How to measure change? An assessment of available Russian language corpora.

Involving a case study: the competition between nominative and instrumental case in predicate constructions with *byt'*.

A thesis presented for the degree of Bachelor of Arts
in Russian and Slavic studies

University of Amsterdam, August 26, 2021; Faculty of Humanities

Student: Damiaan J.W. Reijnaers (10804137)
Supervisor: **dr. A.V. Peeters-Podgaevskaja**
Assessor: **dr. E.R.G. Metz**

University of Amsterdam
Faculty of Humanities

Abstract

The present study was designed to assess the relevance of available language corpora to researchers investigating variations in language characteristics. The ongoing competition between the nominative and instrumental cases in Russian predicate constructions with the copula *byt'* 'to be' was adopted as a case study to capture the complexities of such research and its instruments. Traditionally, corpus-based methods have been employed by researchers for this purpose. This study has identified minor shortcomings with existing such research tools for written language and has thrown up important questions regarding the usage and construction of corpora in general. This thesis has shown that the scientific linguistic community is lacking appropriate spoken language corpora sufficient for investigating ongoing language change. It suggests that artificial intelligence could assist in the process of developing better means of research. This paper has further attempted to shed light on the relevant differences between written and spoken language but was purposely limited in terms of technical understructure. Consequently, it did not evaluate the current state of Russian language with respect to the employed case study. The study, however, found a strong indication that the nominative is preferred for spoken language, contrary to existing academic literature which is predominantly based on written language.

Contents

List of Tables

1 Introduction

Human natural language is an incredible phenomenon. It is highly dynamic and subject to continuous change (Keller 1994: 2). The changes that occur within it manifest *primarily* in spoken language, from which they eventually propagate to written language (Chafe 1985: 122).

There are a number of important differences between written and spoken language[1]. On a pragmatic level, spoken language distinguishes itself by being unplanned ('spontaneous') and informal, whereas written language is well-planned and 'polished to meet formal standards' (Redeker 1984: 44). This difference in 'spontaneity,' in turn, could result in deviations on a morpho-syntactic level. For instance, written language contains a higher number of complement and relative clauses, and participles (Chafe 1982: 44). Differences with respect to word order have also been observed (Smolka 2017: p. 59). In the same vein, Zemskaja (2011: 3) notes that Russian spoken language differs from its written variety, both at the linguistic and extra-linguistic level.

Change in a language can be studied by different means, such as conducting live interviews or social network analysis (Aitchison 2001: 42–51). Corpus linguistics is concerned with the analysis of linguistic phenomena in a real-world context based on extensive collections of authentic texts and utterances. The vast majority of recent studies in linguistics have been quantitative (Joseph 2008: 687). As such, many researchers investigating language change have utilised corpus-based methods (*e.g.* Janda, Nesset, and Baayen (2010) and Nesset and Kuznetsova (2015); rich in such studies are the periodicals *International Journal of Corpus Linguistics* or *Diachronica*). Zakharov (2013) outlines an overview of available corpora in the Russian Language, among which the *Russian National Corpus* (RNC) is most widely used.

However, spoken parts of these corpora, on which such studies are based, are either non-existent (*e.g.* *Uppsala Corpus* does not contain utterances); or relatively small (at the moment of writing, of all tokens in the RNC, only 1.4% of all tokens originate from utterances[2]); or (in the case of *RUSLAN*) consisting of utterances from only a single individual (Gabdrakhmanov, Garaev, & Razinkov 2019: 114–115) or from a relatively small group of people (for the *Odin Rechevoj Den'* corpus) (Asinovsky et al. 2009: 252–253). It is evident that for reasoning about the *development* of a particular language change, it is necessary that the utterances of a large group of people, produced at various points in time, are included in the analysis.

[1] In this thesis, the term 'utterance' is be used to refer to productions of spoken language, while 'text' refers to units of written language. In both cases, a 'token' refers to its smallest contained elements (most often a word).

[2] Nacional'nyj korpus russkogo jazyka, *Statistika*, `https://ruscorpora.ru/new/corpora-stat.html`, accessed on April 24th, 2021

Moreover, the definition of spoken language is not clear-cut; speech rather lives on a 'continuum' defined by the level of 'spontaneity' with which it is uttered (Tannen 1982). Gregory and Carroll (1978: 38–45) come to the same observation. Similarly, McCarthy and Carter (1995: 216) note that oral production can vary in its degree of being 'speakerly' or 'writerly.' In other words, there exists a difference between 'reading aloud a written text' and 'spontaneous conversational interaction' (Chafe 1982: 49). In what was one of the first attempts at gathering a Russian corpus, Zasorina (1977: 8) accounts for colloquial speech by incorporating scripts of drama plays. The spoken subcorpus of the more modern RNC consists almost exclusively of film transcripts and text that was 'written to be spoken' (Grishina (2010: 2953); Grishina (2006: 122)). All of the aforementioned are examples of *scripted* spoken language and are, for the purpose of mapping out the development of ongoing language change, 'linguistic specimens' of questionable appropriateness (Alvarez-Pereyre 2011: 66).

This thesis assesses the relevance of currently available Russian spoken language corpora to studies into the development of ongoing language change. A case study approach is employed to quantitatively characterize the suitability of available corpora for research of such kind: the ongoing competition between instrumental and nominative cases (hereafter to be abbreviated as `INS` and `NOM`, respectively) in the context of nouns and predicate adjectives in combination with the copula *byt'* 'to be' in the Russian language. Therefore, this thesis may serve as a follow-up study of the work by Krasovitsky, Long, Baerman, Brown, and Corbett (2008), in which a thorough analysis is already presented on this subject. Their research was conducted on the basis of statistical analysis on solely written texts, provided mainly by A. A. Barentsen (University of Amsterdam). This thesis investigates whether the paper's results are consistent with analysis on spoken (sub)corpora and expands upon the work by extending the scope to adjectives, and by increasing the measured time period to 2020, in an aim to further lay out the development, as the work concludes that the change is "at an advanced stage" (Krasovitsky et al. 2008: 113).

Furthermore, in an attempt to address the introduced limitations of existing spoken corpora, this thesis proposes an architecture for a new type of language corpus that is inferior in quality but larger in size and the relative proportion of spontaneous speech contained within. Nowadays, the internet is rich in freely available sources that comprise not only written texts (*e.g.* blogs and social media), but also spoken texts (*e.g.* video sharing platforms). Using technology from the field of artificial intelligence, the audio component of clips that are originally without subtitles can be automatically transcribed. Their transcripts can then be linguistically analyzed and used as the contents of a language corpus. This thesis can thus be viewed as having an almost interdisciplinary character, as it contributes to the linguistic community by examining its research tools and proceeding research on a language change; and therewith also describes (and provides) the implementation of a new type of spoken language corpus, of which the conducted case study functions as its evaluation.

The next section of this paper will establish a theoretical framework for the mentioned case study. The third section is concerned with reproducing and elaborating on the work of Krasovitsky et al. (2008) and is composed of two subsections: it first concerns the methodology used for the corpus-based study; and will then go on by outlining and analysing the obtained results. The fourth section likewise presents a methodology and its accompanying results, but now regarding the newly introduced spoken language corpus. The remaining part of the paper proceeds with drawing conclusions on the obtained results and discussing the implications and limitations of corpus-based methods for research on developing language change.

2 Background: A tale of competition between Instrumental and Nominative cases in the Russian predicate.

"Esli drug okazalsja vdrug i ne drug, i ne vrag, a – tak" (if a friend suddenly turns out to be neither a friend, nor an enemy) – a phrase that is acknowledged by many, but *plain ungrammatical*. The copula *okazat'sja* 'to turn out (to be)' only allows the instrumental case for the nominal part of the predicate it connects with the subject of the sentence. Had the introductory quote instead contained the copula *byt'*, allowing *both* the nominative and instrumental cases, the familiar-sounding opening line would have raised a number of existential questions among former Soviet citizens. How *temporal* is a friend?; and, how *accidental* or *incidental* is a friend, who suddenly turns out not to be a friend – and not an enemy? What feelings do we experience; is an enemy *subordinate* to a friend? And, how would we all structure that syntactically; may we ever want to write that down? It is these questions—and many more—that Russian-speakers need to ask themselves when choosing the appropriate grammatical case for predicative nominals with the copula *byt'* 'to be'. At least, if we follow many researchers who have studied this phenomenon and attempted to formulate rules for its usage. Of course, Vladimir Vysotskij lays aside the mishmash of semantic rules and grammatical systems. Just like in the rest of his popular song, *"Pesnja o druge"* (1968), he chose the shortest form (marked by NOM) and favoured rhythm over linguistics.

Since the middle of the 18th century, research investigating the factors associated with grammatical case preference for predicative nominals has emphasized the semantic linguistic perspective. Numerous studies have attempted to formulate a consistent set of rules for case marking by defining lexical classes in which either INS or NOM dominates usage. The vast majority of these works touch upon the idea that INS conveys *temporalness*, *accidentalness*, or *instability*; while NOM indicates that features are *fundamental*, *constant*, or *stable* (*e.g.* Vostokov (1831: 244); Ovsjaniko-Kulikovskij (1902: 167); Peškovskij (1914: 224-226); Bulaxovskij (1958: 301); Timberlake (2004: 282–283)). This semantic difference is made explicit in (1).

(1) a. *My byli druz'jami, a potom vljubilis'*
we were friends.INS.PL and then fell-in-love
'We were friends, but then we fell in love...'

b. *Ne mogu skazat'*, *čto my byli druz'ja*
not can say that we were friends.NOM.PL
'I can't say that we were friends.'

However, on some occasions, resulting findings have the appearance of being inconsistent or, at the least, uncertain. For example, in 1788, Barsov explicitly mentions that nouns such as *otec* 'father', *djadja* 'uncle' and *mat'* 'mother' can not be declined in INS (Barsov, Tobolova, & Uspenskij 1981: 197); whereas Nichols (1981: 151–152) uses precisely these terms to indicate a lexical class—of kinship relations—that, contrastingly, prefers such declension (*e.g. brat'* 'brother'). Furthermore, Mixajlov (2012: 48–49) finds a plausible contradiction to the thesis developed by Potebnja (1888: 521) that states that INS excerpts a "feeling of the current state [of the subject] being hierarchically *subordinate* to other states" subjective to the speaker, and thus "different from the more objective notions of temporalness or accidentalness.[3]"

In the late 20th century, Nichols (1981: 4), who devoted a book to the topic and, as mentioned, introduced such lexical classes herself, stated that "one is unable to commit to a specific formalization". Indeed, as Krasovitsky et al. (2008: 113) later offered, in a 'landscape' of ongoing language competition, the 'pace of change' in a language might differ among its lexical classes. The authors themselves also observed different levels of instrumental case usage for different lexical classes. This insight makes, therefore, the aforementioned *clash* between Barsov and Nichols justifiable, considering that in the two centuries that have passed between their respective publications, speakers' language usage could have further converged. In this light, a formulation of rules defined on such classes is of subsidiary—practical—importance, and the uncertainty and instability encapsulated within these rules merely a natural consequence. In this way, the proposal offered by Krasovitsky et al. can also be viewed as rejecting the semantic approach. Remarkably, more than a century earlier, (Buslaev 1881: 264) already came to a similar conclusion: "It is evident that the nominative and instrumental cases alternate without any visible cause. However, for practical guidance, one can roughly note the instrumental case as marking something as 'non-essential', while the nominative case marks it as 'essential'.[4]"

[3] Original fragment: *'predčuvstviem" sopodčinennosti dannago sostojanija drugim" sostojanijam" [...] upomjanutaja sopodčinennost' dolžna byt' otličaema ot bolěe ob"ektivnoj vremennosti, slučajnosti, nevažnosti priznaka, kak" javlenie čisto ličnoe.'*

[4] Original fragment: *'Očevidno, čto padeži imenitel'nyj i tvoritel'nyj, pri glagolě byt'*, *(...), zaměnjajutsja odin" drugim", bez" vsjakoj vidimoj pričiny. Vpročem, dlja rukovodstva v" praktičeskom" otnošenii, slěduet" zamětit', čto tvoritel'nym imeni suščestvitel'nago označaetsja po bol'šej časti priznak nesuščestvennyj, (...), Imenitel'nym" že padežom" imeni suščestvitel'nago označaetsja priznak" suščestvennyj'*.

To determine the *effects* of this seemingly semantic origin, Krasovitsky et al. (2008) compared the occurrence frequencies of INS in predicate constructions involving different lexical classes. Besides stating that the source of the data was primarily provided by A. A. Barentsen, the authors make no attempt to fully define their employed method. Surprisingly, for the second half of the 20th century, INS was found to extend its scope beyond the lexical groups associated with 'temporalness.' Their statistical indicates that INS might serve as the 'final haven' for predicative nominals after a centuries-long journey through lexico-semantic rules. However, after likewise having conducted statistical analysis into the same matter, Kuznetsova (2013: 59) notes that this assertion does not apply to present tense copula contexts and makes the observation that this, again, translates into the distinction in terms of (in)stability (as the present tense inherently entails 'stability,' in contrast to expressions that relate the subject to a future or past state). However, she does not address the hypothesis of whether the present tense forms—exactly because of their semantic implications—might just be 'lagging behind' in competition and will eventually likewise be marked by INS. Although Kuznetsova subjects the analysis to considerable criticism, the findings by Krasovitsky et al. (2008: 58) at least indicate a 'trend,' favouring INS, on which Kuznetsova agrees regarding non-present forms. A further confirmation of this tendency towards an extended scope for contexts that 'allow' this paradigm is given by Mixajlov (2012: 42, 45–46). Regardless of whether predicates in Russian (with the copula *byt'*) will become a purely syntactic flag (requiring the INS declension paradigm), current-day Russian language teaching resources still mention semantic rules—may it be artefacts—similar to those aforementioned, *e.g.* Podgaevskaja and Honselaar (2018: 457–458) and Wade (2020: 125–126).

In other respects, research has also focused on aspects different from semantics. Bailyn (2012: 186-189) provides a purely syntactical approach and rejects that semantics is the 'driving force' of this change in grammatical case, based on the *possibility* of applying a syntactical approach and, moreover, the *necessity* of it in certain syntactical constructions. An example of such a restricted environment where attraction is motivated by syntactic structure alone is a *null-copula*[5] predicate (in which NOM is guaranteed). Bailyn further finds the contradictory semantic contexts of the constructs in which INS is syntactically required "undermining the semantic approach" (*e.g.* the aforementioned phrases containing a null-copula require declension based on NOM, even if belonging to the previously discussed lexical classes that prefer INS). He subsequently argues that semantic analyses are usually the result of morphological changes – not the cause. He explains the distinction in (1) by an unapparent difference between *primary* and *secondary predicates*[6]: as INS can never occur in primary predicates, Bailyn argues, *byli*, in (1a), cannot

[5] The copula *byt'* 'to be', is *null* (morphologically invisible) in its present tense conjugation.

[6] In *primary predication*, a relation is observed between a *subject* and its *predicate*, i.e. *on student* 'he is a student'; in *secondary predication*, a (non-infinitive) verb is related to exclusively nominal parts of the sentence. In *on kažetsja studentom* 'he looks like a student', *kažetsja* is a primary predicate, while *studentom* is a secondary predicate. In the first example, *on* and *student* are the main predicate's *arguments*; while only *on* is an argument of *kažetsja*. A secondary predicate, thus, provides information not about the subject but about an *argument* of the primary predicate. (Note that this resembles the *complement* of a *predicate phrase*; see footnote 7.)

fill the *head*[7] of the (primary) predicate, as its argument[8] is observed to be in `INS`. Following Bailyn (2012: 128, 182, 194–195), the head of an instrumental predicate phrase "assigns `INS` to its complement[9]" when the head of the predicate phrase is (morphologically) *null* (empty) – *druz'jami*, thus, *'must'* regard the *complement* of a secondary predicate with an empty head, while *byli* is the head of the primary predicate with *my* as its only argument. The example in (1b) results in a primary predication (and thereby 'triggers' `NOM`) as here *byli* acts as an *overt predicator*[10]. This is explained and further expanded upon in earlier work, (Bailyn 2002: 48), that builds on Bowers (1993: 633). Matushansky (2019) challenges Bowers' theory, while Madariaga (2008: 345) also casts doubt on Bailyn's reasoning on the mandatory assignment of `NOM` for constructions where *byt'* acts as an overt predicator and presents a contradiction to the thesis. Richardson (2007: 48) take issue with the contention that the difference in (1) can be entirely described by a relation between a predicate phrase and its head and bases his rejection on observations in other Slavic languages. An alternative syntactical interpretation can be found in Pereltsvaig (2001: 45–46), where a different strategy is offered, likewise based on the syntactic structure of a phrase. Pereltsvaig proposes a system based on *bare small clauses* (which take `NOM`) and *rich small clauses* (taking `INS`)[11], which are concepts originally developed in Moro (2000: 71); this system was further extended by Citko (2008). A more recently published study by Adger and Ramchand (2003) again rejects this bare–rich theory while advocating the aforementioned work by Bowers. Other works in the same spirit include for instance Matushansky (2010) and Den Dikken (2006).

[7] In structural syntax, a sentence is hierarchically composed of *constituents* based on their dependency relationships. *E.g.* the sentence *on student* 'he is a student' can be divided into a subject (*on*) and a predicate (*student*). A constituent is generally a *phrase* or a *clause* and can consist of multiple words. A phrase behaves as a 'single functional unit', *e.g.* a *noun phrase* functions as a single noun, but can, for example, contain determiners and (attributive) adjectives. In *on xorošij student* 'he is a good student', the noun phrase *xorošij student* is functionally equivalent to the same phrase without *xorošij*. *Xorošij* is an *adjunct*, as opposed to a *complement*, as it does not alter the grammatical functioning of the sentence. A clause *combines* a subject and a predicate, whereas phrases, by definition, can not. A clause can thus consist of multiple phrases. Generally, the **head** of a phrase determines its *syntactic function*. Potential other phrase elements are *dependent* on the head, *e.g.* in *xorošij drug* 'good friend', *drug* is the head of the noun phrase on which *xorošij* is structurally dependent. Note that *xorošij* could also fill the head of an (almost always adjunct) *adjective phrase*. Considering secondary predicates: primary and secondary predicates appear in the same clause; a secondary predicate is thus dependent on the main—primary—predicate; it *predicates* ('states something about') the primary predicate. In this sense, one could even view that *any adjunct* serves as a secondary predicate (or contains one). Relating the theory to the case study, a secondary predicate adjunct adjective phrase can be conjugated both in `NOM` and `INS`, as also mentioned in (Bailyn 2012: 178), *e.g. my videli ego* ***bodrym i zdorovym*** 'we saw him vigorous and healthy' (A. M. Dostoevskij, Vospominanija, 1896).

[8] See footnote 6.

[9] See footnote 7.

[10] When thinking only about a predicate's *arguments* (see footnote 6), a *predicator* is the *head* of what would otherwise be considered as the verbal predicate phrase (footnote 7). Here, *byli* is thus a predicator with *my* and *druz'ja* as its arguments. It is an *overt* predicator, as it is *visible*, unlike present tense forms of the verb *byt'*. In the latter case, the 'empty' form of *byt'* still behaves as a predicator, and thus prohibits the assignment of `INS`.

[11] A *small clause* is a structure composed of a subject and predicate (like any clause; see footnote 5) which "lacks tense inflection" (Citko 2011: 748). Pereltsvaig distinguishes *bare* and *rich* clauses by means of its structural embedding: if the *post-copular phrase* (the phrase *after* the copula *byt'*) consists of a determiner phrase (as opposed to a *noun* or *adjective phrase*) the small clause is 'bare,' while otherwise, it is 'rich.'

Although the syntactic approach does aid in guiding grammatical case choice, many of the cited works do not offer an explanation for the semantic differences that are genuinely *felt* by native Russian speakers. However, some works, (*e.g.* Geist (2019)), make an attempt to bridge this gap. The main aim of this concise literature overview was to outline that different theories exist regarding the grammatical interpretation of the differentiation in grammatical case in predicate structures and emphasize that the specifics are still subject to scholarly debate. This fact might, by itself, argue in favour of acknowledging the semantic approach.

Lomtev (1956: 92–93, 96–97), who was well aware of the different ways in which the phenomenon was addressed in the literature of his time, rejects both the semantic and grammatical approaches. He examines the ‘up-growth’ of the instrumental case on a more psychological level, and argues that the differentiation of both cases raised as a consequence of the need for clear communication, influenced by a speaker’s—subjective—point of view. Following his reasoning, case marking ‘expresses the speaker’s thought’ on whether certain properties (conveyed by the predicate) can be deemed as ‘essential’ to the subject. His rationale thus closely parallels the general consensus among scholars who take the semantic approach but discriminates itself by a difference in perspective. Mrazek (1964) and Černov (1986) build on his work. Furthermore, in his discussion, Lomtev also devastatingly criticises the earlier cited work by Potebnja (1888), who likewise takes a psychological stance: “It is completely unacceptable that the instrumental predicative is allegedly triggered by a personal and subjective ‘gut feeling’ of one state being *subordinate* to another.[12]” Although, as mentioned earlier in this section, Lomtev finds support in this from Mixajlov, his further assertion is questioned by Mixajlov (2012: 43). Lomtev states that “if the difference between NOM and INS would be a grammatical *fact*, then it must be the result of an extended historical process.[13]" Mixajlov correctly notices that Lomtev fails to specify *why* a ‘personal and subjective hunch’ can not possibly drive this change in grammatical case.

Having explored the semantic, grammatical and psychological approaches, it is important to note that some scholars view the competition from yet another perspective. Mixajlov (2012: 51–52) describes the *stylistic* factors associated with the choice between the two cases and observes that, at least at the end of the 18th century, stylistic marking propels case preference. In this realm, INS conveys a rather “neutral stylistic atmosphere.” He further notes that, originally, INS was “predominantly acceptable” in productions of ‘lower’ forms of literature, *e.g.* letters, legal documents and satirical prose. Nichols (1981: 154) suggests a similar line of thought. Moreover, prominent lexicographer Švedova indeed notes that the phenomenon of the considered case study is “stylistically not neutral” (Švedova 1980: 42). Moving on now to consider the

[12] Original fragment: *‘Soveršenno nepriemlemym (...) čto tvoritel’nyj predikativnyj vyzvan jakoby ličnym sub”ektivnym predčuvstviem sopodčinennosti dannogo sostojanija drugim sostojaniem.’*

[13] Original fragment: *‘Esli različie meždu vtorym imenitel’nym i tvoritel’nym predikativnym est’ grammatičeskij fakt, to on dolžen byt’ rezul’tatom dlitel’nogo istoričeskogo razvitija (...)’*

origins of the ongoing competition, Madariaga (2007: 110) suggests that exactly such stylistically motivated usage led to the introduction of lexical classes in which the nominal part of the predicate attracts INS. This indeed is a plausible claim as, for example, the mentioned legal genre strictly requires its formulations to be unambiguous. In certain contexts, assuming case agreement, the distinction between the nominal part of the predicate and the subject might become ambiguous, especially since the Russian language has a flexible word order. An example of such ambiguity is shown in (2). This could have paved the way for the the language change in question to take place. Peškovskij (1914: 246–247) provides a similar syntactic account on the origins of instrumental nominals, but brings up an additional contributing factor to the origin of the competition between the two cases: "a general tendency of Indo-European languages to replace parallel constructions into non-parallel ones.[14]"

(2) a. *Mal'čik voditelem byl akkuratnym*
boy.NOM.SG driver.INS.SG was careful.ADJ.INS.SG
'The boy was a careful driver.'

b. *Voditel' mal'čikom byl akkuratnym*
driver.NOM.SG boy.INS.SG was neat.ADJ.INS.SG
'The driver was a neat boy.'

c. *Mal'čik voditel' byl akkuratnyj*
boy.NOM.SG driver.NOM.SG was careful.ADJ.NOM.SG
???

So far this section has focused on the origins of the competition and the different perspectives from which it can be interpreted. This section will now turn to the history of how the instrumental case developed its prevalence in different parts of speech and tense forms. For this purpose, this thesis mainly draws on (Røed 1966). In a way, Røed's work can be regarded as a small corpus study, as he collects the literary works of various popular Russian writers, through different points in time, and conducts statistical research into the behaviour of the two case forms on the material. The time frame considered by the author spans roughly from the end of the 18th century to the middle of the 20th century. In this thesis, I will only consider past, future and present tense forms and do not deal with other tense forms such as infinitives and participles. Moreover, I will only consider *null*-copula present tense forms of *byt'*, instead of the usually omitted form *est'*. I further assume adjective–noun constructions to function as single nouns, as this is observed to be the case in syntactic structures.

We first consider nouns occurring in combination with a past tense conjugation of *byt'*: *byl*, *bylo*, *byla* and *byli*. Røed (1966: 20–22, 31) reports a total of 552 predicate cases which meet these conditions. Based on his data, he notices an increase in instrumental predicate marking from 38% to 75% measured before and after the year 1900. After dividing samples in lexical

[14] Original fragment: *'Opisannoe vytesnenie predikativnogo imenitel'nogo tvoritel'nym možno rassmatrivat' kak častnyj slučaj obščego stremlenii indoevropejskix jazykov zamenjat' parallel'nye konstrukcii neparallel'nymi.'*

groups, he observes an increase in instrumental case in *all* groups. The most significant increases were observed in nominal constructions that express essential properties (3% → 57%); constructions that rather express temporary, accidental properties (to which example (1) belongs; 34% → 64%); and constructions expressing the same treats as the former, but then related to occupations and activities (46% → 81%). In a fourth group, composed of samples containing abstract nouns, the instrumental case was already prevalent (85%), but slightly increased its dominance (to 95%). Krasovitsky et al. (2008: 106) notes that in the early 19th century, inanimate nouns favour INS. Animate nouns favoured INS when occurring in a context of *e.g.* restricted temporality. For future tense constructions, involving *budu*, *budeš'*, etc.; the sample of Røed is too small to draw any conclusions on language change development. The data indicate, however, that INS dominates this syntactic group both before and after 1900 (Røed 1966: 46). This is in line with Nichols (1981: 154), who states that "in the future tense (...) instrumental is volunteered first." Moreover, statistical analysis in Krasovitsky et al. (2008: 106, 107) confirms this observation: according to the author, from 1800 to 1850, 93% of future tense nouns attracted INS. A stark contrast is found with present tense forms – here, NOM is observed to be the only possibility across the entire timespan (Røed 1966: 42). This is consistent with the findings of many of the in this section cited semanticists and grammaticists. While Røed only considers texts dated before the middle of the 19th century, Krasovitsky et al. and Kuznetsova provide statistical findings up until the year 2000. Based on their studies, INS has gained even more terrain in the years after 1950 (Kuznetsova (2013: 58); Krasovitsky et al. (2008: 110)).

Interestingly, before the year 1900, INS was almost completely absent from past tense predicate adjectives; however, after the turn of the century, Røed (1966: 50–51) observes a significant increase in instrumental case marking—albeit smaller than for noun constructions—from 1% to 14%. Until the middle of the 19th century, NOM was still the most widely used form. Indeed, "the nominative adjective," as Peškovskij puts it, "resists the onslaught of the instrumental case longer than nouns.[15]" The same holds for adjectives in combination with a future tense form of *byt'*: although Røed's sample is again too small to provide a reliable insight in the development over time, it is obvious that NOM is preferred not only in old Russian, but also in his contemporary Russian language (Røed 1966: 68). These findings are consistent with (Šmelev 2002: 403), wherein it is noted that the predicative instrumental is "a relatively new phenomenon in Russian" which "does almost not occur in texts of 19th-century writers." Šmelev suggests that instrumental adjective predicates regard properties of the subject , which is in line with the reasoning in the above-mentioned semantic studies. Like substantives, present tense predicate adjectives do only allow declension based on NOM.

[15] Original fragment: *'Imenitel'nyj prilagatel'nogo (osobenno v kratkoj forme) dol'še protivostoit natisku tvoritel'nogo, čem imenitel'nyj suščostvitel'nogo.'*

The previous section has described the approaches used in the investigation into differentiation of instrumental and nominative cases in Russian predicate constructions. As the introductory paragraph sketched out playfully, the issue is still subject to considerable debate. However, all cited studies have in common that they at least *identify* the expansion of the contexts in which a predicative instrumental case may appropriately be used in the Russian language. In an attempt to elucidate the tools available as a basis for such studies, this thesis seeks to answer the following research question: *To what extent are Russian language corpora available and made accessible for adequate research into ongoing language change?* And, in extension to that; *To what extent are available data sufficient to construct a Russian language corpus capable of providing insight into ongoing language change?* This chapter has certainly shown that recent studies, some of which were corpus studies, confirmed a tendency that was already 'manually' recognized centuries earlier. Hence, it is likely that such corpora and/or data are freely and accessibly available to the field of diachronic linguistic research.

Research on the subject has been mostly restricted to texts and not utterances. The current study will therefore also examine a third research question: *How does the competition between nominative and instrumental cases in predicate constructions with the copula byt' manifest itself in spoken Russian language?* In the next two sections, attention will be given to the generalisability of the reported statistics to spoken language. In view of the implicit suggestion made in the first paragraph of the introduction, it could conceivably be hypothesised that spoken language shows an *even higher* tendency to attract `INS` in the considered predicate constructions.

3 Text and speech in the Russian National Corpus

The following part of this thesis moves on to explore the capability of the Russian National Corpus for adequate study on the competition between instrumental and nominative cases in predicate constructions with the copula *byt'*. By extending the study by Krasovitsky et al. (2008), as an additional objective, this chapter also attempts to examine the current position of the Russian language with regard to the mentioned case study. The section begins by discussing in greater detail the recent work by Kuznetsova (2013), wholikewise follows up on Krasovitsky et al., and is concerned with a similar aim. What follows is a description of a research method—alternative to that of Kuznetsova—and (an interpretation of) its corresponding yielded results.

3.1 Methodology

Kuznetsova (2013: 54) already reproduced the study by Krasovitsky et al. not only by using the same corpus but also by using the Russian National Corpus. However, one major drawback of her approach is that sentences were not included in the analysis if they did not *sequentially* contain the subject, copula, and predicate. Based on the statistical findings presented in this thesis, as much as ≈ 84,1% of all predicate constructions with the copula *byt'* 'to be' follow different word order patterns when measured solely on written texts (see appendix A.1). The phrases posed in (2) form an example of possible divergence. It is evident that there are plenty of such occurrences; consider, for example, the basic sentence *on byl'* ***moim*** *otcom* 'he was my father'. A 'real world example' is shown in figure 1. Another serious weakness is the method by which Kuznetsova derives her statistical results. She performs two corpus queries: a noun in `NOM` followed by a conjugation of *byt'*, followed by either (1) another noun taking `NOM`; or a noun declined in `INS`. Essentially, Kuznetsova simply compares the frequency counts of the returned results that meet the conditions of the query. As she is aware of the fact that the corpus' results can contain *noise*[16], she manually validates the first 100 results and obtains by that the frequency of noise she assumes to be present overall in the dataset. For obvious reasons, this method is not 'foolproof.' Although it may be clear that a *tendency* towards case development can be roughly estimated using her method, she explicitly mentions the relatively small size of the corpus used by Krasovitsky et al., while weakening her own research in a similar vein (Kuznetsova 2013: 52–53). Moreover, she mentions that "only one example per author was taken into consideration." Although it is unclear how she decides *which* example to choose, one could reasonably assume that it is (among) the first encountered example(s), as Kuznetsova only manually checks the first one hundred. A slightly better method would be to select an author's *most preferred* case; however, that would be impossible using the type of approach Kuznetsova uses, as that would require processing many thousands of results manually.

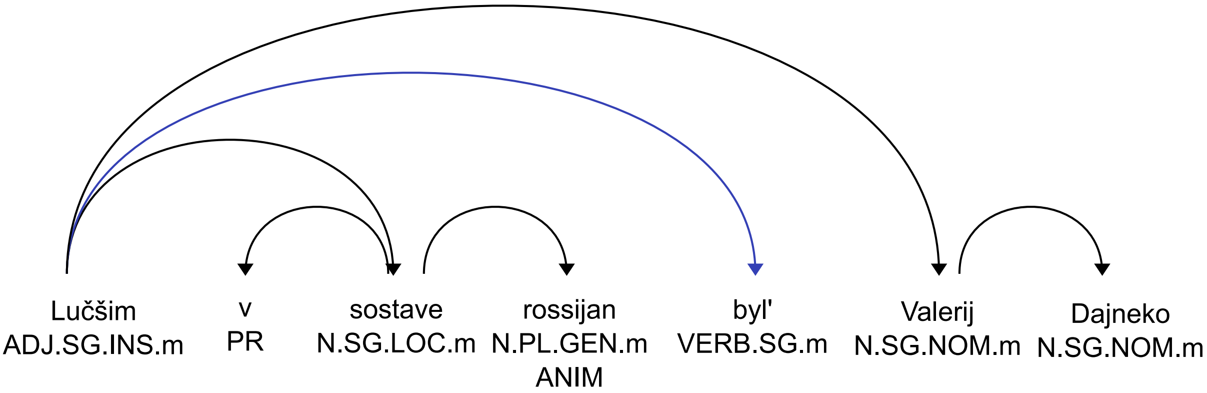

Figure 1: Visualisation of predicted dependency relationships: 'The best among the Russians was Valerij Dajneko'. Parsed on (a part of) a real sentence from the one-million Russian National Corpus database. Arrows denote the dependency relationships predicted by spaCy, while the bottom labels denote the manual annotations by the Russian National Corpus.

[16] With *noise* I mean 'unwanted data.' Data that technically meets the query, but is not consistent with the intention of the query. This could for example be an occasional occurrence of an instrumental form after the copula *byt'*, which does not represent the correct relation between subject and predicate.

To overcome these issues, I have contacted the non-profit partnership that runs the Russian National Corpus and obtained a smaller version of their database, containing about one million samples (Furniss 2013: 200). Instead of manual sampling, I employed an automatic dependency tree parser powered by *spaCy*[17]. A tree parser *predicts* dependency relationships between tokens in a sentence and by this means hierarchically decomposes the sentence into a *tree*-like structure, based on the syntactic dependencies between the words of the sentence. If the sentence was parsed correctly, the copula is *directly* subordinate[18] to the predicative nominal. This way, it can be determined whether there is a subject–predicate relation of interest, even if the tokens do not succeed each other sequentially. An example of a resulting dependency tree has been illustrated in figure 1. (Note that this tree does not resemble a constituency tree as usually observed in linguistics.). The accuracy of this method is about 95%.[19] However, as predicate structures connected by *byt'* are relatively straightforward to derive (precisely because of the overt presence of the copula), it is expected that accuracy is even higher in the considered experimental settings. To increase performance, all data were preprocessed before being 'fed' into the dependency parser (*e.g.* certain symbols were removed). See appendix A for a more detailed analysis and a manual inspection into the method's capacity for classifying these predicate constructions. As the samples from the RNC were manually annotated, the development over time of INS for different tenses and parts of speech could be derived. Note that in this thesis only adjective and noun predicative nominals are considered in combination with strictly overt realisations of finite forms of *byt'* (samples that did not contain a valid construction were discarded). Thus, all participles, short adjective forms (which always take NOM after all), infinitive forms, imperative forms, and other (tense) forms were not taken into account in this study.

3.2 Results and interpretation

The amount of data that the Russian National Corpus provides for 'offline usage' is small (about one million tokens; <1% of the online version). However, the most important limitation lies in the fact that this 'mini RNC' almost entirely consists of texts published after 2000. The result is that this approach does not prove useful in study on language change. The particular time window is, however, perfect for the aim of extending the analysis by Krasovitsky et al. (2008) to 2000 – 2020. Following their method, a distinction is made between (in)animate nouns. All results are reported in table 1 below. Precisely because Kuznetsova (2013) already provides the community with a replication and reproduction of Krasovitsky et al. (2008), it is also less necessary to consider *all* the time periods before the year 2000. Chapter 4 investigates time periods before 2000 solely for the purpose of assessing the quality of a self-constructed corpus.

[17] spaCy is a software development toolkit in the field of *national language processing* within artificial intelligence. The software and its source code are freely available: `https://spacy.io`, accessed on June 22th, 2021.

[18] To exemplify: in figure 1, *sostave* is dependent on *lučšim* (the predicative nominal) and *v* is dependent on *sostave*; here, *sostave* is *directly* subordinate to *lučšim* (just like the copula *byl'*), while *v* is indirectly subordinate.

[19] GitHub, *Release details of the large Russian model for spaCy*, `https://github.com/explosion/spacy-models/releases//tag/ru_core_news_lg-3.0.0`, accessed on June 22th, 2021.

A minority of documents was annotated not only by date of publishing but also by author's date of birth (57 out of 466 from the 'computer-written' texts). As expected, the small sample size of texts written by authors born before 1900 (only 1) neither allows adequate research into the development of a language. When considering the Russian National Corpus, one is thus forced to resort to methods such as employed by Kuznetsova for analysis into language change. Although such research can be effective, the in section 3.1 mentioned limitations are unfortunate.

	Observed grammatical case	
	Nominative	Instrumental
Past tense adjective	40.7% (122) / 22.7% (15)	59.3% (178) / 77.3% (51)
Past tense noun (animate)	31.4% (49) / 20.0% (10)	68.6% (107) / 80.0% (40)
Past tense noun (inanimate)	45.6% (130) / 42.9% (30)	54.4% (155) / 57.1% (40)
Future tense adjective	25.0% (10) / 17.4% (4)	75.0% (30) / 82.6% (19)
Future tense noun (animate)	11.8% (2) / 14.3% (1)	88.2% (15) / 85.7% (6)
Future tense noun (inanimate)	37.2% (16) / 40.7% (11)	72.8% (27) / 59.3% (16)

Table 1: Grammatical case of *byt'*-predicative nominals in sentences from the written part of RNC mini-corpus. Results are rounded to one decimal. Occurrence counts are denoted with parentheses. Small numbers (after the slash) show the results for which for every author only the most frequently observed case was recorded.

It is apparent from table 1 that a preference for INS in written language was observed in all cases. Future tense realisations of *byt'* preferred INS more frequently than past tense realisations. This observation is in harmony with the reported results in Krasovitsky et al. (2008: 107, 110). However, the differences are of strikingly lower magnitude, compared to the mentioned 87%+ frequency (during the most recent measured time period). It is highly unlikely that a reverse phenomenon (a shift back from INS to NOM) has taken place in only two decades (especially since none of the authors, for which the birth year was published, was born after 2000). Instead, the results highlight the inadequacy of the mini-corpus as a tool for investigating language change. The deviant results in table 1 can be partly justified as the median birth year of the authors was 1932 (average 1934.6), which is relatively far in time. Authors could possibly be still accustomed to 'old habits,' and thus more often prefer NOM in cases where later generations would already prefer INS. When correcting the data for author names (see appendix A), the results are similar. No significant differences were found between adjectives and nouns.

The results for the spoken subcorpus are shown in table 2. What is interesting about the data in this table is that, contrary to expectations, the data indicates that NOM prevails more often in spoken language. Surprisingly, NOM was favored over INS almost twice as often in future tense conjugations, in which, in written language, INS overwhelmingly predominated. However, when taking the authors of the texts into account, it becomes clear that the data sample is too small to draw definite conclusions. Again, the mini-RNC appears to be not suitable for investigating subtle changes within a language's development.

This section has identified minor shortcomings with the application of the RNC as a tool for research into language change when proceeding in similar manner as in (Kuznetsova 2013). It has further shown that an approach of the kind is the sole possibility when regarding the RNC. Notwithstanding the relatively limited sample, a surprising finding to emerge from the comparison between written and spoken language in the RNC was that usage of INS in spoken language did not increase in parallel with the tendency observed in written language.

	Observed grammatical case	
	Nominative	Instrumental
Past tense adjective	64.3% (72) / 50.0% (4)	35.7% (40) / 50.0% (4)
Past tense noun (animate)	45.7% (42) / 27.3% (3)	54.3% (50) / 72.7% (8)
Past tense noun (inanimate)	72.9% (78) / 70.0% (7)	27.1% (29) / 30.0% (3)
Future tense adjective	47.2% (17) / 0	52.8% (19) / 100.0% (2)
Future tense noun (animate)	25.0% (2) / 0	75.0% (6) / 0
Future tense noun (inanimate)	63.6% (28) / 57.1% (4)	36.4% (16) / 42.9% (3)

Table 2: Grammatical case of *byt'*-predicative nominals in sentences from the spoken part of RNC mini-corpus. Results are rounded to one decimal. Occurrence counts are denoted with parentheses. Small numbers (after the slash) show the results for which for every author only the most frequently observed case was recorded.

4 A custom-made speech corpus

In this section, an attempt is made to *construct* a Russian language corpus (thus, *without* resorting to pre-existing work). This way, one is not limited to restricted subsets of an original corpus (*e.g.* the mini-RNC-corpus considered in section 3). Additionally, one can exercise control over the contents of the corpus, instead of relying on less optimal approaches such as pursued by Kuznetsova (2013). For the reasons stated in the introduction to this thesis, the aim is to create a corpus that contains as much natural speech as possible. Furthermore, as this work is essentially an investigation into the *usability* of corpus-based methods in general (no matter whether they either need to be obtained from an existing source, or self-generated), a corpus should be possible to construct with minimal effort (and time). Therefore, this thesis proposes an even less extensive approach than expressed in the introduction: instead of describing the processes involved in *creating* or *utilizing* speech-to-text technology (which can be comprehensive), it proposes the use of readily available *output* of such technology, with the least effort required to obtain it. These outputs—the transcripts of speech—are henceforth processed in similar fashion to the method described in section 3. In less abstract terms: in this work, advantage is taken of the auto-speech-to-text module on YouTube by downloading the transcriptions of Russian-spoken videos. As these machine-transcriptions can be downloaded in bulk, a large-scale corpus could be generated in only hours. Data of this kind is also available from other sources; *e.g.* the *Russian open speech-to-text (STT) dataset* contains a wide variety of different types of transcribed speech and was made freely available (Slizhikova, Veysov, Nurtdinova, & Voronin 2021). In order to test the hypotheses with regard to 'spontaneity' that were implicitly sketched in the introduction, the aforementioned dataset is used to facilitate a comparison between semi-spontaneous speech, recorded in transcriptions of broadcasts on YouTube, and highly spontaneous speech, recorded from phone calls derived from the Russian SST dataset.

4.1 Methodology

Historic data was taken from only two channels: *Sovetskoe televidenie* and *Sovetskoe radio*, managed by *Gosteleradiofond Rossii*[20]. The video details were fetched using the *YouTube Data API*[21] using which, for every video, the transcriptions (subtitle tracks) were downloaded with *PyTube*[22]. The resulting transcriptions were converted to *SubRip (SRT)*[23]-format (examples are shown in listings 1–2, 4–11). After conversion, the transcripts were automatically annotated using *PyMorphy 2* (Korobov 2015), which was reported to have achieved accuracy scores of over 90% (Kotelnikov, Razova, & Fishcheva 2018: 139). Automatic annotation was performed *word by word*, as the considered text-to-speech technology also tends to process utterances on a word by word basis. Every 'packet' of words (or word group) is regarded as a 'stand-alone sentence.' No control can be exercised over the organisation of these sentences, as the grouping is done by the speech-to-text module (from which merely the output is extracted). Afterwards, the annotated transcripts were processed in the same manner as described in section 3. The transcripts do not contain any kind of punctuation marks and as such do not require any of the filtering methods outlined in appendix A.1.[24]

The television broadcasts were recorded between 1952 and 1997 while the radio recordings date from 1937 to 1998. Analysing this material not only serves the purpose of attempting to

[20] Gosteleradiofond, *Filial VGTRK Gosteleradiofond*, `https://gtrf.ru/`, accessed on August 5th, 2021. (associated YouTube channels: `https://www.youtube.com/channel/UCiVZttFkdEwMi3QXpRqFTzQ` and `https://www.youtube.com/channel/UCM6oyrdQzBf-egEmlkJyQNg`)

[21] An API is an interface that allows software developers to retrieve and post data. In this case, details (such as identification numbers and URLs) about the videos present in the playlist containing all of a user's videos are retrieved. See `https://developers.google.com/youtube/v3/docs/playlistItems` for technical documentation, accessed on July 27th, 2021.

[22] Like spaCy (see footnote 17), PyTube is a software development toolkit. PyTube provides functionality regarding the downloading of (components of) YouTube videos. Likewise, it is an open source project and is accessible at `https://pytube.io/`, accessed on July 28th, 2021.

[23] SubRip is an elementary and readable format for encoding subtitles. No technical background is needed to read and understand SubRip (`.srt`) files. For some videos, a more advanced format (encoded in XML) had to be converted to SubRip before processing.

[24] The only exception to this are sentences consisting of a single token of the kind *[muzyka]* or *[aplodismenty]*. These sentences are not considered anyway, as they can never contain a valid predicate construction.

replicate the earlier works by Krasovitsky et al. and Kuznetsova; as the type of medium is assumed to provide a higher level of 'spontaneity' (albeit slightly), more importantly, the results could possibly also give insight into whether spoken language is a better indicator of language change than written language. For this reason, the results were also compared to results obtained from utterances recorded from phone calls (which were assumed to offer speech with a higher degree of 'spontaneity'). For simplicity, the text (`.txt`) files contained within the data[25] (these hold the transcribed utterances) were isolated and merged into a single large file which was handled similarly to the YouTube subtitle transcriptions. Furthermore, in order to contribute to previous work by extending the measured time period to ≈ 2020, transcriptions from the last 20,000 videos uploaded by *Pervyj kanal*[26], Russia's most popular television broadcaster[27] were also analysed.

In the previous chapter, a distinction was made between the authors of texts as to prevent the results from being biased towards a single (group of) author(s) whose texts were more greatly represented in the overall dataset (see appendix A.1). In this chapter, no such distinction is made, as this would require more sophisticated technology, which would in turn defeat the purpose of examining the availability of language corpora as a tool for the average linguist.

4.2 Results and interpretation

The used historic sources (*Sovetskoe televidenie* and *Sovetskoe radio*) respectively yielded $7,878$ and $4,417$ transcriptions of videos[28] amounting to respectively $4,044,547$ and $2,952,870$ utterances containing $19,889,109$ and $16,218,741$ tokens in total. When using this method on a larger number of channels, a corpus can be created (with only minimal effort) that is orders of magnitudes larger than the RNC. Especially when considering modern material, it may be clear that the *potential* of automatic transcription technology for linguistic analysis may be huge (*e.g.* more than 500 hours of video material is uploaded on YouTube *every minute*[29] while Russian users are in the top five of most active users[30]). Of the $20,000$ videos analysed from

[25] Data downloaded from Microsoft, *Azure Open Datasets*, `https://docs.microsoft.com/en-us/azure/open-datasets/dataset-open-speech-text?tabs=azure-storage`, accessed on August 9th, 2021.

[26] Pervyj kanal (or 'first channel') can be reached on `https://1tv.ru`. The YouTube channel is found at `https://www.youtube.com/user/1tv`. Both sources accessed on August 9th, 2021.

[27] International Media Distribution, *Channel One Russia*, `https://web.archive.org/web/20140121045622/http://www.imediadistribution.com/news/partner-network-month-channel-one-russia`, accessed on August 9th, 2021.

[28] Not all videos on the channels contained downloadable transcriptions. This occurs when a video is not available to the public ('private'); when a video is blocked in a certain region (in this case, the Netherlands); or simply due to the absence of an automatically generated subtitle track. Videos without a clear marking of date of original publishing were excluded from the sample as well.

[29] TubeFilter, *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*, `https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/`, accessed on August 5th, 2021.

[30] Global Media Insight, *YouTube User Statistics 2021*, `https://www.globalmediainsight.com/blog/youtube-users-statistics/`, accessed on August 5th, 2021.

Pervyj kanal, 16,410 videos contained usable transcriptions, amounting to a total of 8,153,523 utterances consisting of 46,637,938 total tokens. All video material was observed to be heterogeneous in terms of genre, age, and environment of involved speakers and even in degree of 'spontaneity'. More importantly, this heterogeneous mixture of different factors that could affect case preferences in different ways was observed *within* individual items in the dataset as well. A single television or radio broadcast could feature different speakers from different locations and of different ages speaking spontaneously, while being interspersed with highly non-spontaneous scripted pieces of text, poetry, or music. This complicates the process of ensuring that the sampled data is well-balanced.

An obvious weakness of this method is that automatic transcriptions are error-prone. Considering the noise and the age of the material, the accuracy can be reasonably assumed to lie within 60% and 70%, based on Bokhove and Downey (2018: 9). However, the most important limitation lies in the fact that neither of the used sets of utterances were organized into sentences. As the data rather consisted of 'word packets,' with an average size of 5.2 words, long-distance syntactic relationships were neglected during the tagging process which resulted in invalid tokens being interpreted as the nominal part of the predicate (note that modern English has an average sentence length of 15–20 words (Cutts 2020: p. 23)). Two examples of such errors are shown in listings 1–2. Exemplary utterances from phone calls are outlined in listing 3. This phenomenon is especially problematic as the majority of predicate constructions with *byt'* do not follow a strict *subject–copula–predicate* word-order, as mentioned in section 3.1. Indeed, spoken language is less structured in general (Brown & Yule 1983: 15); however, the used data resembles just *word-by-word* transcriptions, packed together merely by a word's relative point in time of utterance (none of the transcripts contained *any* kind of punctuation or other means of sentence demarcation). As Biber, Johansson, Leech, Conrad, and Finegan (1999: 1039) observe: "conversation has no generally recognizable sentence-delimiting marks such as the initial capital and final period of written language." This limitation is so fundamental that it renders irrelevant any results that were based on unorganised data of this kind. After empirical inspection of the results, it is assumed that the majority of the samples were incorrectly classified. Therefore, it was not possible to justifiably extend previous work on the concerned case study. Consequently, the results obtained from analysing the videos by *Pervyj kanal* were omitted. For the same reasons, the results in table 3 do not show occurrence counts. In addition to the listings on the next page, appendix B lists a number of extra examples.

```
795
00:46:42,080 --> 00:46:48,650
vo vremja fevral’skoj revoljucii on byl

796
00:46:44,390 --> 00:46:52,460
junkerom i on i končalos’ ego dežurstvo i
```

Listing 1: Fragment from *Kul’tura russkogo zarubež’ja. Peredača 4. Sud’by i knigi (1990)* uploaded by *Sovetskoe televidenie*. As the predicate construction *on byl junkerom* is broken into two utterances, it is not recognized. Transcription is shown in SubRip-format.

```
71
00:03:50,760 --> 00:03:56,579
monetku v kopilku taksofona nu v svjazi s

72
00:03:54,569 --> 00:03:59,099
knižkoj byli pis’ma čitatelej

73
00:03:56,579 --> 00:04:03,269
byli otkliki i v svjazi s etim s etim
```

Listing 2: Fragment from *Vladyki bez masok. Žan Pol Getti - samyj bogatyj v mire. Čego oni bojatsja. Fil’m 2 (1970)* uploaded by *Sovetskoe televidenie*. Even though *knižkoj* is singular, it is selected as the nominal part of the predicate connected by *byli*. This error emerges from the way the transcription of this excerpt of spoken language is structured. Transcription is shown in SubRip-format.

```
nu my byli škol’nikami kul’turnymi
ty byla uverena čto eto davlenie budet ili čto
net ja budu sljuna prezidentom rossii golosujte za menja
a ty byl znaju takoj izdevaeš’sja bljat’
kogda syn prišel domoj ona byla uže sovsem
on byl kakoj to mužčiny ja tak vas ponjal
nu tol’ko samara kontrolirovat’ kak ona budet etim laborantom
my byli sčastlivy vmeste translirovat’ da
koroče ona budet xata devčonki
```

Listing 3: Example excerpts of phone call data from the Russian SST dataset. Every line represents a different sentence (from a different phone call). Most of the listed constructions are ungrammatical.

	Observed grammatical case					
	Sovetskoe televidenie		*Sovetskoe radio*		Phone calls	
	Nom	Ins	Nom	Ins	Nom	Ins
Past tense adjective	79.6%	20.4%	85.7%	14.3%	81.9%	18.1%
Past tense noun (animate)	51.0%	49.0%	55.8%	44.2%	58.9%	41.1%
Past tense noun (inanimate)	70.3%	29.7%	65.8%	34.2%	87.4%	12.6%
Future tense adjective	78.8%	21.2%	83.4%	16.6%	87.5%	12.5%
Future tense noun (animate)	55.5%	44.5%	45.4%	54.6%	59.4%	40.6%
Future tense noun (inanimate)	69.6%	30.4%	67.5%	32.5%	85.9%	14.1%

Table 3: Grammatical case of *byt'*-predicative nominals in utterances from videos on YouTube channels in blue (*Sovetskoe televidenie* and *Sovetskoe radio*) and from phone calls derived from the Russian SST dataset in yellow. Results are rounded to one decimal.

Despite the large number of incorrect classifications, it could be argued that the *ratio* between these (often false) observations of nominative and instrumental case usage still represents a notion of case preference. In *all but one* of the considered grammatical categories, for both the video and phone call transcripts, NOM was deemed the most appropriate choice. Only future tense animate nouns—which strongly preferred INS in written language with a frequency of above 85% (see table 1)—did show a slight preference for INS. Interestingly, the findings accord with the earlier observations presented in the previous chapter, which indicated that spoken language attracts the use of the nominative case. It can be assumed that the concerned YouTube videos (table 3) contain speech that is more 'speakerly' than the spoken subcorpus of the RNC (table 2). In turn, these videos are still beheld to be more 'writerly' than the utterances stemming from phone calls (table 3). Evidently, the written part of the mini-RNC (table 1) is the most 'writerly' of all the data considered in this work. This combination of findings from different data further supports the implication that NOM is still the dominating case for the considered predicate constructions in spoken language, as opposed to written language. If the results were to be taken as valid, these would indicate that spoken language 'lags behind' in manifesting language change, contradictory to the hypothesis developed in the introduction.

Improving structure and accuracy—as far as current technology allows—requires utilizing more advanced artificial intelligence methods, which, again, makes the method less suitable for the ordinary linguistic researcher. It is unfortunate that (for the same reasons) it was not possible to assess the possibility of the data being more heavily weighted towards the preferences of one or a few authors; therefore, it is unknown whether such bias was present in the reported results. It is worth to note that the constructed database does well in other use cases: the obtained data can support a 'speech search engine' which, when applied, could allow users to search for a (group of) Russian word(s) and retrieve audio fragments of (native) speakers pronouncing the word(s).

In spite of the limitations associated with processing the data, another interesting finding is that the utterances recorded from phone calls seem to contain much less *byt'*-predicate constructions than any other set of data. An average of 3.2 predicates per 10,000 tokens were observed (a number presumably even too high, as most of the selected predicates were deemed invalid), while the written part of the mini-RNC contained 9.1 predicates per 10,000 tokens (the dataset counted 925,058 tokens in total). Although this phenomenon might be partly attributed to the aforementioned lack of structural marking, the low frequency of predicate constructions may also be described as a characteristic of spoken language (Brown & Yule 1983: 17). Whilst most of the selected predicates were indeed deemed invalid, the videos from *Sovetskoe televidenie*—which were burdened by the same limitations—contained 7.7 predicates per 10,000 tokens, which strengthens the claim. However, this finding must still be interpreted with caution because the phone call transcriptions empirically seemed to be slightly better organised with respect to sentence demarcation.

This chapter has analysed the difficulties arising when the construction of a spoken language corpus is attempted. Remarkably, the findings of this chapter further corroborated the tendency towards NOM in spoken language, which was already observed in the spoken data provided by the RNC (see section 3).

5 Discussion

The main goal of the present research was to examine the tools available to the linguistic research community for investigating language change. The second aim of this study was to extend existing research by investigating the development of the change in preference for grammatical case in predicative copula constructions in the Russian language. Relevant to both purposes, this study also set out to determine to what extent the degree of 'spontaneity' of linguistic productions could affect quantitative results of research into language change. This thesis managed to provide a deeper insight into the landscape of Russian language corpora and the data available to construct them. However, as the scope of this study was purposefully limited in terms of technical aspects, the study did not succeed in properly assessing the current developments with respect to the adopted case study. As a consequence, the relevant differences between (spontaneous) spoken language and (not spontaneous) written language could not be examined as initially desired. Whilst this study did not confirm any effect of spoken language, the results, reported in tables 1–3, did partially substantiate contrary to the hypothesis that spoken language accelerates the effects of language change. As these observations may very well be specific to the employed study, I suggest that before any generalisable conclusions about spoken language are drawn, studies similar to this one should be carried out on different phenomena of language change.

On the accessibility of corpora and data This research has confirmed that a large amount of data is required to investigate language change. The freely available 'mini-RNC', the size of one million tokens, is deemed not sufficient in size. Moreover, this study has raised important questions about potential biases that might naturally emerge from working with large-scale sets of data that compose appropriate language corpora. The findings of this study indicate that it is generally hard to completely balance a corpus. Examples of properties that are hard to account for are an author's age (especially if a text or utterance was authored by multiple individuals); the environment within which a text or utterance was produced (including *e.g.* geographical regions); and the distinction between authors in general and their preferences (an author could prefer different grammatical forms in different contexts, while it is also generally hard to ensure that a collection of texts or utterances is not weighted towards a single or a few authors that have a higher presence in the overall dataset). Another challenge is that the type of medium from which the data are derived does generally not provide a stable notion of genre, context, or 'spontaneity' – *e.g.* in a radio show, a completely spontaneous conversation can be interrupted by a prerecorded piece of music. This again exacerbates the difficulty of constructing a well-balanced dataset to do research with. Furthermore, the most obvious finding to emerge from this study is that spoken language, by nature, is hard to transcribe, process, or analyse by automatic means. As spoken language lacks a significant amount of structure, it is challenging to construct a spoken language corpus capable of providing insight into language change.

Notwithstanding these limitations, the study suggests that the currently available tools for research into language change are *cumbersome* but *effective* to a sufficient degree. Their effectiveness has been indicated by, among others, Kuznetsova (2013), who confirms the long observed tendency of the nominal part of predicate constructions with the copula *byt'* 'to be' to shift from `NOM` to `INS`. Taken together, the findings in this paper support a strong recommendation to make available an entirely free to use Russian (spoken) language corpus, at minimum the same size and quality of the RNC. Linguistic research would highly benefit from a large corpus that could be freely obtained and flexibly used. Although the RNC is free to use, any means of research requires navigating through a predetermined search query interface. My thesis has shown that this is subject to limitations. In addition, I have explored several approaches towards the construction of Russian language corpora. Despite its exploratory nature, this study has offered insight into the availability of corpus-based research methods that are in reach of an ordinary researcher in the field of linguistics. I suggest that a greater focus on artificial intelligence could likely aid the production of a suitable corpus at a reasonable cost. Considerably more effort will be needed to construct a corpus of such kind, requiring a more sophisticated technical approach. The contribution of this study confirmed that the creation of an adequate Russian spoken language corpus is *possible* if sufficient energy would be put in. Delivering a state-of-the-art corpus would have been far beyond the scope of this thesis.

On spoken language and the case study Whilst this study did not conclusively confirm the tendency towards declining predicative nominals in NOM in predicate constructions with *byt'* in the Russian spoken language, it did partially substantiate the claim by means of three different analyses on different types of data. The present study has been one of the first attempts to examine the effects of the development of the regarded case study on spoken language. The findings of this thesis have implications for the understanding of how ongoing language change manifests itself in speech. Contrary to the formulated hypothesis, the results of this study indicate that, in spoken language, the 'default grammatical form' (in this case NOM) still prevails over newly preferred forms that have emerged in written language.

A possible explanation for these results may very well be found in *psycholinguistics*: as speech is *unplanned* and produced in *real-time* (Biber et al. 1999: 1048), it is possible to hypothesise that the mental processes involved in producing speech are less sophisticated compared to those involved in producing texts. In general, therefore, it seems that—possibly under the pressure of time—a speaker selects 'the most default form' whenever a construction arises in which a form is not 'typical' or 'routine,' such as is the case when two grammatical forms compete. In a comprehensive study on language production, Levelt (1989: 157) points out: "We must assume that the speaker has at his disposal a set of routine procedures that perform this (...) automatically for whatever the language requires" A further study with more focus on the psycholinguistic aspect of language change and speech production is therefore suggested.

Final notes Ironically, the major limitation of this paper itself was the almost exclusive focus on the main goal of the study. The study was mostly concerned with mapping the *possibilities* of doing research into language change; instead of adopting 'status quo research methods' and actually *researching* language change. As the findings with respect to the research methods were rather disappointing, it was impossible to subsequently employ these methods to assess the research question involved with spoken language and the introduced case study. If the debate on 'spoken language as a manifestor of language change' is to be moved forward, a better understanding of the phenomenon needs to be developed. Further research might, in addition to the aforementioned, also explore how the spoken subcorpus of the (full-version) Russian National Corpus compares to its written parts by conducting a similar research to Kuznetsova (2013). More generally, a natural progression of this work is to analyse whether a decent (spoken) language corpus could be generated when dropping the implicit requirement of the process being replicable for the average linguist. This could likewise provide inside in the debate on spoken language and the concerned case study.

References

Adger, D., & Ramchand, G. (2003). Predication and equation. *Linguistic Inquiry*, *34*(3), 325–359.

Aitchison, J. (2001). *Language change : progress or decay?* (3rd ed.). Cambridge [etc: Cambridge University Press.

Alvarez-Pereyre, M. (2011). Using film as linguistic specimen: Theoretical and practical issues. In *Telecinematic discourse* (pp. 47–67). Amsterdam / Philadelphia: John Benjamins.

Asinovsky, A., Bogdanova, N., Rusakova, M., Ryko, A., Stepanova, S., & Sherstinova, T. (2009). The ord speech corpus of Russian everyday communication "one speaker's day": Creation principles and annotation. In V. Matoušek & P. Mautner (Eds.), *Text, speech and dialogue* (pp. 250–257). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bailyn, J. F. (2002). Overt predicators. *Journal of Slavic Linguistics*, *10*(1/2), 23–52.

Bailyn, J. F. (2012). *The syntax of Russian*. Cambridge University Press.

Barsov, A. A. (1788). *Obstojatel'naja rossijskaja grammatika* (1st ed.).

Barsov, A. A., Tobolova, M. P., & Uspenskij, B. A. (1981). *Rossijskaja grammatika antona alekseeviča barsova*. Moskva: Izd-vo Moskovskogo universiteta.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written english* (1st ed.). Harlow: Pearson Education.

Bokhove, C., & Downey, C. (2018). Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, *11*(2), 1–11.

Bowers, J. (1993). The syntax of predication. *Linguistic Inquiry*, *24*(4), 591–656.

Brian Boniface, M., Cooper, C., & Cooper, R. (2006). *Worldwide destinations*. London: Taylor & Francis.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge university press.

Bulaxovskij, L. (1958). *Istoričeskij kommentarij k russkomu literaturnomu jazyku* (5th ed.). Kiev: Radjans'ka Škola.

Buslaev, F. I. (1881). Istoričeskaja grammatika russkogo jazyka. *Tip. T. Ris*, *2*. Retrieved from `http://elib.shpl.ru/ru/nodes/35996-ch-2-sintaksis-1881#mode/inspect/page/272/zoom/4`

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and written language: Exploring orality and literacy*, 35–54.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. *Literacy, language, and learning: The nature and consequences of reading and writing*, 105–123.

Citko, B. (2008). Small clauses reconsidered: Not so small and not all alike. *Lingua*, *118*(3), 261–295.

Citko, B. (2011). Small clauses. *Language and Linguistics Compass*, *5*(10), 748–763.

Cutler, A. (1982). Idioms: The colder the older. *Linguistic Inquiry*, *13*(2), 317–320.

Cutts, M. (2020). *Oxford guide to plain english*. Oxford: Oxford University Press.

Den Dikken, M. (2006). *Relators and linkers: The syntax of predication, predicate inversion, and copulas* (Vol. 47). Cambridge, MA: MIT press.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, *6*(1), 22–42.

Furniss, E. (2013). Using a corpus-based approach to Russian as a foreign language materials development. *Russian Language Journal*, *63*, 195–212.

Gabdrakhmanov, L., Garaev, R., & Razinkov, E. (2019). Ruslan: Russian spoken language corpus for speech synthesis. In A. A. Salah, A. Karpov, & R. Potapova (Eds.), *Speech and computer* (pp. 113–121). Cham: Springer International Publishing.

Geist, L. (2019). Copular sentences in Russian vs. Spanish at the syntax-semantics interface. *Proceedings of Sinn und Bedeutung*, *10*(1), 99–110.

Gregory, M., & Carroll, S. (1978). *Language and situation: language varieties and their social contexts*. London / Boston: Routledge and Kegan Paul.

Grishina, E. (2006). Spoken Russian in the Russian national corpus (RNC). In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 121–124). Genoa, Italy: European Language Resources Association (ELRA).

Grishina, E. (2010). Multimodal Russian corpus (MURCO): First steps. In *Proceedings of the seventh international conference on language resources and evaluation* (pp. 2953–2960). Valletta, Malta: European Language Resources Association (ELRA).

Janda, L. A., Nesset, T., & Baayen, R. H. (2010). Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling. *Corpus Linguistics and Linguistic Theory*, *6*(1), 29–48.

Joseph, B. D. (2008). The editor's department: Last scene of all... *Language*, *84*(4), 686–690.

Keller, R. (1994). *On language change: The invisible hand in language* (1st ed.). Florence: Routledge.

Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, & V. G. Labunets (Eds.), *Analysis of images, social networks and texts* (pp. 320–332). Cham: Springer International Publishing.

Kotelnikov, E., Razova, E., & Fishcheva, I. (2018). A close look at Russian morphological parsers: Which one is the best? In A. Filchenkov, L. Pivovarova, & J. Žižka (Eds.), *Artificial intelligence and natural language* (pp. 131–142). Cham: Springer International Publishing.

Krasovitsky, A., Long, A., Baerman, M., Brown, D., & Corbett, G. G. (2008, 01). Predicate nouns in Russian. *Russian Linguistics*, *32*(2), 99–113.

Kuznetsova, J. (2013). Diachronie distribution of predicate nouns in Russian. *Russian Linguistics*, *37*(1), 51–60.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lomtev, T. P. (1956). *Očerki po istoričeskomu sintaksisu russkogo jazyka*. Izd-vo Moskovskogo universiteta.

Madariaga, N. (2007). An economy approach to the triggering of the Russian instrumental predicate case. In *Historical linguistics 2005* (pp. 103–117). Amsterdam / Philadelphia: John Benjamins.

Madariaga, N. (2008). *Grammar change and the development of new case relations: The interaction between core syntax and the linguistic periphery in old and present-day Russian* (Unpublished doctoral dissertation). University of the Basque Country, Leioa.

Matushansky, O. (2010). Some cases of Russian. *Proceedings of FDSL*, *7*, 17–65.

Matushansky, O. (2019). Against the predp theory of small clauses. *Linguistic Inquiry*, *50*(1), 63–104.

McCarthy, M., & Carter, R. (1995). Spoken grammar: what is it and how can we teach it? *ELT Journal*, *49*(3), 207–218.

Mixajlov, N. (2012). *Tvoritel'nyj padež v russkom jazyke xviii veka*. Uppsala: Acta Universitatis Upsaliensis.

Moro, A. (2000). *Dynamic antisymmetry*. Cambridge, MA: MIT Press.

Mrazek, R. (1964). *Sintaksis russkogo tvoritel'nogo: strukturno-sravnitel'noe issledovanie* (Vol. 94). Prague: Opera universitatis purkynianae Brunensis, Facultas Philosophica.

Nesset, T., & Kuznetsova, J. (2015). Constructions and language change: from genitive to accusative objects in Russian. *Diachronica*, *32*(3), 365–396.

Nichols, J. (1981). *Predicate nominals: A partial surface syntax of Russian*. Berkeley, CA: University of California Press.

Ovsjaniko-Kulikovskij, D. N. (1902). *Sintaksis russkago jazyka* (1st ed.). Saint Petersburg: Izd. DE Žukovskago.

Pelevin, V., & Bromfield, A. (1997). *The yellow arrow*. New York: New Directions.

Pereltsvaig, A. (2001). *On the nature of intra-clausal relations: A study of copular sentences in Russian and Italian* (Unpublished doctoral dissertation). McGill University, Montreal.

Peškovskij, A. (1914). *Russkij sintaksis v naučnom osveščenii* (1st ed.). Moscow.

Podgaevskaja, A., & Honselaar, W. (2018). *Učebnaja grammatika russkogo jazyka – praktische grammatica van de russische taal* (2nd ed.). Amsterdam: Pegasus.

Potebnja, A. (1888). *Iz zapisok po russkoj grammatike* (1st ed.) (Nos. 1–2). Xar'kov: D.N. Poluextov.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43-55.

Richardson, K. R. (2007). *Case and aspect in slavic*. Oxford: Oxford University Press.

Røed, R. (1966). *Zwei studien über den prädikativen instrumental im russischen.* Oslo: Universitetsforlaget.

Slizhikova, A., Veysov, A., Nurtdinova, D., & Voronin, D. (2021). *Russian open speech to text (stt/asr) dataset.* Zenodo.

Smolka, V. (2017). What comes first, what comes next: Information packaging in written and spoken language. *Acta Universitatis Carolinae. Philologica*(1), 51-61.

Tannen, D. (1982). The oral/literate continuum in discourse. *Spoken and written language: Exploring orality and literacy*, 1–16.

Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge: Cambridge University Press.

Vostokov, A. (1831). *Russkaja grammatika aleksandra vostokova, po načertaniju ego že sokraščennoj grammatiki polnee izložennaja* (1st ed.). Saint Petersburg: V Tip. Imp. Ros. akademii. Retrieved from `https://rusneb.ru/catalog/000202_000006_2574410/`

Wade, T. (2020). *A comprehensive Russian grammar*. Hoboken, NJ: John Wiley & Sons.

Zakharov, V. (2013). Corpora of the Russian language. In I. Habernal & V. Matoušek (Eds.), *Text, speech, and dialogue* (pp. 1–13). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zasorina, L. N. (1977). *Častotnyj slovar' russkogo jazyka: Okolo 40 000 slov*. Moskva: Izd-vo Russkij jazyk.

Zemskaja, E. A. (2011). *Russkaja razgovornaja reč': Lingvističeskij analiz i problemy obučenija* (4th ed.). Moskva: Izdatel'stvo Flinta and Izdatel'stvo Nauka.

Černov, V. I. (1986). Konstrukcii s znamenatel'nymi svjazkami v sovremennom russkom jazyke. *Naučnye doklady vyššej školy. Filologičeskie nauki*(1), 76–80.

Šmelev, D. N. (2002). *Izbrannye trudy po russkomu jazyku*. Moskva: Jazyki slavjanskix kul'tur.

Švedova, N. J. (1980). *Russkaja grammatika, tom 2*. Moskva: Izd-vo Nauka.

Appendices

A Qualitative overview and inspection of dependency parser

A.1 Model architecture specifics

This subsection describes the key decisions made in designing the model architecture, without getting into technical details. The data consists of separate sentences, sampled from different documents (which contain multiple sentences). A single sentence can contain multiple predicate constructions (in the written part of the RNC mini-corpus, only 54 out of $57,608$ sentences contained more than one eligible predicate construction, two of which contained three[31].) A sentence is split into separate tokens. These tokens are either words (again marked by the annotators of the RNC) or punctuation marks (*e.g.* «, **-**, **.**). Certain symbols, that have shown to be not affecting classification results, were stripped from the sentences (*e.g.* **%**, **№**, **"**, **/**), while others were converted (*e.g.* **?!** to **?**, **!!!** to **!** and **...** to **.**). The result of this procedure is a set of sentences consisting of tokens. To derive the dependency relationships between the tokens in a sentence, the resulting set of sentences is inputted into spaCy. As spaCy, for example, regards words containing a hyphen (*e.g. kto-nibud'*) as three separate tokens (*e.g. kto*, - and *nibud'*), these tokens had to be 'retokenized' in order for them to align with the initial scheme of tokenization. Contained in the final dataset were solely sentences for which spaCy's tokenization did exactly match the inputted token sequence (in fact, only one sentence in the written part of the corpus did not[32]). The manual part-of-speech tag annotations by the authors of the RNC were deliberately not passed on to spaCy as the data considered in section 4 does not enjoy the richness of extensive manual tagging. Moreover, as the accuracy of the current architecture was observed to be near-perfect, there was no significant need to do so; especially as it would have added another layer of complexity, while the overarching goal of this thesis to provide a *simplistic* and *comprehensible* approach useful to the entire linguistic research community.

In terms of dependency hierarchies, the closest ancestor to the copula *byt'* is considered to be the predicate. In most cases, such as when regarding the sentence shown in figure 1, the copula has only a single ancestor (which is the predicate). As constructions with prepositions would yield incorrect results using this method (*e.g.* in *on byl s Markom* 'he was with Mark', the 'ancestor' of *byl* would be *Markom*, which would cause the sentence to be interpreted as a past tense noun

[31] An example sentence is: *Moja žizn' byla, čto nazyvaetsja, roman, govorit babuška, i papa tože govorit, čto ego žizn' byla romanom, i mama tože govorit: moja žizn' byla roman.* 'My life was, how to say, a novel, said grandmother; and father also said that is life was a novel; and mother also said: my life was a novel.'

[32] This seems to be the result of a tiny mistake by the RNC annotators: no marked word in the entire database contains a trailing space, while *km* in *Ploščad' – 95,5 tys. kv. km* in one of the documents did; note that other sentences containing *km* also do not contain a trailing space. A simple work-around would be to filter spaces (and possibly other characters) within the boundaries of a marked word (but this would not yield different results).

predicate which is incorrect), constructions containing the prepositions *s*, *meždy*, *pod*, *nad* and *za* directly before the alleged predicative part were discarded. Some sentences contained illegal sequences of characters and were therefore filtered (347 out of 57,955 for the written part of the corpus).[33] Two separate runs were made: one in which no note of author names was taken; and one which allowed only a single document from a single author. From this document, the author's most preferred case was counted once for all the different types of constructions listed in table 1. If a text was authored by multiple persons, the preferences would be attributed to each of the authors, and counted once. If a subsequently processed text was authored by both an 'already encountered' author, and an 'unseen' author, the contents would be attributed for with respect to the first encountered author (working like a 'cascade effect'). The text would thus not be additionally counted. If case usage for a construction occurs equally frequent, no count at all is recorded for the author(s) for that specific type of predicate.

A.2 Examples of model output

In this subsection, several examples are considered that illustrate the workings of the model. Predicates are typeset in bold, while its predicators are underlined. Examples were only taken from the written part of the mini-corpus, as it offered more grammatical and clear sentences. In (3), two constructions are considered that were classified as containing a past tense conjugation of *byt'* in combination with a predicate noun in the nominative case. (3-a) is an adjective–noun construction which functions as a single noun (other examples are (5-a), (7-a) and (7-b); no such pairs have been observed to be not in case agreement). As is visible from (3-b), idioms are included in the output as well. It is worth to note that idioms are more 'syntactically frozen,' depending on their age in the language (Cutler (1982: 319); Fraser (1970: 39, 41–42)). Following this reasoning, idioms thus have a higher tendency to decline in NOM, as it represents the former *status quo*. The only sensible way to filter these structures (which were correctly annotated in every respect), in order to make the resulting statistics more *pure*, is to maintain a *dictionary* of idioms to exclude from the sample. Studies into language change could fair well if they would adopt an approach of this kind.

(3) a. *S raboty uxodili po otdel'nosti, no eto byla* ***bespoleznaja***
from work left by separateness but it was useless.ADJ.NOM.SG
xitrost'.
deception.NOM.SG
'We left work separately, yet the trick did not work.'

b. *Kak vy ponimaete, na to i byl* ***rasčët***.
how you understand on then and was calculation.NOM.SG
'As you understand, it was the plan.'

[33] An example removed sentence is: *A="Šerlok Xolms", B="doktor Vatson", C="sekretar'", D="professor Moriarti", a_1="udostoverenie detektiva", a_2="Z", a_3="pistolet", b_1="ulika", b_2="motiv prestuplenija", c_1="fotorobot", c_2="pokazanija svidetelja", x="priznanie v prestuplenii."*

In (4) below, four examples are shown of past tense adjective predicates. The form of (4-b) (using *bylo by*, or similar) was observed only three times as an instrumental adjective predicate (once using *by byla*). The form appears more often with brief adjectives (which follow NOM). The construction was also observed in combination with noun predicates. In (4-c), only *bogatym* is selected as the nominal part of the predicate (while *studentom* is presumably regarded as an expression of 'manner' or 'circumstance of time' (as in *studentom ja rabotal programmistom* 'as a student, I worked as a programmer')). This results in the sentence being incorrectly regarded as an example of an adjectival phrase rather than a noun phrase. Fortunately, the effects of these false classifications are minor, as in both cases the predicative nominal is correctly interpreted as to be declined in INS. Note that if the em dash would not be present in the sentence, this key piece of the sentence would be ungrammatical, resulting in *povyšennaja stipendija* being regarded as its predicate. This illustrates why 'safe handling' of punctuation marks is required.

(4) a. ***Samym bol'šim*** *v istorii nalogov byl nalog na saxar* (...)
most.ADJ.INS.SG big.ADJ.INS.SG in history tax was tax on sugar
'The largest tax in history was the tax on sugar'

b. *Tak, bylo by* ***estestvennym*** *predpolagat', čto* (...)
so was would natural.ADJ.INS.SG suppose that
'So, it would have been natural to assume that...'

c. *Studentom ja byl* ***bogatym*** *– povyšennaja*
student.INS.SG I was rich.ADJ.INS.SG increased.ADJ.NOM.SG
stipendija (...)
scholarship.NOM.SG
'As as student, I was rich – increased scholarship...'

d. *Pisatelem Robertson byl* ***maloopytnym****, i* (...)
writer.INS.SG Robertson was inexperienced.ADJ.INS.SG and
'As a writer, Robertson was inexperienced and...'

The sentences in (5) are examples of instrumental noun predicates in future tense. (5-a) is an example of an animate predicate noun phrase. In this sentence, the notion of "restricted temporality," mentioned by Krasovitsky et al., is, at the least, questionable. (5-a) is also an example of the subject, copula, and predicate not occurring in a sequence. (5-b), on the contrary, is a prime example of these elements occurring sequentially. In (5-c), *voloskamy* takes on INS, because of the presence of the preposition *meždu*. It is correctly not regarded as a valid predicate.

(5) a. *Pust' knigi etogo* (...) *poeta budut tvoimi* ***samymi***
let-be books this poet will-be your best.ADJ.INS.PL
blizkimi druz'jami!
closest.ADJ.INS.PL friends.INS.PL
'Let the books of this (...) poet be your best and closest friends!'

b. *Ja budu* ***pisatelem!***
I will-be writer.INS.SG
'I will be a writer!'

c. *Togda meždu voloskami budet bol'še vozduxa*, *šubka stanet teplee*.
then between hair.INS.PL will-be more air *šubka* becomes warmer.
'There will be more air between the hairs, so that the *šubka*[34] becomes warmer.'

The two examples below, in (6), were the only animate noun predicates (in future tense) to take on NOM.

(6) a. (...) *ego fil'm – eto budet* ***Čapaev*** *našego vremeni.*
his film it will-be *Čapaev*.NOM.SG our time
'His film – it will be the *Čapaev*[35] of our time!'

b. *Pust' ja budu odin* ***negr*** *na vsju školu!*
let-be I will-be one black-person.NOM.SG to entire school
'Let me be the black person for the entire school!'

Nominative inanimate nouns in future tense were also outnumbered (only 16 occurrences, versus 27 for the instrumental case). Two examples are shown in (7). The first example, (7-a), similar to *e.g.* (6-b) is a 'textbook case' of when there is *no* competition (and only NOM is applicable).

(7) a. *Eto budet* ***pervyj oficial'nyj domašnij***
it will-be first.ADJ.NOM.SG official.ADJ.NOM.SG domestic.ADJ.NOM.SG
matč *rossijskoj sbornoj* (...)
match.NOM.SG Russian national-team
'It will be the first official match of the Russian national team'

b. *Ja ne somnevajus'*, *čto eto budet* ***velikoe nasledie***.
I not wonder what it will-be glorious.ADJ.NOM.SG legacy.NOM.SG
'I do not doubt that it will be a great legacy.'

Slightly more rare than its instrumental counterpart is the sentence in (8-a), which contains a future tense adjective in nominative – of all future tense predicate adjectives (excluding short forms), only 25% were in nominative. The subsequent instrumental example, in (8-b), further illustrates the performance of the model: *takoj* is not misclassified as marked by NOM, but correctly interpreted as to be declined in INS. An interesting case is (8-a), similar to *e.g.* (9-a), where the competition between NOM and INS is highly present.

(8) a. *Togda i čaj budet* ***sladkij***, *i vežlivost' v koridore*.
then and tea will-be sweet and politeness.ADJ.NOM.SG in corridor
"Then we'd have sugar in our tea, and people would behave in the corridor." (Pelevin & Bromfield 1997: 49)

b. (...) *čto sila gravitacionnogo pritjaženija* (...) *budet*
what force.NOM.F.SG gravitational attraction will-be

[34] Russian women's winter jacket

[35] A film about a Russian soldier with the same name.

takoj *že*, (...)
same.PRON.INS.SG again
'that the gravitational force of attraction (...) will be the same...'

(9) shows two examples of past tense adjective predicates put in NOM. Technically, (9-b) involves a pronoun – not an adjective. However, as the RNC marks it as an *adjectival pronoun* (which normally functions like an adjective, *e.g. kotoryj*), I decided to include these parts of speech in the study. The second occurrence of *byt'* in (9-b) (*byla*) does not involve a valid predicative construction, as it is of the form *byt' u kogo-to* 'to have'.

(9) a. *Ty že včera byl* ***p'janyj***, *ne znaeš'.*
that again yesterday was drunk.ADJ.NOM.SG not know.
'You didn't even know that you were drunk yesterday.'

b. *Otvet že glasit, čto eto byl tot*, *u kogo byla samaja*
answer again states what it was that.A-PRON.INS.SG at who was most
bol'šaja golova.
big head.
'The answer again reads that it was the one [person] with the biggest head.'

Finally, (10) shows examples of (in)animate predicative noun constructions declined in INS. (10-a) is a textbook case of an entity (here, the famous mathematician and philosopher *Pythagoras*) that is restricted in terms of temporality. (10-b) likewise explicitly mentions temporal restriction. Instrumental case usage in (10-c) should then imply such restriction *implicitly*, following the cited sources in section 2 (note the difference between (1-b) and this sentence). In similar sense as (10-c), (10-d) restricts the 'unusualness of Inturist as a place [hotel]' implicitly, which makes sense as the tourist agency was privatized following the collapse of the Soviet Union, after being the country's exclusive tour operator since 1929 (Brian Boniface, Cooper, & Cooper 2006: 282).

(10) a. *Govorjat, čto Pifagor byl* ***takim*** ***mužem***, *a posle nego*
say what Pythagoras was such.ADJ.INS.SG man.INS.SG and after him
'They say that Pythagoras was such a man, and, after him...'

b. (...) *on byl* ***direktorom*** *školy, poslednie pjat' let do svoej smerti* (...)
he was director.INS.SG school last five years until his death
'...he was a school director for the last five years of his life...'

c. *Ne mogu skazat', čtoby ja byl ego* ***osobennym*** ***počitatelem*** (...)
not can say whether I was his special.ADJ.INS.SG admirer.INS.SG
'I can't tell whether I was his special admirer...'

d. *Inturist byl ne prosto* ***mestom***, *gde možno ostanovit'sja.*
Inturist was not normal place.INS.SG where possible stay.
'Inturist was not 'just' a place where people could spend the night.'

B Outline and analysis of exemplary transcribed utterances

In this appendix, various listings are presented that illustrate the (in)adequateness of automatic transcripts of spoken language, in combination with automated part-of-speech tagging, for dependency parsing (explained in appendix A). This appendix serves as complementary to listings 1–2, which illustrate different types of errors than contained in this section. All transcriptions are formatted according to SubRip (see footnote 23).

```
422
00:26:58,040 --> 00:27:03,309
[muzyka]

423
00:27:03,520 --> 00:27:10,430
polet sojuza 9 byl novym etapom k

424
00:27:07,310 --> 00:27:14,860
sozdaniju obitaemyx orbital'nyx stancij
```

Listing 4: Fragment from *Leninskim kursom ot s”ezda k s”ezdu. God 1970 (1971)* uploaded by *Sovetskoe televidenie*. An example of a predicate construction correctly classified to have an inanimate past tense noun as its nominal part, declined in the instrumental.

```
178
00:10:54,780 --> 00:11:05,620
esli sud budet demokratičeskim esli
```

Listing 5: Fragment from *V.I.Lenin. Stranicy žizni. VI. I nastupil 1917. Fil'm 2. Vosstanie kak iskusstvo (1990)* uploaded by *Sovetskoe televidenie*. An example of a predicate construction correctly classified to have a future tense adjective as its nominal part, declined in the instrumental.

```
364
00:33:29,289 --> 00:33:35,000
do six por ja ne byl znakom s vami teper'
```

Listing 6: Fragment from *M.Rid. Belaja perčatka. Serija 1 (1968)* uploaded by *Sovetskoe televidenie*. In this example, *znakom* is incorrectly regarded as an instrumental declension of *znak* 'sign.' This is mostly a consequence of word by word annotation, instead of taking context into account, during the annotation process.

```
591
00:49:28,330 --> 00:49:31,850
gospodin major budem kratki kak rimljane
```

Listing 7: Fragment from *JA - 11-17. Serija 1. Telespektakl' po povesti Vasilija Ardamatskogo (1970)* uploaded by *Sovetskoe televidenie*. In this excerpt, the utterances *gospodin major* and *budem kratki kak rimljane* were produced by two different persons. However, in this case, it did not hinder the predicate from being correctly classified as having a future tense nominal adjective following the nominative. Productions of different persons getting mixed up in a single utterance, however, is very error-prone for obvious reasons.

```
25
00:02:37,740 --> 00:02:45,180
no budet budet gavril ja polučil ot nego
```

Listing 8: Fragment from *F.Dostoevskij. Selo Stepančikovo i ego obitateli. Serija 1. MXAT (1973)*. This example illustrates typical spoken language. At the moment the speaker utters *gavril*, the speaker starts a new sentence. The speaker starts with *'budet budet'* to simply interrupt Gabriel and take the floor. However, in this example, *gavril* was incorrectly classified as the nominal part of a predicate.

```
459
00:52:28,000 --> 00:52:35,420
mne ne nužna tvoja pomošč' ja budu podyxat'

460
00:52:32,780 --> 00:52:37,750
s golodu no iz tvoix ruk kuska lepeški

461
00:52:35,420 --> 00:52:37,750
ne voz'mu
```

Listing 9: Fragment from *Četvero iz Čorsanga. Serija 1 (1972)* uploaded by *Sovetskoe televidenie*. Sentences 459–461 of this fragment actually consist of two separate utterances produced by the same person: *mne ne nužna tvoja pomošč'* and *ja budu podyxat' s golodu no iz tvoix ruk kuska lepeški ne voz'mu*. Because every sentence is regarded on a 'stand-alone basis,' *nužna* is incorrectly marked as the nominal part of a predicate connected by *budu*.

```
546
00:30:48,220 --> 00:30:52,860
a samostrely byli vo vre vojny byli byli

547
00:30:51,640 --> 00:30:56,500
samostrely

548
00:30:52,860 --> 00:30:59,620
idioty byli predateli vyrez vse bylo ja
```

Listing 10: Fragment from *Kinopanorama. Interv'ju so Stanislavom Govoruxinym (1990)* uploaded by *Sovetskoe televidenie*. In this excerpt, a man sums up different types of people who were present. In sentence 548, *predateli byli* was wrongly transcribed as *predateli vyrez*. Now, the problem of example (2) arises: were traitors idiots; or were idiots traitors? In any case, the sentence is incorrectly classified to contain a valid predicative construction.

```
779
00:50:52,010 --> 00:50:56,530
bylo važno očen’ byli byli važny točki i
```

Listing 11: Fragment from *Kinopanorama. Peredača posvjaščena tvorčestvu pol'skoj aktrisy Beaty Tyškevič (1991)*. Some sentences contain more than one copula (often sequentally), causing the system to classify the same sentence multiple times. Depending on the context, this can be undesired.