Aspect coding asymmetries of verbs: The case of Russian

Giuseppe G. A. Celano Michael Richter Rebecca Voll Gerhard Heyer
Leipzig University
Natural Language Processing Group
{celano,richter,heyerasv@informatik.uni-leipzig.de}
rebecca_voll@yahoo.de

Abstract

This paper presents preliminary corpus-based evidence from Russian for an "aspectual coding asymmetry". The main research question is: Can different lengths of aspectual verb forms be predicted? We assume that each verb has a default aspectual value and that this value can be estimated based on frequency, which according to Zipf (1936) has a negative correlation to length. Our study provides evidence that the aspectual default value is a better predictor of lengths of verb forms in Russian than frequency. In addition, we observed a positive but weaker impact of information content (Cohen Priva, 2008; Piantadosi et al., 2011), estimated from the verbs' syntactic dependents. A final result is the tendency for the impact of frequency to level out as IC increases.

1 Introduction

The aim of the present study is to present the results of a preliminary research investigating changes in the lengths of word forms in different verb forms in Russian. We choose Russian because of its elaborate aspect system and because aspect is overtly marked in Russian verbs. We compare the effects of the following three predictors on average imperfective and perfective verb form lengths: *aspectual default/non-default coding* (0-coding/1-coding), *frequency*, and *information content* (henceforth IC). IC can intuitively be interpreted as the amount of surprisal when encountering a verb form. Hale (2001) and Levy (2008) define surprisal as the negative logarithm of the (conditional) probability of a word. In this study, we essentially follow this definition (see section 3.1 below). Since there is some empirical evidence that word length correlates negatively with frequency (Zipf, 1936), frequency is chosen as a control predictor and we aim to test the impact of IC, as defined and utilised by (Piantadosi et al., 2011), and the impact of "aspectual default/non-default coding" against this predictor. The method used here is a linear regression analysis employing a mixed–effect model (Bates et al., 2015). Consider the Russian verb сообшать (transliteration 'soobshat') "report". This verb occurs predominantly in the perfective aspect. This means, according to Zipf, that its imperfective forms should be longer. And indeed, for both verbs we find significant differences in length between the shorter perfective and the longer imperfective forms: The average length in the perfective aspect is 7.4 characters and in the imperfective aspect it is 8.2 characters. Our study is in general motivated by observations of coding asymmetries in grammars (e.g. Greenberg (1966); Croft (2012); Haspelmath (2008)), the *Form-Frequency-Correspondence-Principle* (Haspelmath et al., 2014), and the question of whether this principle suffices to explain coding asymmetries in languages. The research questions are in particular based on findings of Piantadosi (2011) and Levshina (2017), who both claim that IC is the strongest predictor of words' lengths in general, utilising unstructured co-occurrences of verbs (Piantadosi et al., 2011), and verb dependents (Levshina, 2017). Like Levshina (2017), we estimate IC from verb dependents, but unlike Piantadosi (2011) and Levshina (2017), we use IC to predict lengths of aspectual verb forms. This leads us to challenge their conclusion regarding IC as the strongest predictor. The central linguistic concept of our study is aspect. Aspect can be connected to time in tenses. However, whereas time places an event on a timeline relative to a given reference point, aspect defines the (temporal) perspective taken on an event (see Velupillai (2012)). The two major aspect categories are perfective and imperfective. Whereas a perfective form expresses reference to a single, completed, particular event, an imperfective form is void of any semantic commitment to whether or not the event is single, completed, or particular. These features allow imperfectives to refer to single, unbounded events that are only partially realized at reference time, to refer to multiple events, and to refer to non-particular (generic) events. It is generally assumed that aspect can be expressed both lexically and grammatically (see Croft (2012) for a detailed analysis). While lexical aspect, also known as *Aktionsart*, is taken to be an inherent property of verb meaning (regardless of any specific grammatical realization), grammatical aspect pertains to the way aspect is encoded by means of grammar (typically morphosyntactically), which can vary depending on how a speaker decides to construe an event. Look at the following example:

(1) a. I love vegetables.
    b. I am loving vegetables.

The lexical aspect of the English verb *love* can be defined as "state" (see Vendler (1967) for a definition of actional classes). Accordingly, in the present it is usually expressed grammatically by the simple present verb form (see 1a), whose function is, among other things, to express imperfective aspect. Notwithstanding, speakers are free to conceptualize the same event differently, and employ, for example, the present progressive (1b). In this case, the event is still in the present, but it is presented as an activity rather than a state. The use of the present progressive is arguably rather exceptional with 'state' verbs, but it is not impossible. The relationship between lexical aspect and grammatical aspect is a notoriously very complex one, both because it is not always easy or even possible to decide the lexical aspect class of a given verb lemma by abstracting from specific instantiations in speech/text and because grammatical tenses such as the present or the present progressive in English are complex structures expressing many bits of information conjointly, whose interpretation can also be strongly dependent on syntactic context. If a way to determine the lexical aspect for each given verb lemma in a language could be found, one could test the hypothesis that, among the word forms of a given verb lemma, those expressing the lexical aspect are, on average, shorter than those expressing any other "non-default" aspect category. In fact, this hypothesis is a particular instantiation of a larger one whereby higher frequency linguistic items are more predictable than lower frequency items and, since predictable content arguably needs less salient formal representation, high frequency items are expected to be expressed by shorter forms. This (binary) opposition, also known in the linguistic literature as "coding asymmetry", has so far been investigated mostly qualitatively in (theoretical and typological) linguistics (see Haspelmath (2008) for a detailed overview). The "aspect coding asymmetry" hypothesis, which is put forward in this article, restricts the notion of coding asymmetry to aspect encoding. We claim that the binary aspect distinction between perfective and imperfective verb forms in Russian is related to word length, in such a way that (i) the most frequent aspect category for each given verb lemma (if it is statistically significant) is assumed to express the verb lexical aspect (henceforth the "default aspect"), and (2) the verb forms of a specific lemma which are associated with its default aspect are expected to be on average shorter than the verb forms of the same lemma which are associated with the other (non–default) aspect category.

2 Related Work

IC is crucial for our language model. It is based on conditional probabilities (the estimation is given in section 3 below) and is a variant of conditional entropy. Language models which are based on conditional probabilities are employed in parsing tasks (cf. Demberg et al. (2013); Hale (2001)) and (Demberg et al., 2013) argue that those probabilistic models have cognitive plausibility since humans tend to make predictions from contexts when they parse natural language (cf. Altmann and Kamide (1999)). Hale (2001) and Levy (2008) use the concept of *surprisal* to describe the effect of conditional probabilities of words on (human) sentence processors. Surprisal is a concept from information theory and, as pointed out above, corresponds to IC. (Piantadosi et al., 2011) investigate the correlation between IC and word length. They show for 10 Indo-European languages that IC, calculated on the basis of syntactic contexts (bigrams, trigrams, and fourgrams), is a much better predictor of word length than frequency. According to them, the effect of frequency is largely due to its correlation with information content. Piantadosi et al. (2011) ascribe the attested correlation of word length and information content to the principle of uniform information density, which says roughly that the information rate of communication over time is kept as constant as possible. They point out that IC is known to influence the amount of time speakers take to pronounce a word and conclude by suggesting Zipf's law in the following way: "the most communicatively efficient code for meanings is one that shortens the most predictable words - not the most frequent words" (Piantadosi et al., 2011).

More recently, Levshina (2017) has investigated whether the length of words can be predicted if the IC score is calculated based on syntactic dependents rather than co-occurrence frequencies (n-grams), as in Piantadosi et al. (2011) . Levshinas's study (2017) confirms the hypothesis that words with higher IC tend to be longer in most languages. As a data resource, we use the Russian treebank SynTagRus (Nivre, Agić, and Ahrenberg et al. (2017) ; see section 3.1 below) which contains information about verbal aspect. Ramm et al. (2017) present a tool for automatic annotation of amongst others aspectual information, however only for English, French and German. These languages are however not in the focus of our study since overt aspect coding is sparse in these languages.

3 Method

3.1 Data

The data for the present study derive from the Russian treebank SynTagRus available in Universal Dependency 2.1 (Nivre, Agić, and Ahrenberg et al. (2017)). The SynTagRus corpus is a relatively large corpus for modern Russian, whose texts belong to a variety of genres, such as contemporary fiction, newspapers, and online news, dated between 1960 and 2016. The corpus contains manually corrected morphosyntactic annotation. Table 1 provides some information about SynTagRus.

SynTagRus	**size**
Tokens	1,107,182
sentences	61,889.
verb lemmas	571
verb tokens	56,828
imperfective tokens	29,514
perfective tokens	27,314

Table 1: The SynTagRus corpus.

For each verb lemma contained in the treebank, we grouped all of its occurrences into imperfective and perfective word forms, and , using a totally automated procedure, chose only those verb lemmas for which the difference between the number of imperfective and perfective forms is statistically significant (relying on a Pearson's Chi-squared test with *p-value* < 0.05). This test provided us with a criterion to determine whether we could consider the more common aspect class as the default one (with the value '0'). Such a line of reasoning is based on the rough assumption that imperfective and perfective word forms can be assigned to the atelic and telic actional classes, respectively. Lexical aspect is a controversial linguistic notion and it is difficult to provide unambiguous (cross-linguistic) criteria for the classification of any given verb lemma. However, a corpus-based query aimed to identify statistically significant frequency differences may provide an operational criterion to determine whether a verb lemma should be expected to be imperfective or perfective (under the assumption that the inherent aspect class for a verb lemma will be the most frequent one). At the present stage of our research it is not yet clear whether or how we can exploit the syntactic annotation contained in our corpus in order to get a more finely grained actional classification. After identification of the default form for each verb lemma, we calculated the average length of default (0) and non-default (1) word forms and created a table where the data for each lemma are conveyed by two rows: one contains the lemma_0 and the other the lemma_1 (associated to a LEMMA and a DEFAULT column variables). Besides the average word length (CHAR_NUMBER variable), we calculated, for each row, the IC (IC variable) and negative log frequency (FREQUENCY variable). For the calculations of IC we used the same formula as Piantadosi et al. (2011):

$$IC = -\frac{1}{N}\sum_{i=1}^{n} log\left(P\left(W = w|C = c_i\right)\right) \qquad (1)$$

The formula is a summation of the logs of a conditional probability of a word w, which, in our case, is represented by a lemma_0 or lemma_1 and by a context c, which we define as the syntactic labels of the direct dependents of a verb. Whereas Piantadosi et al. (2011) define context as n-grams, we were

Syntactic Context	Size
nsubj,obl	5,628
Obl	5,024
obj	4,366
obj,obl	2,322
advmod, nsubj, obl	2,260
nsubj, obj	2,158

Table 2: The six most frequent syntactic contexts.

able to extract syntax by using the explicit (semi-automatic) morphosyntactic annotation provided in the UD treebank. It is to be noted that, although Levshina (2017) also used UD, our IC differs from hers in two important respects: (i) she based her IC on the dependents of any token in UD while we focus only on verbs, and (ii) we decided to apply a filter to verb dependents: since it is known that some syntactic dependents are more difficult to be annotated and would therefore cause inconsistencies within a corpus, we only selected those direct verb dependents that we assume can be annotated more easily (and therefore, arguably, more consistently). The verb dependents that we selected in the end are the direct ones with the following syntactic labels: 'nsubj', 'nsubj:pass', 'csubj', 'csubj:pass', 'obj', 'iobj', 'advmod', 'advcl', 'obl', 'xcomp', and 'ccomp' (see UD documentation for definitions). Since one verb can directly govern more than one dependent carrying one of the above-mentioned syntactic labels, the syntactic con-text has to be thought of as an n-tuple of one or more of these syntactic labels (see Table 2).

After creating a table in the way described above, we used it as an input in order to calculate a linear mixed-effect model, which is described in the following section.

3.2 Analysis

We analysed effects on the average lengths of verb forms (lemma_0 and lemma_1) using linear mixed-effect modeling (R package lme4; see Bates et al. (2015)). The response variable is CHAR_NUMBER, while the predictors are IC, FREQUENCY, and DEFAULT. For FREQUENCY, we took the negative logarithm of the probabilities of word forms that is, the surprisal. We controlled for LEMMA by defining it as a random effect in our model since the absolute difference between lengths depends on the specific lemma. The predictors were negatively tested for collinearity. This indicates that DEFAULT is not a masked FREQUENCY-predictor. We successively extended the models, the criterion being a significant decrease of the deviance score: adding IC, FREQUENCY, DEFAULT and the interaction between IC and FREQUENCY resulted in the best model. We tested three different types of syntactic context: (i) one containing all the direct dependents listed in Section 3.1, (ii) one containing all arguments except for the subject and nothing else ('obj', 'iobj', 'xcomp', and 'ccomp') and (iii) one containing all arguments including the subject ('nsubj', 'nsubj:pass','csubj', 'csubj:pass', 'obj', 'iobj', 'xcomp', and 'ccomp').

4 Results

The model for IC containing all verb dependents in (i) outperformed the remaining two models. The model summary is given in Table 3.

Random Effects:		
Groups	Names	σ^2
LEMMA	(INTERC.)	3.835
Residual		1.347

Fixed Effects:					
	Estim.	ε	df	t	p
Interc.	4.87	1.18	1019	4.13	≈0***
IC	0.47	0.21	867	2.31	0.02*
FREQ	0.47	0.14	978	3.24	0.001**
DEF1	0.44	0.11	973	4.19	≈0***
IC:FREQ	-0.05	0.02	822	-2.05	0.04 *

Table 3: Summary of the best regression model.

The most significant fixed effect is DEFAULT. When DEFAULT has the value '1' the average lengths of verb forms are predicted to increase by .43. FREQUENCY and IC are both significant as well, the former slightly outperforming the latter. Both effects predict an increase of verbs lengths by .47, if their values increase. Figures 1 displays the impacts of the three fixed effects: Longer verb forms correspond with 1-coding, lower frequency and higher IC.

Table 3 discloses that the interaction of FREQUENCY and IC is significant as well, which has

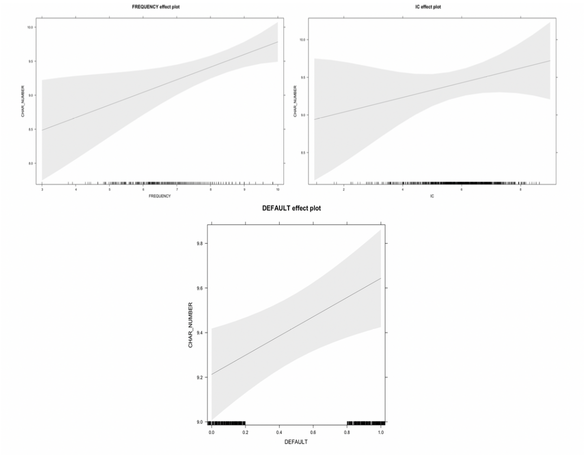

Figure 1: Effects of DEFAULT, FREQUENCY and IC on CHAR_NUMBER.

already been observed by Piantadosi et al. (2011). Figure 2 illustrates the interaction.

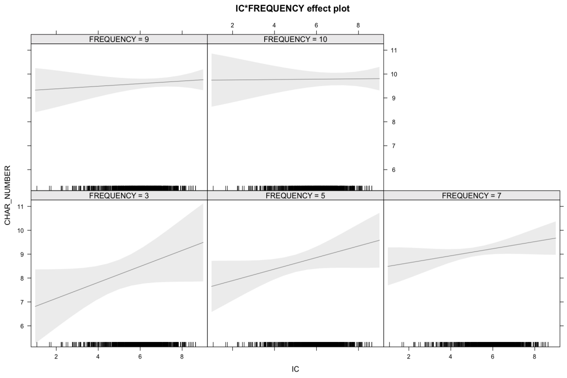

Figure 2: Effect of the interaction of IC and FREQ on CHAR_NUMBER.

The effect of FREQUENCY weakens as IC increases. That is to say, when IC is low, i.e. IC ≈ 3, the impact of FREQUENCY on CHAR_NUMBER is high. In contrast, when IC ≈ 10, i.e., the maximum in our study, the effect of FREQUENCY is almost annihilated, while the length of the verb forms is constantly high. Figure 2 displays that the higher IC, the longer the verb forms and the less the influence of FREQUENCY.

5 Conclusion

We addressed the research question whether the average lengths of verbs in Russian can be predicted based on the aspect distinction default/non–default (with 'default' corresponding to the predominant verb lexical aspect), frequency, and IC estimated from verb dependents. In contrast to Piantadosi et al. (2011), we used the verbs' syntactic dependents as contexts (and not n-grams) and, in contrast to Levshina (2017), we narrowed down our research to verbs. We tested our hypothesis using a linear mixed-effect model. In contrast to earlier studies (Piantadosi et al. (2011) and Levshina (2017)), IC does not outperform frequency in our study. Instead, in our experimental setting FREQUENCY has a slightly higher impact on verb length than IC and the variable DEFAULT is an even stronger predictor of verb length than both IC and FREQUENCY.

An interesting finding is the significant interaction of IC and frequency, when controlling for LEMMA: we found that when verbs have a high IC, the impact of the predictor FREQUENCY on length tends to be small, while with non-informative verbs the impact of FREQUENCY tends to be high. It is known from typological research that frequently expressed meanings tend to be expressed by short forms and this is represented in our model: FREQUENCY is a significant predictor of verb length. Thus, in general, when a verb form is frequent it tends to be short, while long verb forms have a high FREQUENCY-value. This effect has impact on length only with non-informative verb forms. But when IC is high, the verb forms tend to be long, no matter whether they are frequent or non-frequent.

This is economical: An unexpected word form causes an informative surprisal, and the predictor FREQUENCY is not needed. Expected verb forms have to rely on the predictor FREQUENCY, since IC plays a minor role.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: 357550571.

References

Gerry T. M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker.. 2015. Fitting Linear Mixed- Effects

Models using lme4. *Journal of Statistics Software*, 67(1):1–48.

Uriel Cohen Priva. 2008. Using information content to predict phone deletion. Natasha Abner and Jason Bishop (eds.): *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages: 90 – 98, Cascadilla Proceedings Project, Somerville, MA.

William Croft. 2012. *Verbs: Aspect and causal structure*. Oxford University Press, Oxford, UK.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Joseph H. Greenberg. 1966. *Language universals, with special reference to feature hierarchies*. Mouton, The Hague.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, pages: 1–8.

Martin Haspelmath. 2008. Creating economical patterns in language change. In Jeff Good (ed): *Linguistic universals and language change*. pages: 185 – 214. Oxford University Press, Oxford, UK.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog , and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics* , 50(3):587–625.

Natalia Levshina. 2017. Communicative efficiency and syntactic predictability: A crosslinguistic study based on the Universal Dependencies corpora. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*, (UDW 2017). Gothenburg, pager: 72 – 78.

Roger Levy. 2008. Expectation–based syntactic comprehension. *Cognition*, 106(3):1126–117.

Joakim Nivre, Željko Agić, and Lars Ahrenberg et al. 2017. Universal Dependencies 2.0 – CoNLL 2017 Shared Task Development and Test Data, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÃŽFAL) Faculty of Mathematics and Physics, Charles University, Prague.

Lars Johanson. 2000. Viewpoint operators in European languages. In Östen Dahl (ed): *Lense and Aspect in the Languages of Europe*, pages: 127 –187. Mouton de Gruyter, Berlin.

Steven T. Piantadosi, Harry Tily, and Edward Gibson 2011. Word lengths are optimized for efficient communication. *PNAS*, 108(9):3526–3529.

Anita Ramm, Sharid Loáiciga, Annemarie Friedrich and Alexander Fraser. 2017. Annotating tense, mood and voice for English, French and German. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, demo session (ACL), Vancouver, Canada.

Viveka Velupillai. 2012. *Introduction to Linguistic Typology*. John Benjamins, Amsterdam.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca.

George Zipf. 1936. *The Psychobiology of Language*. Routledge, London.