nesset & Anastasia Makarova
UiT – The Arctic University of Norway

Abstract

This article addresses the question of the diachronic development of so-called rival forms, i.e., words or grammatical constructions that appear to be synonyms, based on a detailed empirical analysis of two seemingly synonymous constructions in Russian. Corresponding to the English 'decade construction' *in the twenties*, Russian has two rival constructions, viz. *v dvadcatye gody* [*lit.* "in the twentieth years"] (with the numeral and noun in the accusative) and *v dvadcatyx godax* (with the numeral and noun in the locative case). Three hypotheses about rival forms are considered: leveling (whereby one form ousts its rival), sociolinguistic differentiation (whereby the two rivals survive in different varieties of a language) and semantic differentiation (whereby the two rivals develop different meanings over time). Contrary to what has been suggested in the literature, we find little evidence for semantic and sociolinguistic differentiation. Instead, we demonstrate that leveling is taking place, since the accusative construction is in the process of ousting its rival. While our study shows that corpus data facilitate detailed analysis of the interaction between leveling, sociolinguistic differentiation and semantic differentiation, our analysis also points to limitations, especially when it comes to corpus-based analysis of sociolinguistic and semantic factors.

Keywords: Russian; corpus linguistics; temporal adverbials; leveling; sociolinguistic differentiation; semantic differentiation; CART analysis

1. Introduction: Three hypotheses about rival forms[1]

It has often been noticed that complete synonymy is strongly disfavored to the extent that genuine examples may not exist at all:

(1)   The Principle of No Synonymy: "If two constructions are syntactically distinct, they must be semantically or pragmatically distinct" (Goldberg 1995: 67).

This principle has a number of different names in the literature. Haiman (1980: 516) labels it 'iconicity of isomorphism' about "the commonly accepted axiom that no true synonyms exist, i.e. that different forms must have different meanings", while Clark (1993: 2; see also Croft 2001: 111) refers to the same idea as the 'Principle of Contrast' ("Every two forms contrast in meaning"). Carstairs-McCarthy (1994), who extends Clark's principle to inflection classes in morphology, uses the name 'No Blur Principle'.

In (1), Goldberg focuses on syntactic constructions, and relevant studies couched in the framework of Construction Grammar and other varieties of cognitive linguistics include, e.g., Wulff (2006), Gilquin (2010: 97–143), Klavan et al. (2011) and now Grieve (2016). Important studies of seemingly synonymous syntactic constructions include the so-called dative alternation (*give Jane an apple* vs. *give an apple to Jane*, e.g., Bresnan et al. 2007 and

---

[1] We would like to thank the editor of *Diachronica*, three anonymous referees and the members of our research group CLEAR (Cognitive Linguistics: Empirical Approaches to Russian) for constructive criticism of earlier versions of this article.

Bresnan & Ford 2010) and the locative alternation (*load the hay onto the truck* vs. *load the truck with hay*, e.g., Iwata 2008 and Sokolova et al. 2012). Bolinger (1968: 127) argues that "a difference in syntactic form always spells a difference in meaning".

However, synonymous syntactic constructions still appear to be understudied; it has been more common to relate the ban on complete synonymy to words rather than syntactic constructions. For instance, Cruse (1986: 270) argues that "one thing becomes clear once we begin a serious quest for absolute synonyms, and that is that if they exist at all, they are extremely uncommon." This observation is anything but new. Nida (1958) stated that "there are no complete synonyms within a language", and Bloomfield (1933: 145) says that "there are no actual synonyms." In Saussure's *Cours* (1983 [1916]: 167) we read that "inevitably the phonetic difference which has emerged will tend to acquire significance", which implies that languages tend to avoid complete synonymy.

What are the diachronic implications of the Principle of No Synonymy? As pointed out by Cruse (1986: 270), if absolute synonymy "were to occur, it would be unstable". We may consider (apparent) synonyms to be 'rival forms' (Baayen et al. 2013: 254), i.e., forms that compete for the same semantic or functional 'slot' in a language. If we take the insight that synonymy is unstable seriously, we expect rival forms to undergo diachronic change. We distinguish between three diachronic scenarios (e.g., Cruse 1986: 270; Croft 2000: 177–178; Szymanek 2005: 441; Nuyts & Byloo 2015: 62–63). One form may outcompete its rival so that only one of the rivals survives, or both rivals may survive but in different language varieties, or synonymy may be eliminated because the rival forms develop different meanings. We refer to these scenarios as 'leveling', 'sociolinguistic differentiation' and 'semantic differentiation':

(2)   a.   Leveling: One form ousts its rival over time.
        b.   Sociolinguistic differentiation: The two rivals survive in different varieties of a language.
        c.   Semantic differentiation: The rival forms develop different meanings over time.

It is worth noting that scenarios (2a) and (2c) involve language internal factors, while (2b) is an example of language external factors driving language change.

In historical linguistics, 'leveling' as a term is often reserved for "the complete or partial elimination of morphophonemic alternations within paradigms" (Hock 1988: 168) or "the elimination of (unimportant) morpheme or stem alternations within paradigms" (Hock 1988: 183). We use 'leveling' in a broader sense as a cover term for (complete or partial) elimination of rivalry in language, including rival synonymous forms. For the purposes of the present study, 'leveling' is used for situations where one form survives, while its competitor gradually goes out of use and finally disappears from the language. As an example of leveling in this sense, consider the rivalry between Old Norse *ljúfr* and *kærr*, which both had the meaning "dear" (Bjorvand 2000: 325). The latter, a borrowing from French (cf. *cher*), ousted its rival. An example of sociolinguistic differentiation is *kid* and *child* in English. The former, possibly a Scandinavian borrowing with the original meaning "the young of a goat", is attested for human offspring from the 1590s (Harper 2001–2017). Both words exist in modern English, but they have undergone sociolinguistic differentiation, insofar as *kid* belongs to more informal registers than *child*. An example of semantic differentiation is the rivalry between the two plural forms *brethren* and *brothers* in English, where the former has survived in the specialized meaning "member of a society, profession or sect" (Croft 2000: 177).

Note how tightly intertwined the scenarios in (2a-c) are. Old Norse *ljúfr* "dear" has disappeared from many varieties of Norwegian (including Bokmål), but it has survived in some varieties (including Nynorsk), where it is used as a poetic word in the meaning

"wonderful" (Hovdenak et al. 2001). In these varieties, the word has undergone sociolinguistic differentiation since it has become restricted to poetic registers, and at the same time the change in the meaning from "dear" to "wonderful" testifies to semantic differentiation.

These examples are lexical. However, it is far from obvious that rivalry among lexical items behaves in the same way as rivalry concerning syntactic constructions. In the present article, we investigate the interactions between leveling, sociolinguistic differentiation and semantic differentiation on the basis of two rival syntactic constructions concerning temporal adverbials referring to decades in Russian. Both constructions involve the preposition *v* "in(to)", but the governed noun phrase is in different cases: the accusative or the locative. We will treat the three scenarios in (2) as hypotheses that will be tested empirically against corpus data. Our contribution is fivefold. First, we show that we are dealing with a case of double motivation, insofar as both rivals are well motivated by the grammar of Russian. Second, contrary to what has been claimed in the scholarly literature (Wade 1992), we find no conclusive evidence for semantic differentiation. Third, our data offer little support for sociolinguistic differentiation. Fourth, we demonstrate that the distribution of the two constructions has changed dramatically over the last two centuries and that one of the constructions is in the process of ousting its rival from the language. Finally, our study testifies to the value of diachronic corpus studies of rival forms. We show how corpus data can shed light on the interaction between the three scenarios in (2), but at the same time our study also illustrates some of the limitations of a corpus-based approach – especially with regard to sociolinguistic and semantic differences.

In §2 we present the two rival decade constructions and show that they are an example of double motivation. Sections 3 through 5 explore the three scenarios in (2) individually, before we turn to their interaction in §6. This study's contribution is summarized in §7.

2. The rival decade constructions in Russian: Double motivation

Speakers of Russian have two options for describing events that took place within a decade. Corresponding to the English construction *in the twenties, in the thirties,* etc., speakers of Russian form a temporal adverbial with the preposition *v* "in(to)" followed by a noun phrase consisting of an ordinal numeral and the noun *god* "year" in the plural. As shown in (3), the noun phrase may occur in the accusative case:[2]

(3)   Lučšie svoi knigi on napisal **v 60-e**$_{ACC}$ **gody**$_{ACC}$. [«Zvezda» 2002]
        "His best books he wrote **in the sixties**."

However, *v* governs both the accusative and the locative (prepositional) cases in Russian, and the noun phrase representing the decade may also occur in the locative:

(4)   Lučšie svoi stixi on napisal **v dvadcatyx**$_{LOC}$ **godax**$_{LOC}$. [Berberova 1960–1966]
        "His best poems he wrote **in the twenties**."

As suggested by examples like (3) and (4), where the accusative and locative decade constructions occur in identical contexts, the two constructions are rival forms that appear to be complete synonyms. The question therefore arises as to whether the rival decade

---

constructions involve any of the scenarios of leveling, sociolinguistic differentiation and semantic differentiation discussed in the previous section. However, before we turn to these questions in the following sections, it is important to see how both the accusative and locative constructions are motivated by Russian grammar. We show that decades are different from other time spans; while most time spans show strong preferences for one case, decades involve a high degree of variation between the accusative and the locative cases, which are both motivated by the grammar.

In Russian, the preposition *v* plus a noun phrase indicating a time span is used in temporal adverbials of the type that Haspelmath (1997: 29) refers to as markers of "simultaneous location" and Klein (1994: 149, 2009: 65) calls "temporal adverbials of position". Simplifying somewhat, these adverbials answer the question *when does something happen?* The choice between the accusative and locative cases depends on the properties of the relevant time span. If the adverbial involves a bounded time span longer than a week, such as a month, year, century or millennium, the noun phrase is in the locative (see Nesset 2004: 286–291, also Nesset 2013; Makarova & Nesset 2013; Nesset & Makarova 2015):

(5)   **V ètom**$_{LOC}$ **godu**$_{LOC}$ on kupil tri pary lyž. [«Izvestija» 2002]
        "**This year** he bought three pairs of skis."

If the time span in question is shorter than a week (e.g., a second, minute, part of a day, week day), the noun phrase is in the accusative (Nesset 2004: 287):[3]

(6)   **V sredu**$_{ACC}$ on otpravilsja na jarmarku. [Babel' 1920–1928]
        "**On Wednesday** he left for the market."

Examples (5) and (6) involve bounded time spans, insofar as both a year and a Wednesday have fixed lengths and clearly defined beginnings and ends. For unbounded time spans without clearly defined length, beginning and end, the accusative is used (Nesset 2004: 289):

(7)   U menja **v èto**$_{ACC}$ **vremja**$_{ACC}$ byl roman – s čudesnoj Maškoj. [Belousova 2000]
        "**At this time** I was having a relationship – with the wonderful Masha."

It is not clear exactly when "this time" begins and ends and how long it lasts, so here the accusative is used.

The grammatical category of number has an impact on case usage in Russian temporal adverbials. While (5) shows that long and bounded time spans such as *god* "year" occur in the locative, this only applies to noun phrases in the singular. In (8), where the time span is represented as a noun phrase in the plural, the accusative is used. As shown by Nesset (2004: 302–308), this is a general rule.

(8)   **V èti**$_{ACC}$ **gody**$_{ACC}$ on napisal … povest' «Bašmaki». [Varlamov 2002]
        "**In those years** he wrote the novella *The shoes*."

Nesset (2004) argues that there is a relationship between boundedness and plurality. While "this year" in (5) is a bounded period with a clearly defined beginning, length and end, "those years" in (8) is unbounded: we do not know exactly how many years are referred to, so "those years" is a time span with no clearly defined beginning, length or end. However, we

---

will not discuss whether the generalization about plurality is best understood as a semantic rule (i.e., a consequence of boundedness) or a purely syntactic rule based on the morphosyntactic category of number.

Although this short exposition does not do justice to the complexities of temporal adverbials in Russian, for present purposes the following simple system of generalizations is sufficient:

(9) a. Bounded time span longer than a week: v + locative
     b. Other time span (unbounded or shorter than a week): v + accusative
     c. Time span in the plural: v + accusative

Is the accusative or the locative the default case for Russian temporal adverbials of the type in question? Since (9b) contains a disjunction (unbounded or short time span), we can consider accusative the default case in temporal adverbials of the relevant type, i.e., the case that is used whenever the conditions for the use of the locative are not met. We could conflate (9b–c) and simplify the system by saying that the locative is used for long bounded time spans in the singular, while the accusative is used elsewhere. However, although the notion of 'default' is arguably of importance for the distribution of the two decade constructions, the system in (9) cannot be simplified, since it shows that the rivalry between the accusative and locative cases in the decade constructions is an example of double motivation.[4]

Let us first ask whether the use of the locative with decades is motivated on the basis of the generalizations in (9). The answer is clearly in the affirmative. A decade is a bounded time span, since it has a fixed length (ten years) and a clearly defined beginning and end. A decade is furthermore longer than a week and therefore fulfills the criterion of generalization (9a), which predicts the locative for bounded time spans longer than a week. Therefore the use of the locative for decades is strongly motivated by the grammar of Russian, and it comes as no surprise that the locative is used for decades in examples such as (4) above.

At the same time, we find motivation for the use of the accusative, too. Recall from examples (3) and (4) that the Russian decade constructions involve plural forms. According to generalization (9c), we expect the accusative when the time span is in the plural, and this expectation is borne out by the facts insofar as the accusative is used in examples like (3). In addition, we have seen that the accusative can be regarded as the default for temporal adverbials of the relevant type, so the use of the accusative for decades is in harmony with the default pattern.

In other words, we are dealing with an example of double motivation where the locative is motivated by the length and boundedness of the time span, whereas the accusative is motivated by the fact that the decade construction involves the plural and that the accusative is the default case. The question now arises whether the motivation for the locative or the accusative is stronger. Do length and boundedness provide stronger motivation than plural and default? Does the distribution of the two constructions change over time?

In order to investigate this, we created a database of examples extracted from the Russian National Corpus, a family of electronic corpora freely available online. Our study is based on the main corpus (Russian: "*osnovnoj korpus*"), which consisted of approximately 230 million words when the corpus searches were performed in June 2015. The main corpus includes written texts from 1700 to 2017 and is continuously expanded. A number of genres are represented, including fiction, journalism and scientific/educational texts (see §4.1 for discussion). Each example is provided with detailed information about the title of the text, the name of the author, the genre, the year the text was created, etc.

Our database consists of 5,453 examples (2,670 examples with the accusative construction and 2,783 with the locative). We performed searches for the preposition *v* followed by an ordinal numeral and *god* "year" in the accusative or locative plural. We searched for numerals written as numbers (as in (3)) or as words (as in (4)) and restricted the searches to one example per document.[5] All search results were exported to a spreadsheet and subsequently conflated into one document. We then weeded out irrelevant examples manually and annotated the database for additional parameters that will be discussed in detail in the following sections. The entire database is available at the Tromsø Repository of Language and Linguistics (TROLLing) at http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE. The database contains examples from 1830 to 2012. The distribution of the examples over time is shown in Table 1, which also contains information about the size of the main corpus for different periods.[6] The columns marked with '#' give raw numbers, while the columns with '%' show each period's proportion of the total number of words. As we can see, the number of examples with the decade constructions increases over time. This is not surprising, since the corpus contains more words from recent times.

Table 1. Distribution over time – number of words in the main corpus in general and number of attestations with decade constructions

Period	Main corpus overall		Decade constructions	
	#Words	%	#Examples	%
1701-1750	590,541	0.3	0	0
1751-1800	2,981,803	1.4	0	0
1801-1850	10,380,375	4.8	45	0.8
1851–1900	31,761,447	14.7	448	8.2
1901–1950	53,445,536	24.7	731	13.4
1951–2000	67,252,763	31.0	2,307	42.3
2001–2012	50,231,677	23.2	1,922	35.2
Total	216,644,142	100.0	5,453	100.0

Our database provides a good basis for investigating the development of the two rival decade constructions in the 19th, 20th and 21st centuries. As we will see in the next section, the accusative construction is in the process of ousting the locative construction. This suggests that plural (which motivates the accusative) is becoming a stronger motivation than length and boundedness of the time span. At the same time, the increasing use of the

---

[4] The notion of 'default' is used in different ways in the literature (Fraser & Corbett 1997). We use it to mean 'normal case', i.e., what applies in the absence of blocking information.

[5] The reason why we limited the searches to one (randomly chosen) example per document was to avoid violating assumptions of independence of observations so that we could perform statistical tests (e.g., chi-squared). Statistical models like the chi-squared test are based on the assumption that each observation (i.e., each example) is independent, and in order to perform such statistical tests we cannot have many examples from the same author. We also avoid a skewed dataset where some texts by some authors would be overrepresented (i.e., involve a large number of examples of the constructions under scrutiny), while other texts by other authors would be underrepresented (i.e., involve few examples of the relevant constructions).

[6] The information about the corpus size is taken from http://ruscorpora.ru/corpora-stat.html (accessed May 2017). Numbers given at this webpage are not completely up to date, as the total size of the main corpus is reported to be 216 million words, while it was 230 million words at the time when we performed our corpus searches in June 2015. However, the recent expansion of the corpus is not likely to have changed the overall situation shown in Table 1. Although our data were collected in 2015, our most recent examples are from 2012, and we therefore refer to the final period in Table 1 and other tables as '2001–2012'.

accusative indicates that the decade constructions are moving towards the default case, which, as mentioned above, is the accusative for temporal adverbials of the relevant type.

3. Hypothesis 1: Leveling of form over time

In §1, we saw that languages tend to avoid full synonymy, and we noted that one of the ways this can happen is through 'leveling', whereby one form outcompetes its rival, which gradually disappears from the language. With reference to the competing decade constructions in Russian, we can formulate the hypothesis in (10). This hypothesis is motivated by the Principle of No Synonymy cited in (1), insofar as it represents one way the Russian language can get rid of the synonymy of the two rivaling decade constructions:

(10)   The Leveling Hypothesis:
        One of the decade constructions ousts its rival over time.

We test this hypothesis in two ways. First, for each example the Russian National Corpus provides the year when the relevant text was created. This makes it possible to compare the use of the accusative and locative in texts from different times. Second, the corpus includes information about the year of birth of the author of the relevant texts, which facilitates comparison of case usage of authors born at different times.
        It is important to investigate both parameters. Since authors may change their linguistic habits during their life span, it may in principle be the case that the year when the text was created and author's year of birth give different results. However, as we will see, for the decade constructions both parameters indicate the same development, whereby the use of the accusative increases over time following an S-shaped curve, while the use of the locative decreases.

3.1 Test 1: Date of the texts

To test the Leveling Hypothesis with regard to the year the texts were created, we divided the material into twenty-five year periods, as in Table 2. The table gives the raw numbers of examples with the accusative and locative (columns two and three from the left) and also the percentage of examples with the accusative (the rightmost column). In order to avoid the potential confounding factor of individual author preferences, we have included only one example per author in the table. We will come back to the preferences of individual authors in §4.3.

Table 2. Change in the use of the decade constructions over time – year when publication was created, one example per author

Period	#Accusative	#Locative	%Accusative
1826–1850	1	17	5.6
1851–1875	1	29	3.3
1876–1900	4	58	6.5
1901–1925	14	77	15.4
1926–1950	18	98	15.5
1951–1975	154	220	41.2
1976–2000	356	241	59.6
2001–2012	748	451	62.4
Total	1296	1191	52.1

The data in Table 2 offer strong evidence for the Leveling Hypothesis in (10). In texts from the 1800s the accusative is only sporadically attested, so for this period the locative construction is overwhelmingly dominant. During the 20th century the use of the accusative displays a strong increase and then reaches approximately 60% at the end of the 20th and the beginning of the 21st centuries.

3.2 Test 2: Author's year of birth

The Russian National Corpus does not provide information about the year of birth for all authors. For this reason the dataset reported on in Table 3 is smaller than the dataset discussed in §3.1. However, Table 3 contains 783 data points, which is sufficient to test the Leveling Hypothesis. Table 3 includes one example per author.

Table 3. Change in the use of the decade constructions over time – author's year of birth, one example per author

Period	#Accusative	#Locative	%Accusative
1776–1800	0	8	N.A.
1801–1825	2	32	5.9
1826–1850	2	47	4.1
1851–1875	15	77	16.3
1876–1900	29	80	26.6
1901–1925	80	84	48.8
1926–1950	145	88	62.2
1951–1975	55	35	61.1
1976–2000	3	1	N.A.
Total	331	452	42.3

The data in Table 3 reveal the same tendency as the data discussed in §3.1. For authors born before 1850, we have only a handful of examples with the accusative, which indicates that the accusative construction was possible but marginal for these generations. For authors born after 1850, the use of the accusative increases strongly until it reaches about 60% for authors born between 1926 and 1975. (For authors born before 1801 and after 1975 we have very few data points, so the corpus material does not enable us to draw any conclusions for these age groups.)

3.3 Comparison and discussion: Two parallel S-curves

Figure 1 visualizes the development with regard to both parameters discussed in §3.1 and §3.2. Both show parallel developments, insofar as the two curves have roughly the same shape. However, the solid curve representing author's year of birth is approximately fifty years behind the dashed curve representing the year that the texts were created, which is natural since most texts are presumably written by adults.
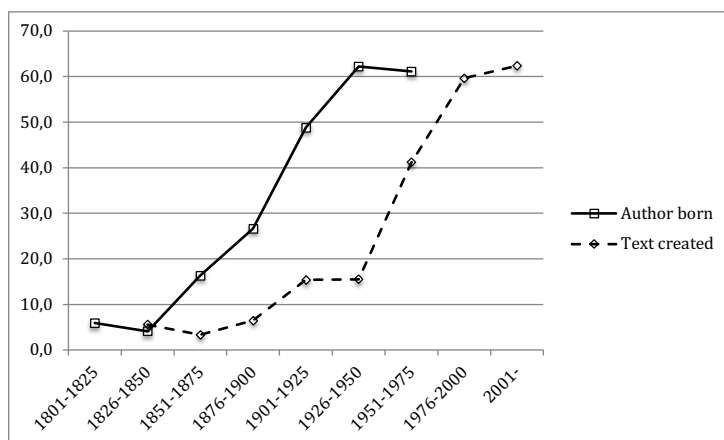
Figure 1. The proportion of examples with the accusative construction for different periods (based on Tables 1 and 2)

The development depicted in Figure 1 follows an S-shaped curve, i.e., a curve beginning with a relatively flat shape, a steep rise in the middle and then flattening out towards the end. Chambers (2002: 361; see also Chambers & Trudgill 1998: 159–165) argues that the S-curve represents a "kind of template for change", and Blythe & Croft (2012: 280), who discuss a large number of changes conforming to the S-shaped curve, state that "there are no clearly documented cases of change going toward completion that follows either a simple linear trajectory or an exponential curve". The fact that the change in the use of the Russian decade constructions follows the standard trajectory of language change provides strong support for the Leveling Hypothesis in (10).

Why is it the locative that is being ousted by the accusative and not vice versa? As we showed in §2, the accusative is the default case for Russian temporal adverbials with *v* "in". The S-curves documented in the present section therefore illustrate a development whereby the default case is in the process of further strengthening its position, while the non-default is losing ground.

This being said, a note of caution is in order. Although it looks like the accusative is in the process of stabilizing at about 60%, it is possible that the accusative will continue to increase in the future and that the locative will be completely marginalized in a couple of generations. The further development of the decade constructions is an empirical question we hope linguists of the future will return to.

4. Hypothesis 2: Sociolinguistic differentiation

Although, as shown in the previous section, the accusative is on the increase, the locative is still used in about 40% of the examples in texts created after 1975. Since the locative is still widely used, it is reasonable to ask whether the distribution of accusative and locative depends on sociolinguistic variables in present-day Russian, i.e., if the scenario of sociolinguistic differentiation mentioned in §1 is relevant for the Russian decade constructions. We will consider the following hypothesis:

(11)   The Sociolinguistic Differentiation Hypothesis:

The decade constructions survive in different varieties of the Russian language.

We use the term 'variety' in (11) in a wide sense so as to include not only sociolects and geographical dialects but also idiolects and variation according to register and genre (see Hudson 1980: 24). Unfortunately, the Russian National Corpus offers limited opportunities to test the hypothesis in (11), since the corpus provides very little sociolinguistically relevant information and no information about regional varieties. It was therefore not possible to carry out an investigation of geographical variation of the type Grieve (2016) did for American English. However, the corpus distinguishes between genres, and it is also possible to infer the gender of the authors. Finally, since the corpus provides the names of the authors, we are in a position to study idiolects, i.e., the preferences of individual language users. We consider genre, gender and idiolect in §4.1 through §4.3 before we sum up our findings in §4.4. Our data do not show effects of genre or gender, and although we find idiolectal differences, these differences seem to be a product of time, insofar as younger authors are more likely to use the accusative. Our findings therefore do not offer support for the Sociolinguistic Differentiation Hypothesis but instead lend additional support to the Leveling Hypothesis discussed in §3.

4.1 Test 1: Genres

We employ the term 'genre' as an English equivalent to *sfera* "sphere", which is used in the Russian National Corpus. The examples in our database are related to eight categories, which we have grouped into three broader genres in order to facilitate statistical analysis:[7]

(12)   a. Journalism:
- *publicistika, bytovaja, nexudožestvennaja* "journalism, everyday, non-fiction"
- *publicistika, nexudožestvennaja* "journalism, non-fiction"
- *reklama, nexudožestvennaja* "advertising, non-fiction"
       b. Fiction:
- *xudožestvennaja* "fiction"
- *xudožestvennaja, publicistika* "fiction, journalism"
       c. Scientific/educational:
- *učebno-naučnaja, nexudožestvennaja* "academic, non-fiction"
- *učebno-naučnaja, publicistika, nexudožestvennaja* "academic, journalism, non-fiction"
- *učebno-naučnaja, cerkovno-bogoslovskaja, nexudožestvennaja* "academic, clerical-theological, non-fiction"

The distribution of the accusative and locative constructions in these three genres is summarized in Table 4 and Figure 2. For each genre, the table provides information about the raw number of examples with accusative and locative, as well as the percentage of accusative. Each row represents the period when the relevant texts were created, and total scores for each genre are provided in the bottom row. The figure plots the development over time (measured in percentage of accusative) as three different curves – one for each genre. Since we have relatively little data before 1876, the figure only maps the development after this year.

---

[7] A small group of examples did not belong to any of these genres and were tagged as 'other' in the database.

Table 4. Changes in the use of the decade constructions for three genres (one example per author)

| Period	Journalism			Fiction			Scientific/educational		
	#Acc	#Loc	%Acc	#Acc	#Loc	%Acc	#Acc	#Loc	%Acc
1826–1850	0	8	N.A.	0	4	N.A.	1	4	N.A.
1851–1875	1	15	6.3	0	7	N.A.	0	7	N.A.
1876–1900	4	34	10.5	0	15	0	0	5	N.A.
1901–1925	8	49	14.0	3	10	23.1	3	16	15.8
1926–1950	9	45	16,7	2	18	10.0	6	32	15.8
1951–1975	98	123	44.3	22	25	46.8	18	51	26.1
1976–2000	251	166	60.2	49	39	55.7	51	31	62.2
2001–2012	582	330	63.8	32	19	62.7	122	89	57.8
Total	953	770	55.3	108	137	44.1	201	235	46.1
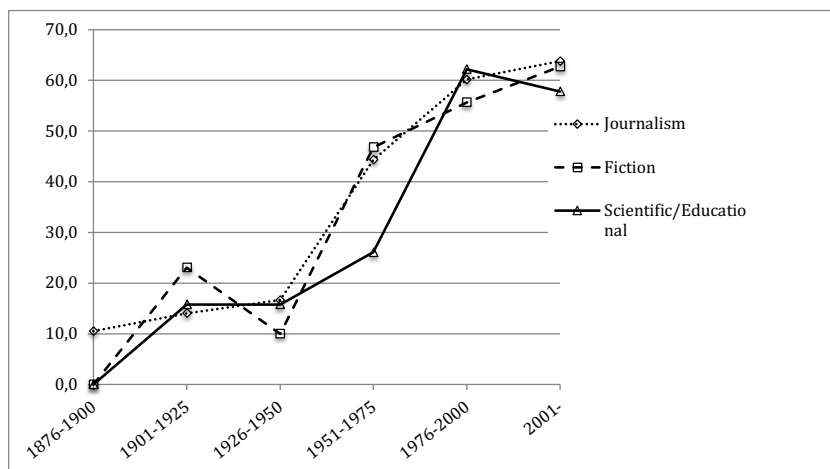
Figure 2. Genre – the proportion of examples with the accusative construction for different periods (based on Table 4)

Table 4 and Figure 2 indicate that the three genres develop in parallel. The total numbers in the bottom row of the table show that journalism has a somewhat higher percentage of examples in the accusative (55.3%) than fiction (44.1%) and scientific/educational (46.1%), but statistical analysis indicates that the observed differences are not robust. Although Pearson's chi-squared test for total numbers yields a p-value that is small enough to indicate significant differences, Cramer's V-value, which is a measure of effect size, is below the threshold of what is considered reportable.[8] We conclude that the data

in Table 4 do not provide support for the Sociolinguistic Differentiation Hypothesis, insofar as we have not been able to identify robust differences between the genres under scrutiny.

4.2 Test 2: Gender

In his classic study, Labov (1972: 303) showed that the "sexual differentiation of speech often plays a major role in the mechanism of linguistic evolution". It is therefore natural to expect gender differences in the use of the two decade constructions over time. Table 5 summarizes the distribution of the accusative and locative constructions for male and female authors born in different periods.[9] As shown, female authors are strongly underrepresented in the corpus, especially in the earlier periods. For this reason, we only plot numbers from 1876 for female authors in Figure 3, which shows the development over time (measured in percentage of accusative for each period). As can be seen from the figure, the two genders display parallel development. The total numbers reported in the bottom row of Table 5 at first glance suggest that female authors are more likely to use the accusative, but a Pearson's chi-squared test indicates that the observed differences are not statistically significant.[10] In other words, our data do not show any reportable differentiation between male and female language users and thus do not lend support to the Sociolinguistic Differentiation Hypothesis. In §6, we present a more sophisticated statistical analysis, which further corroborates the conclusion that gender is of limited importance for the diachronic development of the decade constructions in Russian.

Table 5. Changes in the use of the decade constructions for male and female authors (periods refer to the year of birth of the author; one example per author)

Period	Male authors			Female authors		
	#Acc	#Loc	%Acc	#Acc	#Loc	%Acc
1801–1825	2	29	6.5	0	3	N.A.
1826–1850	2	43	4.4	0	4	N.A.
1851–1875	15	69	17.9	0	7	N.A.
1876–1900	24	72	25.0	5	7	41.7
1901–1925	63	68	48.1	17	16	51.5
1926–1950	127	78	62.0	15	10	60.0
1951–1975	40	28	58.8	14	7	66.7
Total	273	387	41.3	51	54	48.6

---

[8] Pearson's chi-squared test (χ^2 = 19.575, df = 2) gives the p-value 5.6137e-05. Cramer's V-value = 0.09. An R-script detailing the statistical test is available in the Tromsø Repository of Language and Linguistics (TROLLing): http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE. It is important to note that statistical significance (p-value) and effect size (Cramer's V-value) are different measures. Simply put, statistical significance concerns whether it is likely that a result has occurred by chance, while effect size measures how

strong the impact of a factor is. Importantly, even if it is unlikely that a result occurred by chance, the impact of the relevant factor might not be very strong. Cramer's V value can theoretically vary from 0 to 1, but 0.5 is considered high, while 0.3 represents a moderate value and 0.1 a low value (cf. King & Minium 2008: 327–329).
[9] The corpus has the option of carrying out separate searches for male and female authors, but since this would have forced us to perform twice as many searches, we preferred to determine the gender of the authors on the basis of their names, which are provided by the corpus. Most Russian names (first names and last names) are gender specific, and for gender neutral names such as *A. Anikst* we determined the gender of the author by means of web searches. In Table 5, we present data according to authors' date of birth rather than date of text production. We also carried out a comparison based on date of text production, but since this comparison gave very similar results, these data are not included in the article. The fact that authors' date of birth and date of text production yield similar results is expected from the discussion in §3 above.
[10] Pearson's chi-squared test with Yates' continuity correction (χ^2 = 1.6437, df = 1) gave p-value = 0.2, which indicates no statistically significant differences between male and female authors. An R-script detailing the statistical test is available in the Tromsø Repository of Language and Linguistics (TROLLing): http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE.
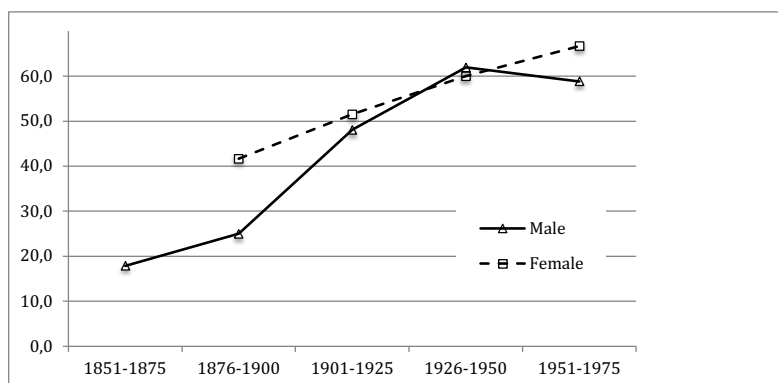
Figure 3. Gender – the proportion of examples with the accusative construction for different periods (based on Table 5)

4.3 Test 3: Individual speaker preferences

In order to find out if individual language users have different preferences, we took a closer look at all authors with four or more examples in our database.[11] The results are summarized in Table 6. The columns indicate to what extent the relevant language users use the accusative construction. In the second column from the left, which is labeled 'A(ccusative) 0%' we find 'locative lovers', i.e., language users who do not use the accusative at all. The next column is for authors who use the accusative in 0–24% of the examples, and so on until we come to 'accusative lovers' in the second column from the right, i.e., language users who only use the accusative. The rightmost column provides total numbers. The rows represent different time periods when the texts were created. Since numbers are too small to be meaningful in the first half of the 19th century, the table only gives numbers from 1851 onwards.

The total numbers in the bottom row of Table 6 indicate that both locative lovers and accusative lovers, as well as the four intermediate categories, are well attested. The locative lovers comprise the largest category (98 out of 344 authors), while all the other categories have between 38 and 59 authors. This shows that different authors show different preferences, so at first glance this seems to support the Sociolinguistic Differentiation Hypothesis, insofar as the two constructions seemingly survive to different degrees in different idiolects.

However, if we look at the distribution over time, a different picture emerges. For texts created in the 1800s locative lovers dominate completely, but in the 1900s the number of locative lovers decrease dramatically, and for texts produced after 2001 we have only three authors who exclusively use the locative construction. In other words, the authors who strongly prefer the locative represent have become a rarity.

The first 'accusative lover' is attested in texts from the period 1926–1950, and the numbers then show a steady increase until the period after 2001, where 29 out of 79 authors (36.7%) use the accusative only. In short, the accusative is gaining ground, while the locative is falling back.

Table 6. Author preferences in raw numbers – number of authors with different preferences for accusative vs. locative in different periods[12]

Period	0% A	1–24% A	25–49% A	50–74% A	75–99% A	100% A	Total
1851–1875	15	0	0	0	0	0	15
1876–1900	24	2	1	0	0	0	27
1901–1925	18	7	0	0	1	0	26
1926–1950	15	8	1	5	2	1	32
1951–1975	11	12	10	9	6	5	53
1976–2000	7	11	16	26	22	24	106
2001–2012	3	3	9	14	21	29	79
Total	98	43	38	54	52	59	344

The situation is presented graphically in Figure 4, showing that fewer and fewer authors use the locative construction consistently, while more and more prefer the accusative. While our results indicate that individual language users have different preferences, the results more than anything testify to the importance of time as the decisive factor for the rivalry between the accusative and locative constructions. In this way, Table 6 and Figure 4 lend stronger support to the Leveling Hypothesis discussed in §3 than to the Sociolinguistic Differentiation Hypothesis.
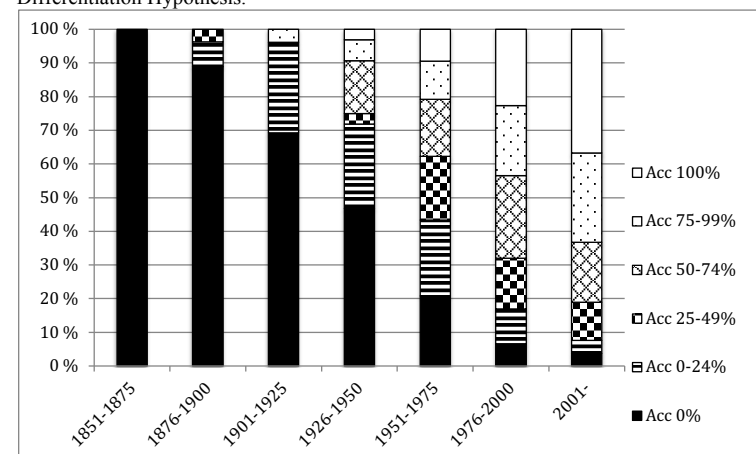

Figure 4. Author preferences in percentages (based on Table 6)

---

[11] We chose to focus on authors with at least four uses of the decade constructions in order to find out whether an author uses one or both constructions. A higher threshold than four would not have been feasible, since there are too few authors with many examples in our database to make statistical analysis possible.

[12] Periods refer to the time when the relevant documents were created (according to the Russian National Corpus). Some authors are represented in two periods; they are counted in both. For instance, on the basis of all examples by A.A. Axmatova we established that this author belongs in the category of 1–24% accusative. Since she has publications for both the 1926–1950 and the 1951–1975 periods, she was included as an author of the 1–24% category in both periods. Other ways of handling authors who were active in multiple periods are possible, but since only 14 out of 242 authors (about 5.8%) figure in two consecutive periods, different approaches are not likely to change the overall picture. We used the time when the relevant documents were created rather than the authors' year of birth. We do not have the birth year of all authors, so the year of birth would not facilitate robust statistical analysis.

4.4 Summing up

To summarize, we have not found strong support for the Sociolinguistic Differentiation Hypothesis – with regard to genre, gender or idiolect. It is of course possible that sociolinguistic differentiation is taking place but that our dataset is too small to capture it. It is furthermore possible that the Russian National Corpus does not facilitate investigation of the relevant sociolinguistic parameters – after all, besides gender, date of birth and name, we do not have any information about the social or geographical background of the authors. While the Russian National Corpus is a powerful tool for the analysis of grammatical constructions, the corpus does not facilitate analysis of geographical variation and has its limitations when it comes to performing analyses on the basis of other sociolinguistic variables.

5. Hypothesis 3: Semantic differentiation

As shown in §1, rival forms may develop different meanings over time. Although, as we have seen in the previous sections, the locative construction is losing ground while the accusative is gradually taking over, the locative is still used in about 40% of the examples in present-day Russian, so it is worth investigating whether the two constructions have developed different 'semantic niches' where they are preferred. A concrete proposal comes from Wade (1992: 453), who claims that "with decades, the accusative is preferred for processes extending over a period", while "the prepositional [locative] is preferred for an event occurring at a point within a decade". Accordingly, we formulate the following hypothesis:

(13)   The Semantic Differentiation Hypothesis:
         The decade constructions develop different meanings, such that the accusative is
         favored for processes extending over time, while the locative is primarily used about
         events that occur at a point within a decade.

We propose two tests for this hypothesis. Unfortunately, neither test lends robust support to the hypothesis.

5.1 Test 1: Perfective vs. imperfective aspect

Testing the hypothesis in (13) empirically is far from trivial. Our searches in the Russian National Corpus did not include parameters that would facilitate testing of the hypothesis directly, so extensive manual annotation of the dataset was required. While it is clear that the hypothesis refers to the aspectual properties of the events that take place within a decade, it is anything but obvious what kind of classification would form the best basis for testing the hypothesis. We decided to pursue two strategies. In the present section, we take morphological aspect as our point of departure, while we turn to a more detailed classification of aspectual properties in §5.2. The advantage of morphological aspect is that it can be determined objectively, but at the same time it represents a very course-grained measure. Conversely, a more detailed classification of aspectual properties offers a more fine-grained basis for testing the hypothesis, but at the same time it introduces a certain level of subjectivity, since the criteria for semantic classification must be interpreted in each individual example.

The rationale behind the use of morphological aspect, i.e., the distinction between perfective and imperfective verbs, is that the perfective aspect is used for events that occur at a point in time, whereas processes that extend over time typically are in the imperfective aspect. In more precise terms, definitions of the Russian perfective aspect often include reference to a boundary (juncture) that occurs at a point in time. For instance, Forsyth (1970: 8) writes that "a perfective verb expresses the action as a total event summed up with reference to a single specific juncture". Similar, but less explicit, are characterizations of the

perfective as the "emergence of a new state" (Russian: "*nastuplenie novogo sostojanija*", Padučeva 2010: 86) and "beginning of a new situation" (Russian: "*vozniknovenie novoj situacii*"; see Bondarko 2001: 153–159). Zaliznjak & Šmelev (2000: 35) state that a perfective verb always describes an "event" (Russian: "*sobytie*"), which they define as a change into a new state (Zaliznjak & Šmelev 2000: 35–36). Although these characterizations do not mention a temporal boundary explicitly, they refer to a change of state, which takes place at a point in time. Arguably, these and similar definitions may not cover ALL uses of perfective verbs in Russian, but it appears uncontroversial that they capture the most typical uses.

It has often been observed that the relationship between the two aspects is asymmetrical and that it is much more difficult to come up with a succinct definition of the imperfective aspect than the perfective (Zaliznjak & Šmelev 2000: 16–17). This has led some researchers, e.g., Jakobson (1984) and Forsyth (1970), to regard the imperfective as the unmarked aspect. While it should be acknowledged that the imperfective may denote events involving a change of state at a certain point in time, it appears uncontroversial that the imperfective TYPICALLY expresses activities or states (in the sense of Vendler 1957) that extend over time or repeated events that also extend over time (Zaliznjak & Šmelev 2000: 36).

Although the perfective-imperfective opposition may not be a perfect measure for Wade's (1992) distinction between 'processes extending over time' and 'events that occur at a point within a decade', we use it as a first test of the Semantic Differentiation Hypothesis. The results are summarized in Table 7 and Figure 5. Notice that the table reports on all examples for authors with birth dates and that we analyze one example per author.[13] We did not include examples where the event was not expressed by a verb, e.g., *terror v tridcatyx godax* "terror in the thirties", since in Russian only verbs display a consistent opposition between perfective and imperfective aspect.[14] A small number of so-called biaspectual verbs were also excluded, because as such verbs do not have a morphological perfective-imperfective opposition. Finally, we excluded verbs in the present tense, since in Russian the imperfective-perfective opposition is neutralized in the present tense where only imperfective verbs are used.

---

[13] Since the investigation of aspect required manual annotation, it was not feasible to use the entire database. We decided to use the subset for which we have the birth year of the author as a subset. Recall from §3 that authors' date of birth and date of authorship give the same results, so the decision to use birth year in the analysis of aspect is not likely to have affected the results. Our decision to analyze one example per author was guided by our wish to avoid biases due to idiosyncratic behavior of specific authors and thus give a representation of the situation in the language in general. As will be shown in §6, statistical analysis of a dataset with one example per author and the dataset without this restriction yields very similar results. In view of this, it appears unlikely that the limitation to one example per author has affected our results.

[14] Admittedly, the situation is more complicated for verbal nouns/nominalizations. However, although the noun *zakrytie* "closing" is derived from the perfective verb *zakryt'* "close" while *rabota* "work" and *žizn'* "life" are related to the imperfective verbs *rabotat'* "work" and *žit'* "live", verbal nouns do not show a morphological contrast between perfective and imperfective aspect, and such nouns were therefore not included. For detailed discussion of aspect in verbal nouns in Slavic, see Dickey (2000: 234–258).

Table 7. Development for verbs of imperfective and perfective aspect (all examples for which the corpus provides date of birth for the author; one example per author)

Period	Perfective verbs			Imperfective verbs		
	#Acc	#Loc	%Acc	#Acc	#Loc	%Acc
1776–1800	0	4	N.A.	0	4	N.A.
1801–1825	1	15	6.3	1	16	5.9
1826–1850	0	18	0	1	23	4.2
1851–1875	5	37	11.9	7	26	21.2
1876–1900	10	35	22.2	12	32	27.3
1901–1925	36	38	48.6	36	38	48.6
1926–1950	75	48	61.0	56	28	66.7
1951–1975	30	20	60.0	20	13	60.6
1976–2000	0	0	N.A.	3	1	N.A.
Total	157	215	42.2	136	181	42.9

The aggregate numbers for all periods given in the bottom row of the table show almost exactly the same proportion of accusative for perfective (42.2%) and imperfective verbs (42.9%). Clearly, these results do not support the Semantic Differentiation Hypothesis. If we consider the development over time, the use of the accusative displays a parallel increase for both aspects. This is shown in Figure 5, where the solid curve representing perfective verbs and the dashed line for imperfectives are very close for all periods. Once again, we conclude that the perfective-imperfective distinction does not lend support to the Semantic Differentiation Hypothesis.
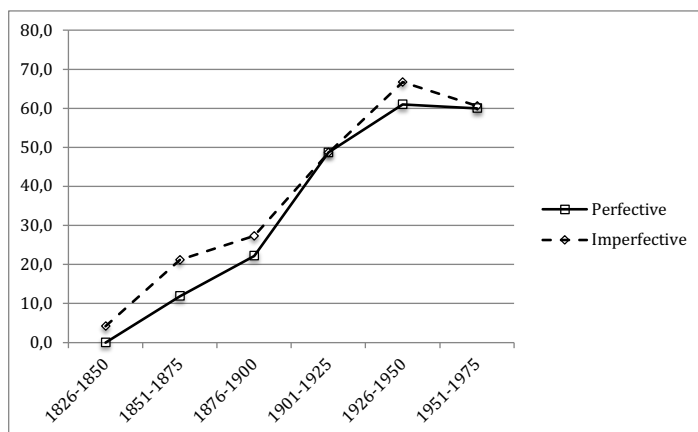

Figure 5. Aspect – the proportion of examples with the accusative construction for different periods (based on Table 7)

5.2 Test 2: Aspectual types

In order to carry out a more detailed test of the Semantic Differentiation Hypothesis, we tagged our database manually for ten different categories presented. Included are all examples

created after 1976, which suffices to give a good overview of the situation in present-day Russian. In order to avoid skewed results due to individual author preferences, we classified only one example per author.

'Accomplishment' is used in the Vendlerian sense (Vendler 1957) about events such as *postroit'* "build" in example (14), which take time and culminate in a change of state (the house being completed):[15]

(14) Ètot dom *byl postroen*$_{PF}$ … **v desjatyx**$_{LOC}$ **godax**$_{LOC}$. [Višneveckaja 1997]
"This house *was built* **in the 1910s**."

'Activity' is also used in the sense of Vendler about verbs like *rabotat'* "work", which do not culminate in a change of state:

(15) Každyj iz nas **v tridcatye**$_{ACC}$ **gody**$_{ACC}$ *rabotal*$_{IPF}$ po-svoemu. [Kozincev 1940–1973]
"**In the thirties**, each of us *was working* in our own way."

In our database, we have numerous examples of the following type, where the beginning of something occurred at a point in time within a decade and was followed by a state or activity that extended over time:

(16) **V semidesjatye**$_{ACC}$ **gody**$_{ACC}$ *načali*$_{IPF}$ *èmigrirovat'* moi druz'ja. [Dovlatov 1984]
"**In the seventies**, my friends *started to emigrate*."

We decided to single out such examples as a separate category, since they involve both something taking place at a point in time (the beginning itself) and the resulting state or activity, which extends over time.

In the literature on Russian aspect, the term 'general factual' is used about situations where an imperfective verb is used "simply to confirm the occurrence of an action" (Dickey 2000: 95):

(17) «Segodnja my pereživaem vremja, — *otmečal*$_{IPF}$ **v 30-e**$_{ACC}$ **gody**$_{ACC}$ poèt-filosof Indopakistana Muhammad Ikbal (1877–1938), — sxodnoe so vremenem protestanskoj revoljucii v Evrope». [Istorija vostočnoj filosofii 1998]
"Today we are going through a time, *remarked* the Indo-Pakistani philosopher Muhammad Ikbal (1877–1938) **in the thirties**, that resembles the protestant revolution in Europe."

We decided to treat the general factual as a separate category since it goes against the general pattern described in the previous section whereby events occurring at a point in time are expressed by perfective verbs. In (17), we are most likely dealing with one single remark at one point in time, but the imperfective aspect is nevertheless used.

Constructions with modal predicates such as *moč'*$_{IPF}$/*smoč'*$_{PF}$ "be able" and *prixodit'sja*$_{IPF}$/*prijtis'*$_{PF}$ "have to" are difficult to include in the other categories, and we therefore singled them out as a separate category:

(18) Mne *prixodilos'*$_{IPF}$ neodnokratno vstrečat'sja s nim i pozže, **v 60-x**$_{LOC}$ **godax**$_{LOC}$. [«Soldat udači» 2004]

---

[15] Since this section explores the relationship between the predicates in the sentence and the temporal adverbials for the convenience of the reader we use italics for the predicates and boldface for the temporal adverbials. We have added subscript indexes for imperfective ($_{IPF}$) and perfective ($_{PF}$) verbs. Only verbs relevant for the argument are supplied with indexes marking aspect.

"I *had to* meet him several times after that, **in the sixties**."

Examples with non-verbal predicates (nouns or adjectives) instead of verbs were also treated separately:

(19) Fanxio (Argentina) **—** pjatikratnyj *čempion mira v gonkax «Formula-1»* **v 50-e**$_{ACC}$ **gody**$_{ACC}$. [Karapetjan 2000–2002]
"Fangio (Argentina) is the five-time *Formula 1 champion* **in the fifties**."

Since events that in Wade's (1992: 453) words occur "at a point within a decade" are of particular relevance for the Semantic Differentiation Hypothesis, we treat punctual events of the following type as a separate category:

(20) Kirill *rodilsja*$_{PF}$ **v 30-x**$_{LOC}$ **godax**$_{LOC}$ XII veka. [«Žurnal Moskovskoj patriarxii» 2004]
"Cyril *was born* **in the thirties** of the 12th century."

'Repeated event' is used as a label for situations where the same event takes place several times, as in the following example:

(21) **V 80-ye**$_{ACC}$ **gody**$_{ACC}$ XVIII veka on dvaždy *izdaval*$_{IPF}$ polnoe sobranie ego sočinenij. [«Vestnik SŠA» 2003]
"**In the eighties** of the 18th century it *published* his collected works twice."

We use the term 'state' in Vendler's sense about situations with no dynamism that are stable over time:

(22) **V 20-e**$_{ACC}$ **gody**$_{ACC}$, naprimer, *suščestvoval*$_{IPF}$ obyčaj, po kotoromu xozjajka mogla priglasit' gostej … [«Vokrug sveta» 1989]
"**In the twenties**, for instance, there *existed* a custom according to which a hostess could invite guests …"

Our final category 'summarized repeated' captures examples where a number of sub-events are summed up and 'packaged' as one event by means of a perfective verb:

(23) **V 90-x**$_{LOC}$ **godax**$_{LOC}$ na rossijskuju džazovuju scenu *vyšlo*$_{PF}$ novoe pokolenie molodyx trubačej. [«Rossijskaja muzykal'naja gazeta» 2003]
"**In the nineties** a new generation of young trumpet players *entered* the jazz scene."

In this example, we are dealing with a number of trumpet players whose careers started at different points in time, but the use of the perfective verb *vyjti* "come out" facilitates a construal of this as one event, which as a whole had an impact on the jazz scene in the nineties.

These ten categories form a fine-grained basis for testing the Semantic Differentiation Hypothesis. As mentioned above, the classification involves a certain degree of subjectivity, but we have based ourselves on well-understood Vendlerian categories, which were augmented by some categories of particular importance for the hypothesis we are testing. Nevertheless, even with a fine-grained classification that was fine-tuned for the purposes of testing the Semantic Differentiation Hypothesis, our results summarized in Table 8 and graphed in Figure 6 do not offer robust support for the hypothesis. The accusative construction is used in about 60% of the examples for all ten categories. In the figure, we have organized the ten categories according to an increasing percentage of the accusative from left to right. As the figure shows, the differences are small.

Table 8. Distribution of accusative and locative with different aspectual types (examples from texts created after 1976)

Period	#Accusative	#Locative	%Accusative
Accomplishment	214	145	59.6
Activity	97	50	66.0
Beginning	153	118	56.5
General factual	25	15	62.5
Modal predicate	24	13	64.9
Non-verbal predicate	55	34	61.8
Punctual	179	139	56.3
Repeated event	117	51	69.6
State	189	94	66.8
Summarized repeated	51	34	60.0
Total	1104	693	61.4

Admittedly, punctual events receive a somewhat lower score for the accusative (56.3%) than the remaining categories, which on its face looks like weak support for the Semantic Differentiation Hypothesis. However, although statistical analysis indicates that the differences between punctual events and all other categories is statistically weakly significant, the effect size is below the threshold for what is reportable.[16] In other words, the observed differences do not provide robust support for the hypothesis.

At the other end of the scale, repeated events received the highest score for the accusative construction (69.6%). Again, this looks like weak support for the Semantic Differentiation Hypothesis, but statistical analysis shows that although the observed differences are weakly significant, the effect size is below the threshold for what is reportable.[17] In sum then, we have not managed to find robust support for the hypothesis.

---

[16] Pearson's chi-squared test ($\chi^2 = 4.3193$, df = 1) yields p-value = 0.04. Cramer's V = 0.05. (http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE)
[17] Pearson's chi-squared test ($\chi^2 = 5.2688$, df = 1) yields p-value = 0.02. Cramer's V = 0.05. (http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE)
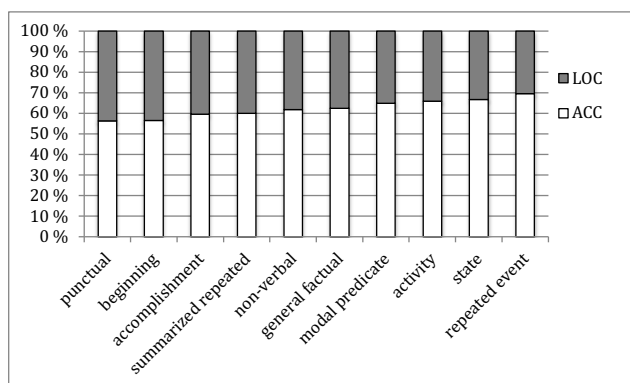
Figure 6. Distribution of accusative and locative with different aspectual types (percent, based on Table 8)

That we have not identified any semantic differentiation between the two Russian decade constructions does not preclude the existence of such differences. The tests we have suggested may not be sufficient to shed light on the Semantic Differentiation Hypothesis, and there may be other semantic differences than the ones we have been looking for. While a large electronic corpus like the Russian National Corpus is a powerful tool in the hands of a (historical) linguist, corpora have their limitations. Large datasets like the one we have analyzed in the present article do not lend themselves to in-depth semantic analysis of individual examples, since reliable automated tools for such analysis are not available. At present, there are therefore limits on how far a corpus can take us in pinpointing subtle semantic distinctions.

6. Interaction of the three hypotheses: CART and Random Forest analysis

How do the three scenarios of leveling, sociolinguistic differentiation and semantic differentiation interact? In order to find out, we carried out a CART (Classification And Regression Tree) analysis, as well as a Random Forest analysis. Our results reinforce the conclusion from the previous sections that leveling is the dominant scenario, although sociolinguistic differentiation (gender and genre) also plays a certain role.

CART and Random Forest (Strobl et al. 2009) are methods for statistical analysis of the interaction of predictor values, i.e., situations where a number of variables potentially influence the choice between two or more outcomes (e.g., the accusative or the locative case in Russian decade constructions). The methods were introduced to linguistics by Tagliamonte & Baayen (2012; see also Levshina 2015: 291), and Baayen et al. (2013) have shown that by most measures CART and Random Forest perform almost identically compared to traditional regression models.

Since manual annotation was required for variables pertaining to semantic differentiation, it was not feasible to use the entire database for the purposes of the CART and Random Forest analysis. For this reason, we decided to use the same partly manually tagged dataset as in §5, which is large enough to facilitate robust results.

We included the following independent variables (values in parentheses):

(24)   a. Author's year of birth (25-year periods; see §3.2)

  b. Genre (journalism vs. fiction vs. scientific/educational vs. other; see §4.1)
  c. Gender of author (male vs. female; see §4.2)
  d. Aspect (perfective vs. imperfective; see §5.1)
  e. Grammatical form of predicate (syntactic category and inflectional features)

Variables (24a-d) have already been discussed. Variable (24e) refers to the word that denotes the event that is located within the relevant decade. Once these words were annotated for aspect it was relatively easy to provide their syntactic category (noun, adjective, pronoun or verb) and specify their inflectional features. Since inflectional features have been shown to interact closely with aspect in Russian (Janda & Lyashevskaya 2011), we decided to include this factor in the analysis.

The results of the CART analysis are summarized in the tree diagram in Figure 7, based on the variables in (24).[18] Simply put, what the model does is to take the whole dataset and divide it into two subsets based on the variable that best predicts the choice between the accusative and locative in the decade constructions. Then the model splits each subset into further subsets based on the best predictor variable in each subset. This process is repeated many times until a better result cannot be obtained. However, the tree diagram in Figure 7 only represents the first three splits, which involve the most important variables. Each node represents one variable, and the branches from the nodes are labeled with the values that form the basis of the partitioning of the data. For instance, node no. 1 at the top of the diagram represents the authors' year of birth (which the Russian National Corpus somewhat misleadingly refers to as 'birthday', hence the label 'BirthdayPeriod' in the diagram). The left branch that connects node no. 1 with node no. 2 covers authors born after 1900, while the right branch leading to node no. 9 is about authors born between 1776 and 1900. For each node the model provides a p-value indicating statistical significance.

If we consider the left part of the diagram further, we see that the next split also refers to authors' year of birth. Below node no. 2 there is a split between authors born between 1901 and 1926 on the one hand and authors born later on the other. For authors in the former group, the model makes a further split between female and male authors below node no. 3. The terminal nodes (the 'leaves') in the tree are bar diagrams showing the proportion of accusative and locative in the relevant subparts of the dataset. Node no. 4 indicates a little less than 40% accusative (light gray shading) and more than 60% locative (dark gray shading) for male authors born between 1901 and 1925. Comparison of the bar diagrams in nodes no. 4 and 5 shows that female authors of the relevant generation are less likely to use the accusative than male authors. For each terminal node, the model provides in parentheses the number of examples covered by the relevant node. For instance, "n = 181" indicates that node no. 4 accommodates 181 examples altogether, while node no. 5 covers 689 examples. As mentioned in §4.2, female authors are underrepresented.

For authors born after 1925, node no. 6 indicates that genre is relevant. (Again, the corpus uses the *sfera* "sphere" to indicate genders, and 'sphere' therefore occurs in the label. 'Conflated' refers to the fact that we have collapsed some of the genres provided in the corpus, as explained in §4.1). Comparison of nodes no. 7 and 8 shows that the values 'fiction' and 'other' yield a somewhat lower proportion of the accusative than the values 'journalism' and 'scientific/educational'.

Let us now move to the right half of the tree, which pertains to authors born between 1776 and 1900. For this group, the model creates a split between the oldest authors and the

---

[18] The R script is available in TROLLing at
(http://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QKHCVE).

authors born after 1875. As shown from node no. 10, the latter group has a higher proportion of the accusative than older authors accommodated in nodes no. 11–13.

Taken together, the CART analysis corroborates the results reported in §3 through §5. By far the most important factor for predicting the choice between the accusative and locative in the decade constructions is time (here measured as author's year of birth). For authors born in the 18th and 19th centuries this is the only factor that appears in the diagram. For authors born in the 20th century the sociolinguistic variables of gender and genre are relevant to some degree in certain periods. Since aspect does not appear in the tree at all, the CART analysis does not provide any support for semantic differentiation.
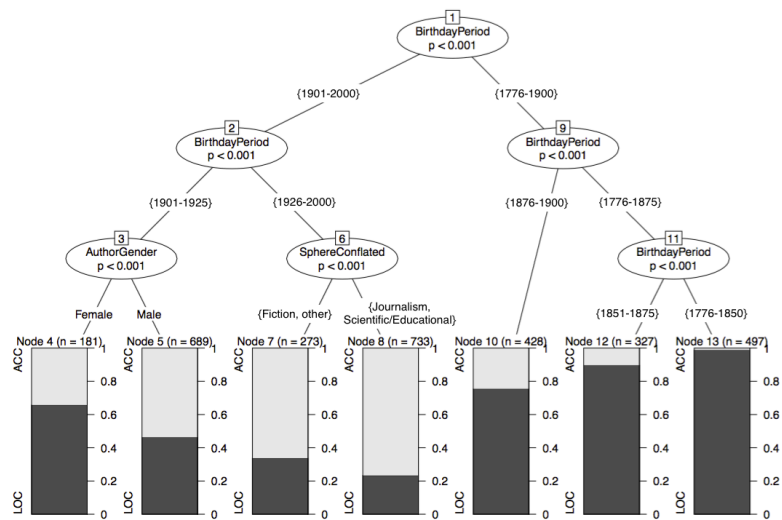

Figure 7. Decision tree from CART analysis of all examples for which the corpus provides the date of birth of the author (multiple examples per author)

The tree diagram in Figure 7 is based on all examples for which the corpus provides birth year for the authors. In order to control for the potentially confounding factor of having more than one example per author, we also performed a CART analysis for a dataset with only one example per author. As shown in Figure 8, this analysis reinforces the conclusions drawn above. The only variable represented in the tree, which has only three terminal nodes, is author's year of birth ('BirthdayPeriod'). In other words, it is clear that time is the best predictor of case in the decade constructions.
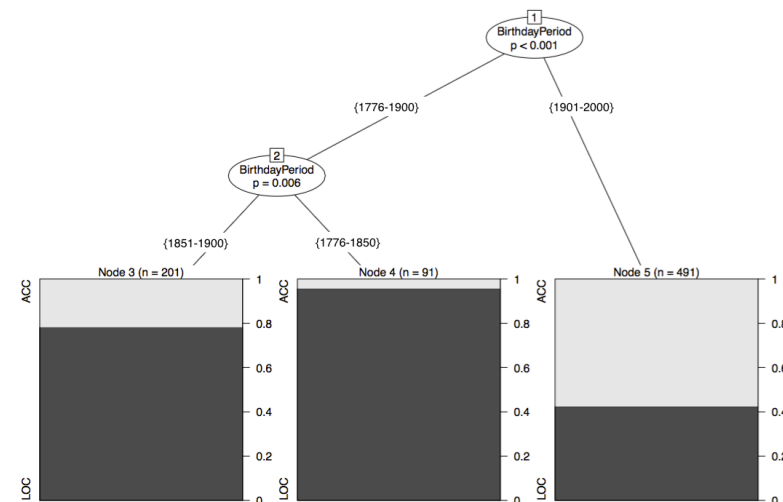

Figure 8. Decision tree from CART analysis of all examples for which the corpus provides the date of birth of the author (one example per author)

In order to measure the relative importance of all the variables in the model, we performed a Random Forest analysis (Breiman 2001). Random Forests is a validation technique that creates random subsets of the data and grows a tree for each subset. These trees are then compared, and the relative importance of the variables is measured. The results for our study are summarized in Figure 9. Each bar represents a variable, and the longer the bar of a variable, the more important the variable is. As shown in the figure, author's year of birth ('BirthdayPeriod') is by far the most important variable, while the other variables are of only marginal importance. This analysis is based on all examples with the author's year of birth. It was not necessary to carry out a Random Forest analysis for the dataset with one example per author, since, as shown in Figure 8, only one variable appears in the decision tree.
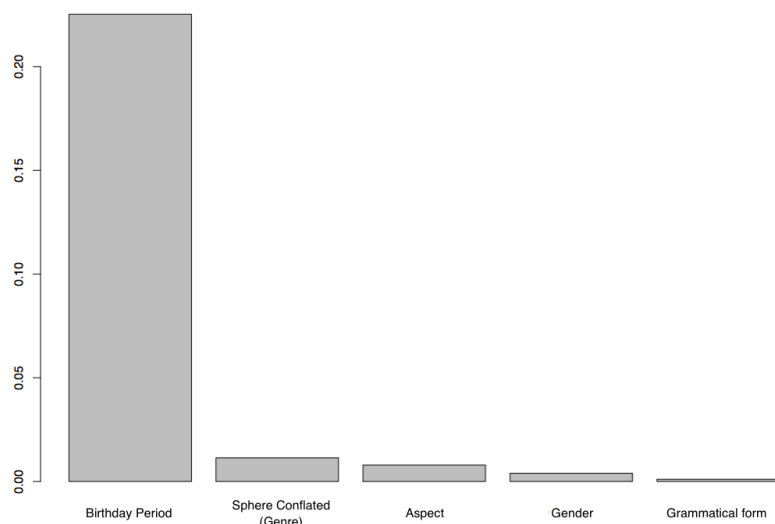
Figure 9. Random Forest analysis of all examples for which the corpus provides the date of birth of the author

7. Concluding remarks

Our detailed corpus investigation enables us to draw the following conclusions. First, we have shown that both constructions are well motivated by Russian grammar. The fact that the locative prevails for bounded time spans longer than a week motivates the use of the locative for decades, while the use of the plural in the decade constructions provides motivation for the accusative. Second, our data indicate that the distribution of the two constructions has changed radically over the past two hundred years. The accusative construction was marginal in the 19th century but showed a steep increase in the 20th century and is used in about 60% of recent examples. Third, we have found very little evidence for sociolinguistic differentiation with regard to gender, genre and idiolect. Fourth, contrary to claims in the literature, we have not been able to detect any semantic differentiation between the two decade constructions.

Although the present corpus-based study has not found strong evidence for sociolinguistic or semantic differentiation, other kinds of data might add nuances to this picture. For instance, it would be interesting to gather data from different geographical regions, and data from psycholinguistic experiments might give indications of sociolinguistic or semantic differences between the two Russian decade constructions. However, studies along these lines are left for future research.

Why are we witnessing the development documented in this article? While future research may demonstrate the relevance of sociolinguistic (external) factors, our research points toward language internal factors. As we pointed out in §2, the accusative is the default case in temporal adverbials with *v* "in(to)". In view of this, a possible explanation is that the development we have explored is part of a larger trend, whereby the locative is being marginalized in temporal adverbials with *v* "in(to)". If this explanation is on the right track,

we expect further marginalization of the locative in other temporal constructions, not just the decade constructions under scrutiny in the present study.

The historical development we have discovered and described follows the shape of an S-curve and therefore testifies to the importance of S-curves in diachronic linguistics. However, although we observe stabilization in the distribution of the decade constructions at the end of the 20th and the beginning of the 21st centuries, we cannot exclude the possibility that the growth of the accusative will continue in the future. In particular, the fact that we did not find conclusive evidence for sociolinguistic or semantic differentiation may indicate that the locative has not (yet) found a niche where it could survive and thrive in the future. At this point all we can do is appeal to the linguists of the future to revisit the decade constructions and find out whether the accusative will take over completely or whether the locative will be able to establish a niche for itself.

The present study has not only provided new insights about the diachrony of Russian temporal adverbials but has also demonstrated the value of corpus studies in diachronic linguistics in general. Although most corpora incorporate limited amounts of metadata and therefore have limitations, e.g., for the purposes of analyzing sociolinguistic and semantic differentiation, corpora provide us with detailed information about language change that could not have been obtained otherwise. Corpora have a lot to offer the historical linguist.

References

Baayen. R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37(3). 253–291.

Bjorvand, Harald. 2000. Diakron lingvistikk. In Rolf Theil Endresen, Hanne Gram Simonsen & Andreas Sveen (eds.), *Innføring i lingvistikk*, 307–339. Oslo: Universitetsforlaget.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.

Blythe, Richard A. & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88(2). 269–304.

Bolinger, Dwight. 1968. Entailment and the meaning of structures. *Glossa: An International Journal of Linguistics* 2. 119–127.

Bondarko, Aleksandr B. 2001. *Osnovy funkcional'noj grammatiki: jazykovaja interpretacija idei vremeni*. St. Petersburg: Izdatel'stvo SPbGU.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45(1). 5–32.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Carstairs-McCarthy, Andrew. 1994. Inflection classes, gender, and the principle of contrast. *Language* 70(4). 737–788.

Chambers, J.K. 2002. Patterns of variation including change. In J.K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *Handbook of language variation and change*, 349–372. Oxford: Blackwell.

Chambers, J.K. & Peter Trudgill. 1998. *Dialectology*, 2nd edn. Cambridge: Cambridge University Press.

Clark, Eve V. 1993. *The lexicon in acquisition*. Cambridge: Cambridge University Press.

Croft, William. 2000. *Explaining language change*. Harlow, UK: Longman.

Croft, William. 2001. *Radical construction grammar*. Oxford: Oxford University Press.

Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Dickey, Stephen M. 2000. *Parameters of Slavic aspect*. Stanford: CSLI Publications.

Forsyth, James. 1970. *A grammar of aspect: Usage and meaning in the Russian verb*. Cambridge: Cambridge University Press.

Fraser, Norman M. & Greville G. Corbett. 1997. Defaults in Arapesh. *Lingua* 103. 25–57.

Gilquin, Gaëtanelle. 2010. *Corpus, cognition and causative constructions*. Amsterdam: John Benjamins.

Goldberg, Adele E. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge: Cambridge University Press.

Haiman, John. 1980. The iconicity of grammar. *Language* 56(3). 515–540.

Harper, Douglas. 2001–2017. *Online etymology dictionary*. http://www.etymonline.com/index.php (October 17, 2016.)

Haspelmath, Martin. 1997. *From space to time: Temporal adverbials in the world's languages*. Munich: Lincom Europa.

Hock, Hans Henrich. 1988. *Principles of historical linguistics*. Berlin: Mouton de Gruyter.

Hovdenak, Marit, Laurits Killingbergtrø, Arne Lauvhjell, Sigurd Nordlie, Magne Rommetveit & Dagfinn Worren. 2001. *Nynorskordboka*. Oslo: Det norske samlaget.

Hudson, Richard A. 1980. *Sociolinguistics*. Cambridge: Cambridge University Press.

Iwata, Seizi. 2008. *Locative alternation: A lexical-constructional approach*. Amsterdam: John Benjamins.

Jakobson, Roman O. 1984. Structure of the Russian verb. In Linda R. Waugh & Morris Halle, *Russian and Slavic grammar: Studies 1931–1981*. Berlin: Mouton.

Janda, Laura A. & Olga N. Lyashevskaya. 2011. Aspectual pairs in the Russian National Corpus. *Scando-Slavica* 57(2). 201–215.

King, Bruce M. & Edward W. Minium. 2008. *Statistical reasoning in the behavioral sciences*. Wiley: Hoboken.

Klavan, Jane, Kaisa Kesküla & Laura Ojava. 2011. The division of labour between synonymous locative cases and adpositions. In Seppo Kittilä, Katja Västi & Jussi Ylikoski (eds.), *Case, animacy and semantic roles*, 111–134. Amsterdam: John Benjamins.

Klein, Wolfgang. 1994. *Time in language*. London & New York: Routledge.

Klein, Wolfgang. 2009. How time is encoded. In Wolfgang Klein & Ping Li (eds.), *The expression of time*, 39–82. Berlin: Mouton de Gruyter.

Labov, William. 1972. *Sociolinguistic patterns*. Oxford: Blackwell.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam: John Benjamins.

Makarova, Anastasia & Tore Nesset. 2013. Space-time asymmetries: Russian v "in(to)" and the North Slavic temporal adverbial continuum. *Russian Linguistics* 37(3). 317–345.

Nesset, Tore. 2004. Case assignment and image schemas: Russian temporal adverbials. *Studies in Language* 28(2). 285–320.

Nesset, Tore. 2013. How Russian became typologically unusual: The history of Russian temporal adverbials with v "in(to)". *Scando-Slavica* 59(1). 32–57.

Nesset, Tore & Anastasia Makarova. 2015. Prostranstvo vo vremeni? Asimmetrija predloga *v* v prostranstvennyx i vremennyx konstrukcijax. In Andrej A. Kibrik & Aleksej D. Košelev (eds.), *Jazyk i mysl': sovremennaja kognitivnaja lingvistika*, 388–410. Moscow: Jazyki slavjanskoj kul'tury.

Nida, Eugene A. 1958. Analysis of meaning and dictionary making. *International Journal of American Linguistics* 24(4). 279–292.

Nuyts, Jan & Pieter Byloo. 2015. Competing modals: Beyond (inter)subjectification. *Diachronica* 32(1). 34–68.

Padučeva, Elena V. 2010. *Semantičeskie issledovanija. Semantika vida i vremeni v russkom jazyke. Semantika narrativa*. 2nd edn. Moscow: Jazyki slavjanskoj kul'tury.

Russian National Corpus: Available at www.ruscorpora.ru

Saussure, Ferdinand de. 1983 [1916]. *Course in general linguistics*. Translated and annotated by Roy Harris. London: Duckworth.

Sokolova, Svetlana, Olga Lyashevskaya & Laura A. Janda. 2012. The locative alternation and the Russian 'empty' prefixes: A case study of the verb *gruzit'* "load". In Dagmar Divjak & Stefan Th. Gries (eds.), *Frequency effects in language representation*, 51–86. Berlin: Mouton de Gruyter.

Szymanek, Bogdan. 2005. The latest trends in English word-formation. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of word-formation*, 429–448. Dordrecht: Springer.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Vendler, Zeno. 1957. Verbs and times. *Philosophical Review* 66. 143–160.

Wade, Terence. 1992. *A comprehensive grammar of Russian*. Oxford: Blackwell.

Wulff, Stefanie. 2006. Go-V vs. go-and-V in English: A case of constructional synonymy? In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 101–125. Berlin: Mouton de Gruyter.

Zaliznjak, Anna A. & Aleksej D. Šmelev. 2000. *Vvedenie v russkuju aspektologiju*. Moscow: Jazyki russkoj kul'tury.

Résumé

Que peut faire un corpus pour le spécialiste en diachronie? Comment les données d'un coprus peuvent-elles jeter de la lumière sur le développement diachronique des formes soi-disant rivales, c.-à-d. des mots ou des constructions grammaticales qui paraissent synonymes ? Le présent travail aborde ces questions sur la base d'une analyse empirique détaillée de deux constructions russes, synonymes à première vue. Correspondant à la construction anglaise qui renvoie à la décennie, *in the twenties* "dans les années vingt", le russe possède deux constructions rivales, viz. *v dvadcatye gody* (lit. *in the twentieth years* "dans les vingtièmes années", avec l'adjectif numéral et le nom en accusatif), et *v dvadcatyx godax* (avec l'adjectif numéral et le nom en locatif). Trois hypothèses portant sur les formes rivales sont prises en considération : nivellement (où une des formes supplante la forme rivale), différentiation sociolinguistique (où les deux formes persistent, mais dans différentes variétés de la langue), et différentiation sémantique (où les deux formes rivales graduellement développent des significations différentes). Contrairement à ce qui s'est proposé dans les travaux antérieurs, nous trouvons peu de preuves de différentiation sémantique, ni ne trouvons-nous d'indications nettes de différentiation sociolinguistique. Nous démontrons plutôt l'existence d'un nivellement, vu que la construction accusative est en voie de supplanter son rival. Si notre étude montre que les données de corpus facilitent une analyse détaillée de l'interaction entre nivellement, différentiation sociolinguistique et différentiation sémantique, notre analyse révèle également certaines limitations, surtout en ce qui concerne les analyses, sur la base de corpus, de facteurs sociolinguistiques et sémantiques
.

Zusammenfassung

Was kann ein Korpus für historische Sprachwissenschaftler tun? Wie können Korpusdaten die diachrone Entwicklung sogenannter rivalisierender Formen erhellen, also von Wörtern oder grammatischen Konstruktionen, die scheinbar Synonyme sind? Dieser Artikel befasst sich mit diesen Fragen auf der Grundlage einer detaillierten empirischen Analyse zweier derartiger Konstruktionen im Russischen. Dieses besitzt zwei rivalisierende Konstruktionen, die der englischen „Jahrzehntkonstruktion" wie *in the twenties* entsprechen, nämlich *v dvadcatye gody* (wörtl. 'in den zwanzigsten Jahren' mit Numeral und Nomen im Akkusativ) und *v dvadcatyx godax* (mit Numeral und Nomen im Lokativ). Wir diskutieren drei Hypothesen hierzu: Nivellierung (eine Form verdrängt ihren Rivalen), soziolinguistische Differenzierung (die beiden Rivalen bestehen in verschiedenen Varietäten der Sprache fort) und semantische Spaltung (die beiden Formen erwerben im Laufe der Zeit unterschiedliche Bedeutungen). Im Gegensatz zur einschlägigen Literatur finden wir weder ausreichend Evidenz für eine semantische Spaltung noch eine soziolinguistische Differenzierung. Stattdessen demonstrieren wir hier den Fall einer Nivellierung da die Akkusativkonstruktion dabei ist ihren Rivalen zu verdrängen. Während unsere Studie zeigt, wie hilfreich Korpusdaten bei einer detaillierten Analyse des Verhältnisses zwischen Nivellierung, soziolinguistischer Differenzierung und semantischer Spaltung sein können, zeigt sie aber auch deren Grenzen auf, insbesondere bei der korpusbasierten Analyse soziolinguistischer und semantischer Faktoren.

Authors' addresses

Tore Nesset
Department of Language and Culture
UiT The Arctic University of Norway
NO-9037 Tromsø
NORWAY

tore.nesset@uit.no

Anastasia Makarova
Department of Language and Culture
UiT The Arctic University of Norway
NO-9037 Tromsø
NORWAY

anastasia.makarova@uit.no