REDUNDANT FEATURES ARE LESS LIKELY TO SURVIVE: EMPIRICAL EVIDENCE FROM THE SLAVIC LANGUAGES

ALEKSANDRS BERDICEVSKIS, HANNE ECKHOFF

Department of Language and Linguistics, UiT The Arctic University of Norway
Tromsø, Norway
aleksandrs.berdicevskis@uit.no, hanne.eckhoff@uit.no

We test whether the functionality (non-redundancy) of morphological features can serve as a predictor of the survivability of those features in the course of language change. We apply a recently proposed method of measuring functionality of a feature by estimating its importance for the performance of an automatic parser to the Slavic language group. We find that the functionality of a Common Slavic grammeme, together with the functionality of its category, is a significant predictor of its survivability in modern Slavic languages. The least functional grammemes within the most functional categories are most likely to die out.

1. Introduction

Many explanations of language evolution and change involve (either explicitly or implicitly) the concept of *redundancy*, especially morphological redundancy. The assumption that redundant features are more likely to disappear has played an important role in historical linguistics for decades (see Kiparsky 1982: 88–99 for an example; Lloyd, 1987: 33–35 for a brief overview). More recently, several influential theories have emerged (Sampson et al., 2009; Lupyan and Dale, 2010; Trudgill, 2011) that refine this assumption, claiming that it does not apply in equal measure to all languages. It is hypothesized that languages under certain sociocultural conditions (such as large population size or a large share of adult learners) will tend to shed excessive (i.e. redundant) complexity.

A serious problem with the notion of redundancy, however, is that it is difficult to operationalize and measure quantitatively, which means that theories such as those cited above must to some extent rest on assumptions or indirect qualitative estimates. In this paper, we improve on a method of measuring morphological redundancy proposed by Berdicevskis (2015).

The key idea behind the method is that the identification of syntactic structure by an automatic parser can be taken as a model of how human beings understand meaning (i.e. identify semantic structure). While the model is not necessarily ecologically valid (parsers and humans process information differently), it is externally valid: given the same input (text to process) as humans, parsers can approximate the output (correct structure) very well. The main benefit of the model is that it makes it possible to run experiments, manipulating the input. If we, for instance, artifically distort the input, removing the information about a given morphological feature, and then compare the performance of the parser *before* and *after* removal, we can estimate how important the feature is for the identification of the underlying structure, how necessary for the understanding of the meaning and hence, how functional (non-redundant).

Importantly for the study of language change and evolution, this ablation technique can be applied both to extant and extinct languages, as long as there exists a decent treebank. We present a case study where we apply the method to the Slavic language group. We estimate the functionality of morphological categories and grammemes in Common Slavic and test how well this information predicts the survival and death of those features in modern Slavic languages.

2. Materials and methods

2.1. *The Slavic group*

The Slavic language group is divided into three branches: South, West and East. All extant languages have rich inflectional morphology, mostly inherited from Common Slavic. In this section, we describe how Common Slavic grammemes survive across Slavic languages (Table 1). In the following two sections we describe how we measure the functionality of these grammemes.

The earliest Slavic texts were written in Old Church Slavonic (OCS), a literary language based on a South Slavic dialect of Late Common Slavic. We use OCS as a proxy for Common Slavic, as is often done in historical linguistics.

We exclude the following from the analysis: mood, finiteness, voice, degree of comparison, adjective long/short form, synthetic future tense (which exists only for the verb 'be'), non-indicative and non-finite verbal forms.[a] The tense grammeme coded as **res** in Table 1 stands for the Common Slavic perfect, pluperfect and conditional that consisted of an auxiliary verb and a so-called **res**ultative participle. We do not take into account any morphological

[a] The reasons for exclusion range from theoretical (there is no unified view on the structure of some categories, e.g. finiteness) to methodological (our experiments in their current form do not work with binary categories, e.g. adjective form, or categories that are too heterogeneous, e.g. mood).

innovations. Decisions represented in Table 1 largely follow Comrie & Corbett (1993).

Table 1. Common Slavic grammemes across modern Slavic languages

Cate-	Gram-	CF	GF	freq	South branch				West branch						East branch		
gory	meme	$\bullet 10^{-3}$		$\bullet 10^{3}$	bul	mkd	hbs	slv	ces	slk	hsb	dsb	pol	csb	rus	bel	ukr
Case	nom	36	6.3	20.5	1	1	1	1	1	1	1	1	1	1	1	1	1
Case	acc	36	5.2	16.1	0	0	1	1	1	1	1	1	1	1	1	1	1
Case	dat	36	4.8	8.4	0	0	1	1	1	1	1	1	1	1	1	1	1
Case	gen	36	4.2	11.1	0	0	1	1	1	1	1	1	1	1	1	1	1
Case	ins	36	2.8	3.1	0	0	1	1	1	1	1	1	1	1	1	1	1
Gend	m	4	2.5	35.4	1	1	1	1	1	1	1	1	1	1	1	1	1
Pers	1	3	2.0	4.8	1	1	1	1	1	1	1	1	1	1	1	1	1
Pers	2	3	2.0	6.1	1	1	1	1	1	1	1	1	1	1	1	1	1
Pers	3	3	2.0	22.4	1	1	1	1	1	1	1	1	1	1	1	1	1
Gend	n	4	2.0	10.3	1	1	1	1	1	1	1	1	1	1	1	1	1
Numb	pl	4	2.0	20.2	1	1	1	1	1	1	1	1	1	1	1	1	1
Numb	sg	4	2.0	60.7	1	1	1	1	1	1	1	1	1	1	1	1	1
Tens	res	8	2.0	0.4	1	1	1	1	1	1	1	1	1	1	1	1	1
Case	loc	36	1.7	3.3	0	0	0	1	1	1	1	1	1	1	1	1	1
Case	voc	36	1.7	0.9	1	1	1	0	1	0	1	0	1	1	0	0	1
Gend	f	4	1.5	10.6	1	1	1	1	1	1	1	1	1	1	1	1	1
Tens	pres	8	1.3	15.6	1	1	1	1	1	1	1	1	1	1	1	1	1
Numb	du	4	1.0	2.1	0	0	0	1	0	0	1	1	0	0	0	0	0
Tens	aor	8	0.7	7.4	1	1	1	0	0	0	1	0	0	0	0	0	0
Tens	impf	8	0.7	2.1	1	1	0	0	0	0	1	0	0	0	0	0	0

CF = category functionality, GF = grammeme functionality (see section 2.3), freq = absolute frequency (in OCS). The table is sorted first by GF (descending), then by CF (ascending), i.e. in the approximate descending order of survivability (see section 3). Languages are denoted by their ISO 639-3 codes. 0 means that a grammeme is (almost) extinct, 1 means that it is extant.

2.2. *Treebank and parser*

We extracted OCS data from the Tromsø Old Russian and OCS Treebank (TOROT),[b] using the two largest documents, the Codex Marianus and the Codex Suprasliensis, both dated to the beginning of the 11th century. The joint TOROT file contains 13308 manually annotated (and double-checked) sentences.

[b] https://nestor.uit.no/

The TOROT is a dependency treebank with morphological and syntactic annotation according to the PROIEL scheme (Haug et al., 2009). For our experiments, we converted the files to the CONLL format (Table 2).

For the parsing experiments we used MaltParser (Nivre et al., 2007), version 1.8.1.[c] The parser was optimized using MaltOptimizer (Ballesteros and Nivre, 2012), version 1.0.3,[d] optimization was performed on the original text, before any changes (see section 2.3).

Table 2. Example OCS sentence ('He said to them', from Matthew 12:11) in the PROIEL scheme and CONLL format.

ID	Form	Lemma	CPOS	FPOS	Features	Head	DREL
1	on″ *he*	on″	P	Pd	NUMBs|GENDm|CASEn	3	sub
2	že *but*	že	D	Df	INFLn	3	aux
3	reče *say*	reŝi	V	V-	PERS3|NUMBs|TENSa|MOODi|VOICa	0	pred
4	im″ *them*	i	P	Pp	PERS3|NUMBp|GENDm|CASEd	3	obl

CPOS/FPOS = coarse/fine-grained part-of-speech tag; DREL = dependency relation. OCS words are transliterated using the ISO 9 system.

Contrary to standard practice in computer science, we do not create separate training and test sets, and thus perform all operations, including optimization, on the whole dataset. The reason for this solution is that our goal is not to evaluate how accurately a given parser can analyze a given text, but how its performance is affected by certain changes in the annotation of the input data. As regards absolute measures of performance, we want them to be as high as possible, in order to approximate human performance and thus increase the validity of the model. Training and parsing on the same set allows us to reach a LAS (labelled attachment score) of 0.938,[e] while parsing of unfamiliar test sets usually results in a LAS in the high seventies at best.

2.3. *The ablation experiments*

We perform two experiments. In the first one, we estimate the functionality of the morphological categories listed in Table 1 (column 1), in the second one, the

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

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

[e] This score is reached when the parser is trained and tested on the whole dataset prior to any changes. In the actual experiments, parsing is done on relevant subsets (see section 2.3) and reference LASes vary, but typically lie around 0.900.

redundancy of grammemes (column 2). Prior to both experiments, we remove all information about word *forms* from the input, replacing every form with the corresponding lemma (see below for the rationale).

In the first experiment, all information about the given category is deleted from the "Features" column (see Table 2). If, for instance, we are interested in the OCS NUMBer category (which includes three grammemes: **s**ingular, **d**ual and **p**lural), then a subset containing all sentences which have at least one token with number among its features is created. The parser is trained on the whole dataset and tested on this subset, providing the reference LAS (0.907 for number). After that, the strings *NUMBs*, *NUMBd* and *NUMBp* get deleted (in Table 2, that would affect rows 1, 3 and 4) in both sets. If the removal of the category leaves the "Features" column of a given token empty, the string *INFLn* is inserted (i.e. the token is marked as non-inflecting). After the deletion, the parser is trained again on the whole dataset and tested on the "number" subset. The difference between the reference LAS and the LAS after the deletion (0.907 - 0.903 = 0.004) serves as the measure of functionality. This measure is reported in Table 1 (column CF).

In the second experiment, we deal with grammemes. Simple deletion is not a suitable solution both for technical and theoretical reasons. The disappearance of a grammeme almost always means that this grammeme *merged* with another one (as the dual merged with the plural in most Slavic languages). We model this process in the following way: every grammeme within a category (say, s within number) is successively merged with every other grammeme within the same category (d and p in this case). Technically, it means that the string *NUMBs* is always replaced by *NUMBd* during the s-d merger (mergers are symmetric: s-d is equivalent to d-s) and *NUMBp* during the s-p merger. As with category deletion, a subset is created, which contains all relevant sentences (for the s-d merger, all sentences which have at least one token either in the singular or in the dual form), and both before and after the merger the parser is trained on the whole dataset and tested on the subset (for the s-d merger, the reference LAS is 0.9065). The differences in parser performance for each merger are summed and divided by the number of mergers (for s, that would mean summing across the s-d and s-p mergers and dividing by two: ((0.9065 - 0.905) + (0.9065 - 0.904))/2 = = 0.002), the result is considered a measure of grammeme functionality. This measure is reported in Table 1 (column GF).

Note that these changes affect only the "Features" column. When we merge s and d, we only change their morphological descriptions. We are, however, unable to merge the word forms in a reasonable way (partly due to large form variation within OCS, partly due to the absence of form-generating software).

Thus, if we leave the forms as they are, the merger would not be complete: the parser would still potentially be able to see that there are systematic differences between the singular and the dual forms and use this information. In order to enhance the merger's impact, we perform all experiments after deleting all information about word forms (see above).

3. Results and discussion

Results are presented in Table 1 (see also supplementary materials for more detailed data). Most of the zeroes are clustered in the lower part of the table, and it seems that GF (grammeme functionality) positively correlates with the grammeme's survivability, while for CF (category functionality) the correlation is negative. The former observation is expected, but the latter one is quite surprising. A possible explanation is that the most functional categories are also the largest ones in terms of grammeme number (case has 7, tense has 4). This means that the competition between the grammemes can be higher, or, to put it another way, the sheer probability of a merger is higher. In addition, these categories have more resources to sacrifice, both in terms of grammemes and functionality.

A notable exception from the general trend are the Bulgarian and Macedonian cases, located high in the table. One of the important reasons for their loss most likely is the intense long-term language contact within the Balkan Sprachbund (Wahlström, 2015).

A reviewer asked whether other Slavic languages had experienced less contact. The answer depends on what type of contact we have in mind. In a sprachbund, the contact is long-term, co-territorial and likely to involve child bilingualism. This specific type of contact can favor complexification through additive borrowing (Trudgill, 2011). While Bulgarian and Macedonian did lose the nominal cases, they are the only ones among Slavic languages that developed a definiteness category. It is likely that its development has contributed to the case loss and, in its turn, has been facilitated by the contact (Wahlström, 2015).

Going back to the question, Bulgarian and Macedonian have definitely experienced much more sprachbund-type contact than any other Slavic language. However, as regards shorter-time contacts that involve adult bilingualism and thus are likely to favor simplification (Trudgill, 2011), these two languages score relatively low. For example, Bentz & Winter (2013) estimate the proportion of non-native speakers as 21% for Bulgarian, 42% for Russian, 52% for Serbian (but 26% for Croatian, which is listed separately), 3% for Polish.

It is remarkable that the resultative tense, which was very infrequent in OCS, but survived in all Slavic languages (and became the only past tense in many of them), gets high CF and GF values.

We tested the correlations by means of the mixed effects logistic regression with survival (1 or 0) as the dependent variable, using R (R Core Team, 2015) and *lme4* (Bates et al., 2015). Due to the small size of the dataset and the nature of the dependent variable (categorical and not continuous) we face severe convergence problems, which makes it difficult to apply Barr et al.'s (2013) recommendation to "keep it maximal" and, in some cases, to apply likelihood ratio tests (Wald-z statistic is used instead). We try to keep the model as simple as possible, including only the most important predictors as fixed effects, viz. (centered) CF and GF. Since the maximal model does not converge, we exclude random slope for CF (which is presumably a less important predictor). The final model (henceforth Model 1) includes language and grammeme as random effects with by-language and by-grammeme random slopes for GF. More complex theoretically-justified models either do not improve the goodness of fit (AIC) or do not converge.

According to Model 1, GF increases the logit estimate by 10.37 ± 3.30 standard errors (Wald-z = 3.14, p = 0.0017), while CF decreases it by -2.21 ± 0.62 standard errors (Wald-z = -3.59, p = 0.0003).

The random slopes and intercepts are highly correlated both for grammemes and languages. The analysis of the coefficients does not reveal any interesting patterns for grammemes. As regards languages, Bulgarian and Macedonian get much lower coefficients than others both as intercept (2.66 for both languages vs. mean 8.41, sd 3.21) and as slope (0.61 vs. mean 9.34, sd 4.92). This means that the impact of GF on survivability is much lower in these languages, and the average survivability of grammemes is lower. The real effect behind this is the fate of the case system, which is poorly predicted by our measures. Going back to the discussion of contact types, it can be that the sprachbund influence somewhat shields Bulgarian and Macedonian from the pressure to shed redundant complexity, while imposing some other pressures such as the influence of the neighboring languages.

One might wonder if the random slope coefficients correlate positively with the population size or the share of L2 speakers. This correlation would imply not only that the impact of functionality is different across languages, but also lend support to the language complexity theories cited in Section 1. We find, however, no significant correlation between (modern) population sizes and random slopes. We did not collect data about the shares of L2 speakers or other social parameters that might be important for language complexity, though this is a promising avenue for future studies.

It is an important question whether the same results can be achieved using simpler predictors, first of all frequency (Table 1, column “freq”). Frequency and GF are not collinear, but seem to have similar impact. Simply replacing the fixed effect of GF by that of frequency in Model 1 (henceforth Model 2) results in slightly worse goodness of fit (AIC 127 vs. 120), but the change is not significant. If, however, we try to abandon GF and CF altogether and build a model using frequency only, then the best model (frequency as a fixed effect, random intercepts for grammeme and language, by-language random slope for frequency, no random correlation for grammeme, henceforth Model 3) is significantly worse than Model 1 ($\chi^2(3)=53.05$, $p < 0.0001$). Somers' D_{xy} and the C index are also slightly worse, resp. 0.978 and 0.989 for Model 1, 0.974 and 0.987 for Model 2, 0.923 and 0.961 for Model 3.

4. Conclusion

We show that the functionality of a morphological feature, measured as the importance of this feature for an automatic parser, is a significant predictor of survivability of the feature. Higher functionality of a grammeme increases its chances to survive, while higher functionality of a category, unexpectedly, decreases the chances of the respective grammemes to survive, although the slope is not as steep as in the former correlation.

The ablation technique that we described has several limitations, and ignores some potentially important factors. These problems can probably be partly overcome, but even in its current form the method can explain some of the morphological variation observed in modern languages and is thus a useful tool to test theories about language change, evolution and diversity.

Acknowledgements

The authors were supported by the Norwegian Research Council grant “Birds and Beasts” (222506). We are grateful to the CLEAR research group for discussing earlier versions of the paper.

References

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

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package ver. 1.1-9.

Barr, D., Levy, R., Scheepers, C. & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3), 255–278.

Bentz, C. & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change* 3, 1–27.

Berdicevskis, A. (2015). Estimating Grammeme Redundancy by Measuring Their Importance for Syntactic Parser Performance. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 65–73. Association for Computational Linguistics.

Comrie, B. & Corbett, G. (eds.) (1993). *The Slavonic Languages.* London: Routledge.

Haug, D., Jøhndal, M., Eckhoff, H. Welo, E., Hertzenberg, M., & Müth, A. (2009). Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *Traitement Automatique des Langues 50(2),* 17–45.

Kiparsky P. (1982). *Explanation in phonology*. Dordrecht: Foris.

Lloyd, P. (1987). *From Latin to Spanish: Historical phonology and morphology of the Spanish language*. American Philosophical Society.

Lupyan, G. & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE* 5(1):e8559.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering 13(2)*, 95–135.

R Core Team (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sampson, G., Gil, D. & Trudgill, P. (eds.) (2009). *Language complexity as an evolving variable*. Oxford: Oxford University Press.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press.

Wahlström, M. (2015). *The loss of case inflection in Bulgarian and Macedonian.* Helsinki: Slavica Helsingiensia.