Estimating Grammeme Redundancy by Measuring Their Importance for Syntactic Parser Performance

Aleksandrs Berdicevskis
UiT The Arctic University of Norway
Department of Language and Linguistics
aleksandrs.berdicevskis@uit.no

Abstract

Redundancy is an important psycholinguistic concept which is often used for explanations of language change, but is notoriously difficult to operationalize and measure. Assuming that the reconstruction of a syntactic structure by a parser can be used as a rough model of the understanding of a sentence by a human hearer, I propose a method for estimating redundancy. The key idea is to compare performances of a parser on a given treebank before and after artificially removing all information about a certain grammeme from the morphological annotation. The change in performance can be used as an estimate for the redundancy of the grammeme. I perform an experiment, applying MaltParser to an Old Church Slavonic treebank to estimate grammeme redundancy in Proto-Slavic. The results show that those Old Church Slavonic grammemes within the case, number and tense categories that were estimated as most redundant are those that disappeared in modern Russian. Moreover, redundancy estimates serve as a good predictor of case grammeme frequencies in modern Russian. The small sizes of the samples do not allow to make definitive conclusions for number and tense.

1 Introduction

Explanations of historical language change often involve the concept of redundancy, especially grammatical (morphological) redundancy.

One important example is a family of recent theories about linguistic complexity (Sampson et al., 2009), including those known under the labels "sociolinguistic typology" (Trudgill, 2011) and "Linguistic Niche Hypothesis" (Lupyan and Dale, 2010). The key idea behind these theories is that certain sociocultural factors, such as large population size or a large share of adult learners in the population can facilitate morphological simplification, i.e. increase the likelihood that the language will lose some morphological features, which are often described as "complex" and "redundant".

It is, however, often difficult to determine (and provide empirical evidence in favour of the chosen decision) whether a certain feature is indeed redundant, or to what extent it is redundant and to what extent it is functional. Some conclusions can be drawn from indirect evidence, e.g. typological (cf. Dahl's (2004) notion of *cross-linguistically dispensable* phenomena). For modern languages, redundancy can be studied and measured by means of psycholinguistic experiments (e.g. Caballero and Kapatsinski, 2014), but this approach is not applicable to older language stages and extinct languages.

I propose a computational method to estimate the functionality (and, conversely, redundancy) of a grammeme (that is, a value of a grammatical/morphological category) that can potentially work for any language for which written sources are available or can be collected.

I describe the philosophy behind the proposed method and its relevance to cognitive aspects of language evolution in section 2. Section 3 provides the necessary background for a particular instance of language change that will be used as a case study. Section 4 describes how the experiment was performed, section 5 provides the results. Section 6 discusses possible interpretations of the results, and section 7 concludes.

2 Using parsers to measure morphological redundancy

In the most general terms, morphological redundancy can be described as follows: if a message contains certain morphological markers that are not necessary to understand the message fully and correctly, then these markers can be considered (at least to some extent) redundant.

The problem with operationalizing this intuition is that it is unclear how to model *understanding* (that is, the reconstruction of the semantic structure) of a message by human beings.

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, pages 65–73,
Lisbon, Portugal, 18 September 2015.

In the method I propose, syntactic structure is taken as a proxy for semantic structure, and a reconstruction of syntactic structure by an automatic parser is taken as a model of how a human hearer understands the meaning.

The assumption that these processes have enough in common to make the model adequate is bold, but not unwarranted. It is generally agreed that a correct interpretation of syntactic structure is necessary to understand the meaning of a message, and that humans use morphological cues to reconstruct syntactic structure. Parsers, obviously, do the latter, too. Crucially, the model does not require the assumption that parsers necessarily process the information in exactly the same way as humans. It is enough that they, using the same input, can approximate the output (i.e. syntactic structures) well enough, and modern parsers usually can. Furthermore, parsers also rely heavily on the morphological information, not unlike humans.

The key idea is then to take a morphologically tagged treebank of a language in question and parse it with an efficient parser, *artificially removing* morphological features (either grammemes or categories) one by one. Changes in the parser's performance caused by the removal of a feature can serve as a measure of its redundancy. In other words, if the removal of a feature causes a significant decrease in parsing accuracy, the feature can be considered important for extracting syntactic information and thus functional. If, however, the decrease is small (or absent), the feature can be considered redundant.

Obviously, it is not necessary that this approach will provide an exact and comprehensive measure of morphological redundancy; there are numerous potential sources of noise and errors. We can, however, expect that at least some real redundancy will be captured. The method can then be applied to make rough estimates and thus be useful, for instance, in large-scale typological studies, or in language change studies, or any studies aiming at understanding why languages need (or do not need) redundancy. Understanding that, in turn, will help to reveal the cognitive biases that influence language learning.

It has been shown by means of computational modelling and laboratory experiments that strong biases which affect the course of language change can stem from weak individual cognitive biases, amplified by iterated learning over generations (Kirby et al., 2007; Reali and Griffiths, 2009; Smith and Wonnacott, 2010) and communication within populations (Fay and Ellison, 2013). Thus, if it is shown that there is a diachronic bias towards eliminating redundant grammemes, it will be possible to hypothesize that this bias stems from individual speakers' preference to avoid overloading their speech with excessive complexity.

Importantly for diachronic studies, the method can be applied to extinct languages, provided that large enough treebanks exist.

In the following sections, I will exemplify the method by applying it to a particular case of language change (Proto-Slavic —> Contemporary Standard Russian). I also use the case study to test whether the resulting redundancy estimates are plausible. Following a common assumption that more redundant grammemes are in general more likely to be lost (Kiparsky 1982: 88–99, see also references above), and that Russian has been under considerable pressure to shed excessive complexity (see section 3), I make the prediction that the grammemes that did disappear were on average more redundant than those that were kept, and that the "remove-and-reparse" method should be able to capture the difference.

In order to be explicit about the assumptions behind the current study and its limitations I want to highlight that the study attempts to test two independent hypotheses at once: first, that redundant grammemes are more likely to disappear or become less frequent; second, that parsing is an adequate model of human language perception, since what is redundant for a parser is redundant for a human as well. This can be problematic, since we do not really know whether either of these hypotheses is true.

Let us look at the experiment from the following perspective: if it turns out that there is a strong correlation between importance of the grammeme for parser performance and grammeme survivability, then this fact has to be explained. A plausible explanation which fits well with the existing linguistic theories would be the one outlined above in the form of the two hypotheses: under certain sociocultural conditions speakers tend to abandon redundant grammemes; grammemes that are not important for the parser are redundant. If there is no correlation, however, this absence would not tell us whether both hypotheses are false or only one of them (and which one) is.

In addition to the main prediction, I make a secondary one: assuming that more redundant grammemes will tend to become less frequent, and more functional grammemes will tend to become more frequent, we can expect that the

functionality of grammemes in Proto-Slavic should serve as a good predictor of their frequency in modern Russian. I will test this prediction as well, though the possibilities for this test offered by the current study are limited. In addition, the prediction itself relies on stronger assumptions (redundancy is not necessarily the only, nor even the most important predictor of frequency).

3 From Proto-Slavic to Russian

In this section, I briefly describe the relevant morphological changes that occurred in the period from Proto-Slavic (alias Common Slavic, a reconstructed protolanguage that existed approx. from the 5th to 9th centuries AD) to Contemporary Standard Russian (CSR). Old Church Slavonic is used as a proxy for Proto-Slavic (see section 4.1).

CSR has been chosen for the pilot study for the following reasons. First, Russian is the largest Slavic language with a total of about 166 million speakers (Lewis et al., 2015). Second, its contact with other languages has been quite intense. Bentz and Winter (2013) use 42% as an estimate for the ratio of L2 speakers to the number of all speakers of CSR (their absolute estimate is 110 million). According to the linguistic complexity theories cited in section 1, these factors make pressure towards simplification stronger, i.e. redundant morphological features are more likely to be lost.

Russian has not lost any Proto-Slavic morphological category completely, though many have been very significantly restructured. Some grammemes, however, did disappear.

Proto-Slavic had seven nominal cases: **n**ominative, **a**ccusative, **g**enitive, **d**ative, **i**nstrumental, **l**ocative and **v**ocative. Russian has preserved the former six, but lost the vocative and is now using the nominative in its place. It should be noted that some scholars do not consider the vocative a real *case* (Andersen, 2012: 139–143). In addition, the vocative was relatively infrequent, and often coincided with the nominative already in Proto-Slavic. Still, there is a clear distinction between Proto-Slavic (where a separate obligatory vocative form existed) and CSR (where there is no such form). The fact that CSR developed several novel marginal cases, including the so-called "new vocative", does not affect the general picture in any relevant way.

Proto-Slavic had three numbers: **s**ingular, **d**ual and **p**lural, of which the dual is not present in CSR: the plural is used instead (the dual, however, left visible traces in the morphosyntax of the numerals and the formation of plural forms).

Proto-Slavic had five basic verbal tenses: present (also called *non-past*), aorist, imperfect, perfect and pluperfect.[1] The perfect and pluperfect were analytical forms, consisting of resp. present and imperfect[2] forms of an auxiliary ('be') and a so-called resultative participle. Later, the aorist, imperfect and pluperfect went out of use, while the former perfect gradually lost the auxiliary verb. As a result, in CSR the only means to express indicative past tense is the former resultative, which has lost most of its participial features and is treated on a par with other finite forms. In the current study, I will consider four morphologically distinct tenses: **p**resent, **a**orist, **i**mperfect and re**s**ultative. The label "resultative" will cover all uses of the resultative participle, both in the perfect and pluperfect, both with and without an auxiliary. Non-indicative verbal forms (except for the resultative) will be ignored (i.e. the present and past tense of participles, imperatives, infinitives and subjunctive). To sum up: we will focus on the four tenses listed above, of which two (aorist and imperfect) disappeared, replaced by the resultative.

Finally, a Proto-Slavic verbal grammeme called supine also disappeared, but it will ignored in the current study, partly since its frequency in Old Church Slavonic is very low, partly since it is not entirely clear what grammatical category it belongs to.

4 Materials and methods

4.1 Language data

The oldest Slavic manuscripts were written in Old Church Slavonic (OCS), a literary language based on a South Slavic dialect of late Proto-Slavic. OCS is not a direct precursor of CSR (nor of any other modern Slavic language), but it is the best available proxy for Proto-Slavic, and is commonly used in this role.

4.2 Treebank and parser

I extracted OCS data from the Tromsø Old Russian and OCS Treebank,[3] limiting myself to one document, the Codex Marianus, which has been thoroughly proofread and submitted to compre-

[1] The verb 'be' also has a separate synthetic future tense, which is ignored here.
[2] Sometimes also aorist or perfect.
[3] https://nestor.uit.no/

hensive consistency checks (Berdicevskis and Eckhoff, 2015). The Codex Marianus is dated to the beginning of the 11th century. The TOROT file contains 6350 annotated sentences.

The TOROT is a dependency treebank with morphological and syntactic annotation in the PROIEL scheme (Haug, 2010, Haug et al., 2009). For the purposes of the experiment, I converted the native PROIEL format to the CONLL format (see Table 1).

For the parsing experimens I used MaltParser (Nivre et al., 2007), version 1.8.1.[4] The Codex Marianus was split into a training set (first 80% of sentences) and a test set (last 20% of sentences). The parser was optimized on the training set using MaltOptimizer (Ballesteros and Nivre, 2012), version 1.0.3.[5] Optimization had been performed before any grammemes were merged or any morphological information was deleted (see section 4.3).

Parsing the TOROT with MaltParser faces several difficulties. First, the PROIEL scheme uses secondary dependencies – for external subjects in control and raising structures, and also to indicate shared arguments and predicate identity. Since MaltParser cannot handle secondary dependencies, all this information was omitted. Second, the PROIEL scheme also systematically uses empty verb and conjunction nodes to account for ellipsis, gapping and asyndetic coordination. Since MaltParser cannot insert empty nodes, they were explicitly marked in both the training and test sets (with form and lemma having the value *empty*; part-of-speech marked as resp. verb or conjunction, and morphological features having the value *INFLn* 'non-inflecting', see Table 1, token 14).

The LAS (labelled attachment score) for parsing the test set was 0.783. Parsing took place before merging grammemes, but after removing person and gender information from verbs (see section 4.3).

4.3 Merging grammemes

When linguists say that a grammeme *disappeared*, they usually mean that the grammeme merged with another one, or that another grammeme expanded its functions, replacing the one that *disappeared*. As described in section 3, disappearances that occurred in the (pre)history of Russian were actually mergers: vocative > nominative; dual > plural; aorist and imperfect > resultative.

I will illustrate how I model grammeme mergers using the example of the number category. The category has three values: **s**ingular, **p**lural, and **d**ual, their absolute frequencies in the Codex Marianus are resp. 28004, 10321 and 942. Every grammeme is consecutively merged with the other grammemes in the same category. When, for instance, the s>p merger takes place, the string *NUMBs* in the FEATURE column (see Table 1) is replaced with *NUMBp* (see below about the number of occurrences that are replaced). After that, the original values are restored, and s>d merger follows: *NUMBs* is being replaced with *NUMBd*. Later, p>s, p>d, d>s and d>p mergers take place in the same way.

After every merger, the Codex Marianus is split into the same training and test sets, and parsed anew, using the same optimization settings. The difference between the original LAS and the resulting LAS (delta) shows how strongly the merger affected parser performance. For every grammeme, the sum of deltas for all its mergers (for s, that would be the sum of deltas for the mergers s>p, s>d) is taken as a measure of its functionality, or non-redundancy. The higher this number is, the more important the grammeme is for parser, and the less redundant it is.

The frequency of grammemes can vary greatly, as the number category illustrates. It can be expected that if we always merge all the occurrences of every grammeme, then the deltas will tend to be higher for more frequent grammemes, because the larger number of occurrences is affected. On the one hand, frequency is an important objective property of any linguistic item, and it is legitimate to take it into account when estimating redundancy and functionality. On the other hand, very high frequencies can skew the results, making the functionality estimate a mere correlate of frequency, which is undesirable. In order to test whether redundancy/functionality is a useful measure, we need to disentangle it from potential confounding factors. To address this issue, the experiment was run in two conditions.

In condition 1, all occurrences of every grammeme are merged (that is, the s>d merger results in 28946 *NUMBd* strings and 0 *NUMBs* strings, while the d>s merger results in 28946 *NUMBs* strings and 0 *NUMBd* strings). It is reasonable to expect that this condition will have a bias for more frequent grammemes: they will get higher functionality scores.

[4] http://www.maltparser.org/

[5] http://nil.fdi.ucm.es/maltoptimizer/index.html

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	i *and*	i	C	C-	INFLn	10	aux
2	aŝe *if*	aŝe	G	G-	INFLn	10	adv
3	k″to *anyone*	k″to	P	Px	NUMBs|GENDq|CASEn	4	sub
4	poimet″ *forces*	pojati	V	V-	NUMBs|TENSp|MOODi|VOICa	2	pred
5	tja *you*	tja	P	Pp	PERS2|NUMBs|GENDq|CASEa	4	obj
6	po *by*	po	R	R-	INFLn	4	adv
7	silě *force*	sila	N	Nb	NUMBs|GENDf|CASEd	6	obl
8	pop′riŝe *mile*	pop′riŝe	N	Nb	NUMBs|GENDn|CASEa	14	adv
9	edino *one*	edino	M	Ma	NUMBs|GENDn|CASEa	8	atr
10	idi *go*	iti	V	V-	PERS2|NUMBs|TENSp|MOODm|VOICa	0	pred
11	s″ *with*	s″	R	R-	INFLn	10	obl
12	nim′ *him*	i	P	Pp	PERS3|NUMBs|GENDm|CASEi	11	obl
13	d′vě *two*	d″va	M	Ma	NUMBd|GENDn|CASEa	10	adv
14	empty (*go*)	empty	V	V-	INFLn	4	xobj

Table 1. Example sentence (Matthew 5:41, 'If anyone forces you to go one mile, go with them two miles') from the Codex Marianus in the PROIEL scheme and CONLL format. OCS words are transliterated using the ISO 9 system (with some simplifications). Columns: 1 = token ID; 2 = form; 3 = lemma; 4 = coarse-grained POS tag; 5 = fine-grained POS tag; 6 = features; 7 = head; 8 = dependency relation. For the reader's convenience, an English gloss is added under every form (in italics). Note the absence of the *PERS3* feature for token 4. While it had originally been there, it was removed in order to facilitate the mergers of indicative and participial forms (see main text). It is, however, kept for those verb forms which will not be affected by any mergers (e.g. token 10, which is in the imperative).

In condition 2, the number of merged occurrences is *constant for all grammemes* in the category, and equal to *the frequency of the least frequent grammeme*. For number, that would be dual with its frequency of 942. Here, the s>d merger results in 1884 *NUMBd* strings (942 original + 942 merged) and 27062 *NUMBs* strings (28004 original - 942 merged), while the d>s merger results in 28946 *NUMBs* strings (28004 original + 942 merged) and 0 *NUMBd* strings (942 original - 942 merged). This condition can potentially create a bias for less frequent grammemes: while the absolute number of the affected occurrences is always the same, their share in the total occurrences of the grammeme that is being merged can be very different. The d>s merger, for instance, empties the dual grammeme fully, while the s>d merger removes only a small share of the singular occurrences. This potential bias can, however, be expected to be weaker than the reverse bias in condition 1, and the results can then be expected to be more reliable.

The occurrences to be merged are selected randomly. Since the resulting change in parser performance may depend on the sample of selected occurrences, the process is repeated 10 times on 10 random samples, and the average of 10 functionalities is taken as the final measure.

Note that in both conditions, mergers always affect two grammemes: the source (i.e. the one that is being merged) and the target one. However, I consider only the former effect and ignore the latter: for instance, the change of LAS after s>d merger is added to the functionality of s, but

not of d. Technically, it is possible to take into account the respective delta when calculating the functionality of d, too, but it is not quite clear whether this is theoretically justified. The rationale behind adding the delta to the functionality of s is that s has been (partially) removed, and we are investigating how this removal affected the possibility to restore syntactic information. No instances of the target value, however, have been removed, and while the grammeme has been somewhat changed by its expansion, it is not clear how to interpret this change. Besides, I assume that the influence of the expansion of the target grammeme is small (compared to that of the removal of the source one) and ignore it in the current study.

Case is processed in exactly the same way as number (each case is consecutively merged with the six others), but tense represents an additional substantial problem. Remember that the present, imperfect and aorist are typical finite forms, which means that they have the features person, number, tense, mood (the value is always **i**ndicative) and voice, while the resultative is a participle (the mood[6] value is always **p**articiple), and does not have the feature person, but does have the features gender, case and strength.[7] By the OCS period, however, the resultative has already lost most of its original participial properties, and case is always nominative, while strength is always **s**trong. The problem is that when we merge, for instance, the present with the resultative, we have a feature mismatch: the present has one extra feature (person) that the resultative never has, but lacks the three other features (gender, case, strength); in addition, the mood feature is different. Obviously, the merger in the other direction faces the inverse obstacle.

I solve this problem in the following way. Since there is no means to reconstruct information about person when merging resultative to the three indicative tenses and no means to reconstruct information about gender when merging in the other direction, I remove person and gender features from all relevant verbal forms. This is done prior to any other operations. The initial LAS (0.783) is calculated after this removal. Without it, LAS would have been 0.785. When a resultative > {present | aorist | imperfect} merger occurs, information about case and strength is removed, and mood is changed from **p** to **i**. When a merger in the other direction occurs, information about case and strength is added (resp. **n** and **s**), and mood is changed from **i** to **p**. While these changes are pretty artificial, they do ensure that we perform a full merger that affects all relevant properties of a grammeme, and not only changes its label.

5 Results

Results of the experiment for both conditions are presented in Table 2. Grammemes within each category are first sorted in descending order by their functionality in the condition 2 (which is supposed to be a more reliable measure), then by their functionality in condition 1.

Zero values for vocative in columns 3 and 4 do not mean that merging vocative with other cases never affects the parser performance at all, but that the changes are negligibly small, represented as 0 after rounding to three decimal places. Negative functionality values (for number grammemes) mean that merging this grammeme with others on average leads to *increase* of the LAS, not decrease. These results can be interpreted in the same way as positive and zero values: lower functionality (which in this case means larger increase in parsing accuracy) implies higher redundancy (so high that its removal facilitates the restoration of the syntactic structure instead of inhibiting it).

Absolute frequencies of every grammeme are provided for OCS (the Codex Marianus) and CSR. The CSR frequencies were calculated using the manually disambiguated part (≈6 million words) of the Russian National Corpus[8] (RNC). While it is known that ranking the CSR grammemes by frequency may sometimes provide different results depending on the chosen corpus (Kopotev 2008), the general picture can be assumed to be adequate and stable, since the RNC is a relatively large and well-balanced corpus.

6 Discussion

As can be seen, in both conditions the vocative gets identified as the most redundant case. This fits nicely with the fact that CSR lost it, while preserving the other six cases.

[6] The mood category in the PROIEL scheme for OCS has broader coverage than the traditional mood category. It has the grammemes indicative, imperative, subjunctive, infinitive, participle, gerund and supine (i.e. covers both mood and finiteness).

[7] Strength here refers to the distinction between long and short forms of Slavic adjectives and participles, remotely similar to the Germanic distinction between weak and strong adjectives.

[8] http://ruscorpora.ru/

Category	Grammeme	Functionality (condition 1)	Functionality (condition 2)	Frequency (OCS)	Frequency (CSR)
CASE	n	0.039	0.009	9812	1026131
	g	0.017	0.008	4470	731435
	a	0.017	0.006	7657	539768
	d	0.006	0.004	3694	180131
	l	0.008	0.001	1671	265701
	i	0.005	0.001	1050	271531
	v	0	0	400	0
NUMBER	s	-0.004	0	28004	2861455
	p	-0.004	-0.001	10321	886420
	d	-0.002	-0.002	942	0
TENSE	s	0.009	0.009	199	458820
	p	0.009	0.001	4452	231946
	a	0.007	0.001	3772	0
	i	0.003	0.001	1121	0

Table 2. Results of the merging experiment for the two conditions.

Moreover, most modern Indo-European languages have lost the original Proto-Indo-European vocative. Most Slavic languages, however, have retained it. Outliers here are Bulgarian and Macedonian, which have lost all the cases but the vocative. These two Slavic languages, however, are exceptional in many respects (possibly due to the influence of the Balkan Sprachbund).

Importantly, the functionality ranking of cases does not seem to be a mere reflection of their frequency ranking in OCS. In condition 1, the genitive and the accusative[9] have the same functionality (while the accusative is noticeably more frequent), and the dative is less functional than the locative, while being more frequent). In condition 2, the genitive is more functional than the accusative, despite lower frequency.

As regards the second prediction, functionality scores do turn out to be a good predictor for CSR frequency. Pearson correlation coefficients[10] are 0.96 ($p < 0.001$) in condition 1, and 0.92 ($p = 0.004$) in condition 2. Importantly, in both conditions functionality is a better predictor than plain OCS frequency. The Pearson coefficient for the OCS and CSR frequencies is 0.86 ($p = 0.012$).

Absolute differences between the functionality of cases are larger in condition 1, which can probably be explained by a frequency effect.

For number, the situation is different. In condition 2, the singular gets the highest functionality score and the dual the lowest, which again fits with the historical development of the Slavic languages: all except Slovene and Sorbian have lost the dual form (the same holds for most other Indo-European languages). In condition 1, however, the results are opposite: the dual is the most functional grammeme, while the singular and the plural are the most redundant ones.

Functionality is a poor predictor for CSR frequency in condition 1 ($r = -0.73$, $p = 0.471$). It is better correlated (though still insignificant) in condition 2 ($r = 0.98$, $p = 0.14$), but loses out to OCS frequency ($r = 1$, $p = 0.026$). The extremely small sample size, however, makes the Pearson test unreliable.

Within the tense category, the resultative is at the most functional end of the scale, while the aorist and the imperfective are at the least functional end in both conditions. The absolute values, however, differ, as does the position of the present: in condition 1, it has the same value as the resultative (slightly higher than the aorist), whereas in condition 2, its functionality is equal to that of the aorist and the imperfect. Importantly, the least frequent tense (the resultative) gets the highest functionality score in both conditions.

For tense, OCS frequency is the worst predictor of CSR frequency ($r = -0.39$, $p = 0.611$). Functionality has larger coefficients and smaller p-values, though they do not reach significance (in condition 1 $r = -0.74$, $p = 0.259$; in condi-

[9] Both in OCS and CSR the accusative case of some animate nouns is identical to the genitive. In the TOROT, these genitive-accusatives are annotated as genitives. For consistency's sake, I coded them as genitives when calculating the CSR frequencies as well.

[10] It can be questioned whether it is legitimate to use Pearson product-moment correlation, or a non-parametric method like Spearman rank correlation should be preferred. Given that the data are on the interval scale and that they answer the Shapiro-Wilk normality criterion, I opt for Pearson.

tion 2 r = -0.87, p = 0.132). Again, small sample size prevents any definitive conclusions.

It is not quite clear why the present scores so low in the condition 2: it is frequent enough, it has survived in all Slavic languages, and can be expected to be quite functional. It can be a consequence of the complicated corrections that were performed to compensate for the morphological mismatch between participle and indicative (see section 4.3).

It is remarkable that the two tenses that get the lowest scores in both conditions are those that have disappeared in CSR: the aorist and the imperfect. They have not survived in other Slavic languages either, with the exception of Bulgarian, Macedonian and partly Bosnian-Serbo-Croatian, where its use is restricted to certain genres and dialects (Dahl 2000: 101). The decline of the imperfect usually happens before the decline of the aorist in Slavic languages (including the East Slavic group, to which the CSR belongs), and, remarkably, the imperfect gets lower functionality score in condition 1.

The difference between the scores of the most and the least functional grammemes is largest for case and lowest for number in both conditions. This fits with the functionality values of the categories themselves measured in a separate experiment, where the changes of LAS were measured after deleting all information about a particular category (for instance, removing all strings *NUMBs*, *NUMBd* and *NUMBp* from the FEATURE column). Case turned out to be the most functional category (0.030), which is unsurprising, given that cases are typically assumed to mark the syntactic role of an argument in a sentence, and hence can be expected to be crucial for the reconstruction of the syntactic structure. Tense got second place (0.014) and all other categories scored noticeably lower, from 0.004 to 0 (for number the value is 0.003). This difference can account for the contradictory results that the two conditions return for number: given that the total functionality of the category (from parser's perspective) is relatively small, the proposed method can be less sensitive to real performance changes caused by mergers and more vulnerable to random fluctuations.

7 Conclusion

While the results vary across categories and conditions, the general trend is quite clear: grammemes that did disappear in the course of language history tend to get lowest functionality scores in the present case study, in other words, the main prediction holds. If we follow the assumption that the most redundant morphological features tend to disappear first, especially under conditions that facilitate morphological simplification (see section 1), then the results confirm the validity of the proposed method.

The secondary prediction holds for case grammemes, where functionality allows to make better predictions about the frequencies that the grammemes will have after almost a thousand years than plain frequency. It does not hold for number and tense, but small sample sizes (i.e. the number of grammemes within a given category) can be the reason.

The fact that the functionality scores for case correlated with the CSR frequencies suggests that the method can predict grammeme development, at least in some cases. It seems to be able to capture the "functional potential" of a grammeme, which can influence its frequency in the future: the lower it is, the more likely the frequency is to decrease. However, given the small differences in correlation coefficients, the small number of datapoints and the problematic situation with number and tense, the support for this hypothesis at the moment is rather weak.

It is not quite clear which of the two conditions gives better predictions. It is possible that the best way to calculate functionality is to combine the results of both conditions in some way. The method should be tested on larger language samples in order to solve this and other potential issues and find its strengths and limitations. One immediate development of this study would be to take into account *all* modern Slavic languages to find out how likely a given Proto-Slavic grammeme (or category) was to disappear or to stay. Intermediate language stages (Old Russian, Old Bulgarian etc.) can, of course, also be considered. Given that some amount of noise (for instance, peculiarities of a specific treebank, specific document or a chosen parser) will always affect the performance of the method, larger language samples can also lead to more stable and more interpretable results.

Looking from another perspective, this study is an attempt to model how human speakers process linguistic information and which features are least informative for them. While the processing itself is not expected to be entirely isomorphic to what happens in a human mind (and the model in general is somewhat of a black box, unless we use a fully deterministic parser), the

output gives us some information about human cognition and existing learning and usage biases.

The method can be applied not only to language change or older stages of language, but also to modern languages, and the results can be tested against existing psycholinguistic or typological evidence about redundancy.

Obviously, it is necessary to test how robust the results are with respect to the choice of the parser, annotation scheme, merging procedures and languages.

The results can have some practical value, too, as they provide information about which features are most and least useful for parsers.

Acknowledgments

I am grateful to Hanne Eckhoff, Laura Janda and three anonymous reviewers for their valuable comments, and to Ilya German for technical assistance. This work has been supported by the Norwegian Research Council grant 222506.

References

Henning Andersen. 2012. The New Russian Vocative: Synchrony, Diachrony, Typology. *Scando-Slavica*, 58(1):122–167.

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: A System for MaltParser Optimization. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 23–27 May 2012*. European Language Resources Association.

Christian Bentz and Bodo Winter. 2013. Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, 3:1–27.

Aleksandrs Berdicevskis and Hanne Eckhoff. 2015. Automatic identification of shared arguments in verbal coordinations. *Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue"*, 14:33–43.

Gabriela Caballero and Vsevolod Kapatsinski. 2014. Perceptual functionality of morphological redundancy in Choguita Rarámuri (Tarahumara). *Language, Cognition and Neuroscience*, DOI: 10.1080/23273798.2014.940983

Östen Dahl (ed.) 2000. *Tense and Aspect in the Languages of Europe*. Mouton de Gruyter, Berlin, Germany.

Östen Dahl. 2004. *The growth and maintenance of linguistic complexity*. John Benjamins, Amsterdam, The Netherlands.

Nicolas Fay and T. Mark Ellison. 2013. The cultural evolution of human communication systems in different sized populations: usability trumps learnability. *PLoS ONE* 8(8):e71781.

Dag Haug. 2010. PROIEL guidelines for annotation. http://folk.uio.no/daghaug/syntactic_guidelines.pdf

Dag Haug, Marius Jøhndal, Hanne Eckhoff, Eirik Welo, Mari Hertzenberg and Angelika Müth. 2009. Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues* 50(2):17–45.

Simon Kirby, Mike Dowman and Thomas L. Griffiths. 2007. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences* 104(12):5241–5245.

Mikhail Kopotev. 2008. K postroeniju chastotnoj grammatiki russkogo jazyka: padezhnaja sistema po korpusnym dannym. *Slavica Helsingiensia* 34:136–151.

M. Paul Lewis, Gary F. Simons and Charles D. Fennig (eds.). 2015. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas.
Online version: http://www.ethnologue.com.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5(1):e8559.

Daniel Nettle. 2012. Social scale and structural complexity in human languages. *Phil. Trans. R. Soc. B* 367:1829–1836.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(2):95–135.

Florencia Reali and Thomas L. Griffiths. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition* 111:317–328.

Geoffrey Sampson, David Gil and Peter Trudgill (eds.) 2009. *Language complexity as an evolving variable*. Oxford University Press, Oxford, UK.

Kenny Smith and Elizabeth Wonnacott. 2010. Eliminating unpredictable variation through iterated learning. *Cognition* 116:444–449.

Peter Trudgill. 2011. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford, UK.