A mixed-effects model for the imperfective-perfective coding asymmetry with a random effect for lemma

Anonymous EMNLP submission

Abstract

The article presents the results of a mixed-effects model fit with a random effect for lemma aimed to investigate, in 18 UD treebanks, the relationship between the imperfective-perfective aspect difference and word length. Default aspect values for each lemma in a treebank are identified via χ^2 tests and, on the basis of them, the hypothesis is tested whether there exists a coding asymmetry such that default aspect values relate to shorter word lengths. The study shows that such a coding asymmetry cannot be substantiated as a language universal. The mixed-effects model also regresses word lengths on frequency and information content, the latter being defined by a selection of syntactic labels of verb direct dependents. For 6 treebanks no significant relationships are either found for these predictors, which seems to suggest that an aspect-related coding asymmetry may actually not exist.

1 Introduction

In aspect-related studies a bidimensional approach is common (Persohn, In press), whereby a distinction is made between *aktionsart*, which is a lexical semantic notion and grammatical aspect (i.e., aspect as a morphosyntactic device). Grammatical aspect and lexical aspect interact (Persohn, In press; Sasse, 2006; Velupillai, 2012), even if their semantics is different (Johanson, 2000). The grammatical aspect of a verb/predicate tends to correlate with the verb/predicate's aktionsart. One such pattern of interaction is the correlation between telicity and perfectivity. Telicity describes whether or not a verb/predicate has an inherent goal or endpoint. Telic verbs/predicates do have an inherent goal or endpoint, whereas atelic verbs/predicates do not. Regarding (im)perfectivity, the perspective of the perfective aspect is on the whole event, while that of the imperfective aspect is within the event. (Velupillai, 2012, p. 209–210).

One can therefore expect that verbs that are intrinsically telic occur in the perfective and verbs meaning states or processes in the imperfective (Timberlake, 2007, p. 292-293). More precisely, telic verbs are not expected to always occur in the perfective and atelic verbs always in the imperfective; rather, telic verbs are expected to occur more frequently in the perfective than in the imperfective, while atelic verbs are expected to occur more frequently in the imperfective than in the perfective. This asymmetry, which has been described by various scholars (e.g. Persohn, In press; Sasse, 2006; Velupillai, 2012; Timberlake, 2007), is a frequency asymmetry.

In order to capture this asymmetry in our study, we identify default aspect values for each lemma in 18 UD treebanks on the basis of statistically significant differences between a lemma's occurrences as imperfective and perfective, and investigate whether such a variable is a predictor for word length. If this proved to be true, we could argue for the existence of an aspect-related coding asymmetry, whereby default aspect values are expected to relate to shorter word length.

Indeed, coding asymmetries are a special instantiation of the well-known grammatical form-frequency correspondence principle as formulated by Haspelmath et al. (2014), which ultimately goes back to Greenberg (1966): "When two grammatical patterns that differ minimally in meaning (i.e. patterns that form a semantic opposition) occur with significantly different frequencies, the less frequent pattern tends to be overtly coded (or coded with more coding material), while the more frequent pattern tends to be zero-coded (or coded with less coding material)". Applied to our case, Greenberg's principle would suggest that, if there is a statistically significant frequency difference

between imperfective and perfective word forms for a given lemma, the verb forms which are associated with the most frequent aspect category, i.e., the ones showing default aspect, should be (on average) shorter than the ones having non-default aspect.

The grammatical form-frequency correspondence principle can also be interpreted as an instantiation of the more general form-frequency correspondence known as Zipf (1935)'s law, according to which "languages tend to use less coding material for more frequent expressions" (Haspelmath et al., 2014, p. 23). Zipf's law has been been argued to apply to a great number and variety of languages (Bentz and FerreriCancho, 2016). Earlier studies (Piantadosi et al., 2011; Levshina, 2017) have also tried to specify Zipf's law further, suggesting that information content (IC) conveyed by words is an even better predictor for word length than frequency.

In the present paper, therefore, we present the results of a mixed-effects model fit, where word lengths, calculated as the average lengths of all imperfective and perfective word forms within a lemma, are modeled as a dependent variable to regress on four predictors: default aspect value, frequency, information content, and the interaction between (relative) frequency and information content. In Section 2 we detail the data and the method employed. We report the results in Section 3, while in Section 4 we discuss them.

2 Data and Method

Our research is based on the data contained in UD treebanks (v. 2.1), which currently represent the largest annotated and typologically diverse corpus, comprising 102 treebanks and 60 languages. In a previous pilot study with UD_Russian-SynTagRus (under review), we fitted a mixed-effects model (Baayen et al., 2008) to investigate word length differences of word forms belonging to the same lemma but differing in the aspect feature (imperfective vs perfective): the underlying hypothesis was that, if one could identify a default aspect for each verb lemma, i.e., the aspect a verb is primarily associated with on the basis of its lexical value, then one could try to test whether average word lengths of imperfective and perfective word forms can be regressed on such a default value, while controlling for lemma as a random effect.

Since the experiment was successful, we decided to extend it to all the UD treebanks where aspect is encoded as a verb morphological feature. If we could successfully find models for most/all languages, this would represent an argument for the existence of an aspect-related coding asymmetry, whereby, in the binary opposition between imperfective and perfective, the shorter word form is associated with the default aspect value, while the longer word form with the non-default aspect.

In the UD corpus only 45 treebanks contain annotation for the morphological feature "Aspect". 12 of these treebanks, however, cannot be used for our study: on the one hand, all Czech treebanks (UD_Czech, UD_Czech-CAC, UD_Czech-CLTT, UD_Czech-FicTree, UD_Czech-PUD), UD_Gothic, UD_Sanskrit, UD_Hungarian, and UD_North_Sami have different lemmas for imperfective and perfective verb forms, and therefore it is impossible to compare their word length differences while keeping the lemma as a constant; on the other, UD_Arabic-NYUAD does not contain word forms for copyright reasons, while the UD_Chinese treebanks (UD_Chinese and UD_Chinese-PUD) have aspect encoded not on verbs.

In order to render the data crosslinguistically uniform, aspect oppositions within each treebank are reduced to the imperfective-perfective one, which can be found in all languages displaying morphological aspect. More precisely, the habitual and progressive aspects are subsumed under the imperfective and the resultative aspect under the perfective, while verb forms in the prospective aspect have been ignored, its value not being clear with respect to the imperfective and perfective opposition[1].

This aspect reduction concerns the following treebanks (the original aspect categories are reported within parentheses): UD_Basque ("Perf", "Imp", "Prog", "Prosp"), UD_Buryat ("Imp", "Hab", "Perf", "Prog"), UD_Hindi-PUD ("Perf", "Imp", "Prog"), UD_Kurmanji ("Prog", "Perf"), UD_Marathi ("Imp", "Perf", "Prosp", "Hab"), UD_Old_Church_Slavonic ("Perf", "Res", "Imp"), and UD_Turkish-PUD ("Perf", "Hab", "Prog", "Imp", "Prosp"). In our preliminary investigation, we also exclude UD_Slovak ("Perf", "Imp", "Imp,Perf") because of the unclear "Imp,Perf"

[1] one can also argue that the prospective as defined in UD (http://universaldependencies.org/u/feat/Aspect.html) is not really a type of aspect, in that its definition centres around relative time in reference to another action

label and UD_Turkish ("Perf", "Prog", "Imp", "Rapid", "DurPerf", "ProgRapid") because of the categories "Rapid", "DurPerf", and "ProgRapid".

After all aspect categories are leveled to the imperfective-perfective opposition, all verbs within a treebank are grouped by lemma and, subsequently, by aspect. This allows us to count all the imperfective and perfective word forms and perform Pearson's χ^2 tests to determine whether differences between the two aspect categories of each lemma are statistically significant ($p<0.05$): if so, we take the most frequent aspect category as the default one for a given lemma and calculate the (character) word length of the imperfective and perfective aspects as the average of the lengths of all the word forms within each aspect category.

We identify default aspect categories on the basis of a statistically significant frequency difference because it is not clear whether it is always possible to establish a default aspect category notionally (and, in any case, it would be infeasible). While for certain (prototypical) lemmas, such as the ones meaning "reach" or "love", lexical aspect can arguably be easily identified (they are an achievement and a state, respectively), for other verbs lexical aspect seems to be a gradient rather than a binary category.

In providing an empirical criterion, a χ^2 test does therefore seem to be well suited to address this question. On the theoretical side, we expect the lemmas passing the χ^2 test to have played the major role in determining the aspect coding asymmetry (if any). A drawback of this approach is however that many lemmas and even languages are filtered out. For example, Belarusian and Buryat cannot be included in our study because there are no lemmas were the difference between imperfective and perfective word forms can be proved to be significant. Table (1) summarizes all the treebanks that can be used for our study.

The average imperfective and perfective word form lengths for each lemma are calculated because a direct comparison of lengths of verb word forms agreeing in any morphological feature other than aspect - which would be ideal - is not always possible, not all imperfective word forms been attested in a treebank also as perfective (and viceversa). Since word length can be to a great extent affected by non-aspect related morphemes (such "person" or "number", which can change in verb conjugation), average length allows us to maximize the number of verb lemmas available for comparison.

UD treebanks	lemmas
UD_Ancient_Greek	42
UD_Ancient_Greek_PROEIL	205
UD_Arabic	181
UD_Arabic_PUD	3
UD_Basque	82
UD_Bulgarian	6
UD_Greek	58
UD_Hindi	62
UD_Latin	9
UD_Latin-ITTB	68
UD_Latin-PROIEL	164
UD_Latvian	6
UD_Marathi	6
UD_Old_Church_Slavonic	18
UD_Polish	4
UD_Russian-SynTagRus	571
UD_Slovenian	3
UD_Urdu	25

Table 1: UD treebanks with the number of lemmas for which the frequency difference between perfective and imperfective word forms is statistically significant ($p<0.05$).

We also calculate information content (IC) associated with imperfective and perfective word forms, it having been argued to be a better predictor for word length than frequency (Piantadosi et al., 2010; Levshina, 2017). Piantadosi et al. (2010)'s formula (derived from Cohen (2010)) has been adapted to our case study:

$$y_{ij} = -\frac{1}{N}\sum_{j=1}^{N} logP(A = a_{ij}|C = c_j). \quad (1)$$

In formula (1) the negative log conditional probability of a_{ij} is calculated, i.e, of lemma *i* with aspect *j* (imperfective or perfective) given the context c_j, which is defined syntactically as the set of the syntactic labels of the direct dependents of a_{ij} which are "advmod", "advcl", "obl", "nsubj", "nsub:pass", and "csubj:pass". We restrict IC to only the preceding labels because, on a theoretical basis, we expect them to be the most representative ones for each a_{ij} (i.e., we basically exclude conjuncts and dependents broadly definable as function words, which are notoriously not verb-specific). As UD syntactic annotation is not meant

to (exactly) capture verb argument structure (e.g., oblique dependents can correspond to either adjuncts or arguments), we include the labels "advmod", "advcl", and "obl" in our IC definition: in the pilot study on UD_Russian-SynTagRus (under review), this selection proved to provide the best results with our mixed-effects model, and therefore we stick to it. On the basis of the occurrences of c_j the summation is calculated and the result is divided by *N*, i.e., the total frequency of a_{ij}.

Once default aspects and their frequency and IC have been calculated for the lemmas in each treebank (both frequency and IC being calculated as $-log$), we fit a mixed-effects model with a random effect for lemma (Baayen et al., 2008) for each treebank, which is determined by the formula

$$y_j = \beta_0 + \beta_1 Def_j + \beta_2 Freq_j + \beta_3 IC_j + \beta_4 (Freq_j \times IC_j) + W_j + \epsilon_j. \quad (2)$$

In Equation (2), the average word form length of imperfective and perfective of each lemma *j* is regressed on the dummy variable *Def* (default vs non-default), *Freq*, i.e., relative frequency, *IC*, and the interaction between *Freq* and *IC*. The model also contains a by-lemma adjustment to the intercept (W_j), which represents the random effect. As the length of a verb word form is highly affected by its base form/root, which in turn affects lemma length, adding a random effect for lemma is expected to sensibly improve the model.

3 Results

The models computed for each treebank in Table (1) show that there is a significant relationship between aspect word length and default value in only 6 of the 18 treebanks (UD_Ancient_Greek-PROIEL, p<0.001; UD_Basque, p<0.001; UD_Hindi, p<0.001; UD_Marathi, p<0.05; UD_Polish, p<0.001; UD_Russian-SynTagRus, p<0.001).

Of these treebanks 5 show also a significant relationship between aspect word length and frequency (UD_Ancient_Greek-PROIEL, p<0.001; UD_Basque, p<0.001; UD_Marathi, p<0.01; UD_Polish, p<0.001; UD_Russian-SynTagRus, p<0.01). 4 treebanks turn out to have a significant relation only between aspect word length and frequency (or frequency:IC): UD_Greek (p<0.001), UD_Latin-ITTB (p<0.01), UD_Latin-PROIEL (p<0.05),

Figure 1: Character number as a function of default value for UD_Ancient_Greek-PROEIL

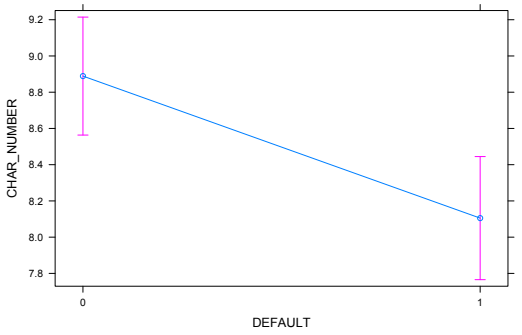
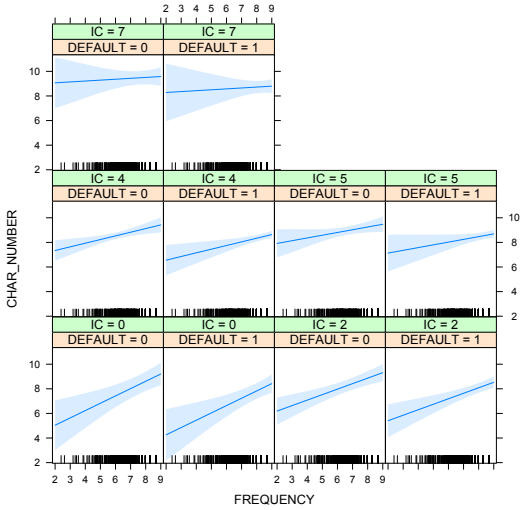


Figure 2: Mixed-effects fit for UD_Ancient_Greek-PROEIL



UD_Urdu (p<0.05); UD_Ancient_Greek and UD_Old_Church_Slavonic have a significant relationship between not only aspect word length and frequency (p<0.05 and p<0.001, respectively) but also aspect word length and IC (p<0.05 and p<0.01, respectively); UD_Ancient_Greek-PROIEL, UD_Polish, UD_Russian-SynTagRus show significant relationships for default value, frequency, and IC.

4 Discussion and Conclusion

The results we get from our mixed-effects model are, as far as language universals are concerned, rather poor. 6 of the 18 treebanks do not show any significant relationship between aspect word length and any predictor (UD_Latin, UD_Bulgarian, UD_Slovenian, UD_Arabic, UD_Arabic-PUD, and UD_Latvian). We hypothesized that the default aspect value

Figure 3: Character number as a function of default value for UD_Basque

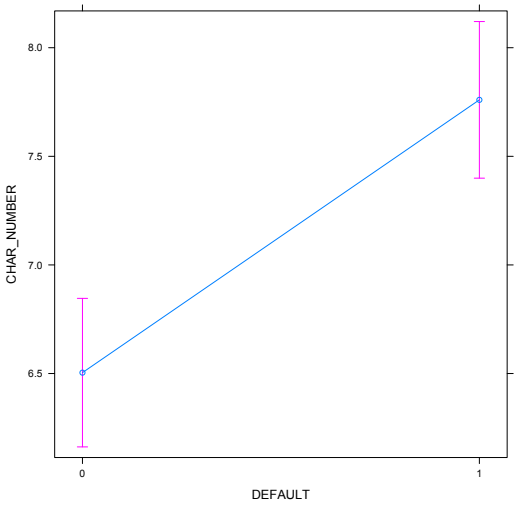


Figure 4: Mixed-effects fit for UD_Basque

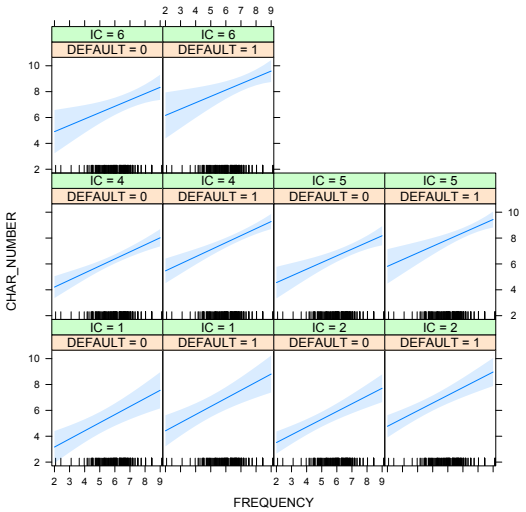


Figure 5: Character number as a function of default value for UD_Hindi

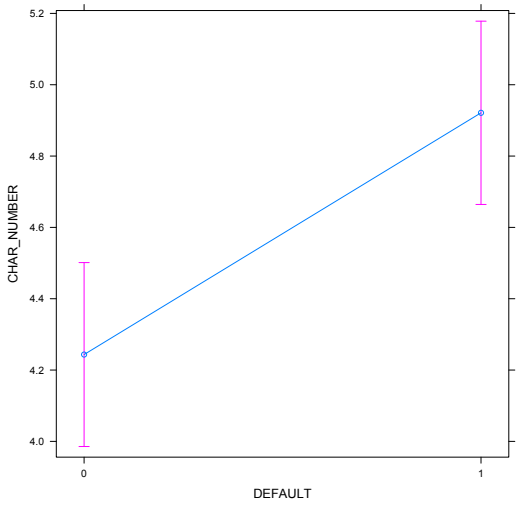
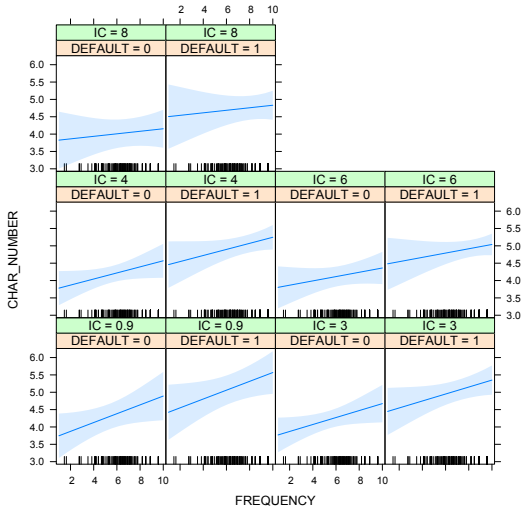


Figure 6: Mixed-effects fit for UD_Hindi



is associated with aspect word length, but only 6 treebanks provide evidence for such a relationship ($p<0.05$). Moreover, among the languages showing a statistical significant relationship between aspect word length and default value, UD_Ancient_Greek-PROEIL and UD_Polish present a negative correlation: the negative coefficient for the dummy variable DEFAULT (where 0 means "default" and 1 "non-default") provides evidence for the non-default value to be associated with a shorter word length (see Figures 1 and 9).

Interestingly, a significant relationship between frequency and aspect word length is also attested only in 11 languages, where in all but UD_Marathi the longer the word length is, the less frequent it is (see Figure 8; the slope is negative because we calculate the $-log$ value for frequency).

It has been argued that the long-studied Zipf's law concerning the relationship between word length and frequency (the longer a word is, the less frequent it is) holds for more than 900 languages (Bentz and FerreriCancho, 2016). If we therefore assume the validity of Zipf's law, the fact that it does not apply in our study to 7 languages could be taken as being in agreement with the overall lack of relationship between aspect word length and default value. Indeed, coding asymmetries can be interpreted as specifications of Zipf's law, in that a (shorter) default value in a binary linguistic opposition should be expected to occur more

Figure 7: Character number as a function of default value for UD_Marathi

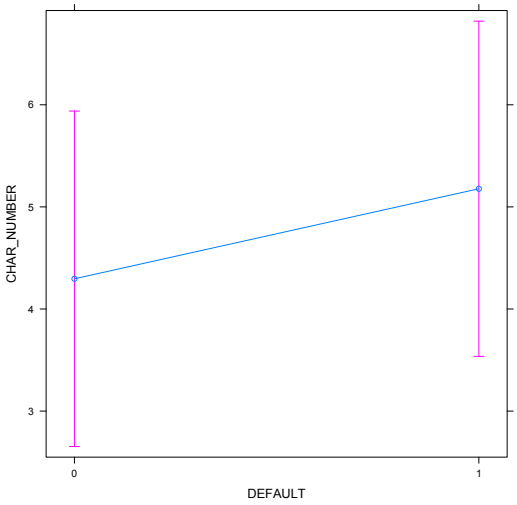


Figure 8: Mixed-effects fit for UD_Marathi

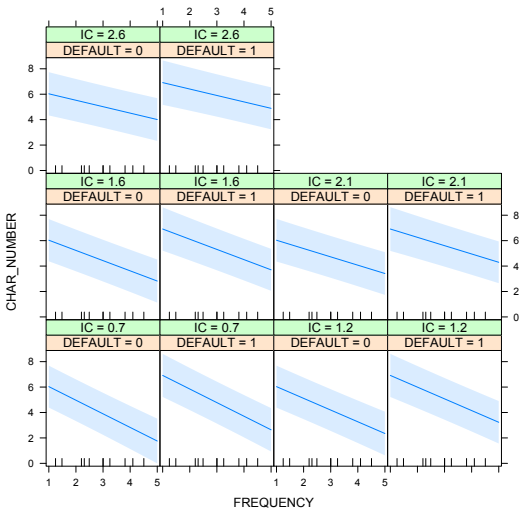


Figure 9: Character number as a function of default value for UD_Polish

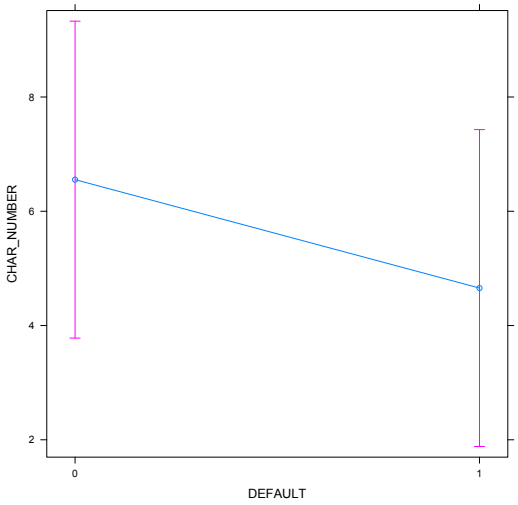
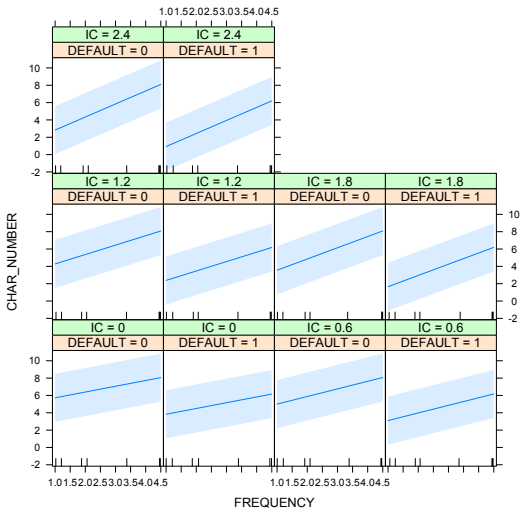


Figure 10: Mixed-effects fit for UD_Polish



frequently than the (longer) non-default one. It should be therefore not surprising that, if Zipf's law does not overall apply in our study, the related coding asymmetry does not either.

It has to be noted that, if positing an aspect-related coding asymmetry is theoretically - on pair with other clearly identifiable linguistic pairs - sound, it does not necessarily follow that such a coding asymmetry (or even, for that matter, Zipf's law) should hold: such a great variety of principles cooperate and compete to determine language form that none of them is expected to always prevail. In language universals research principles are sought which are expected to hold for a statistically significant sample of languages rather than all languages. Similarly, in coding asymmetry research, an hypothesis about a specific coding asymmetry which is not attested in a significant sample of languages can prove false without the general hypothesis about coding asymmetries being proved so. One needs more corpus-based research on different coding asymmetries before a stance can be taken.

We plan to investigate the aspect coding asymmetry further. The questions still remain open as to whether a different modelling or a different sampling could return a positive result and whether and how a default aspect value can be identified for lemmas at all. The χ^2 test we used in the first place to select lemmas has showed that a significant difference in use between imperfective and perfective word forms does not hold for a con-

Figure 11: Character number as a function of default value for UD_Russian-SynTagRus

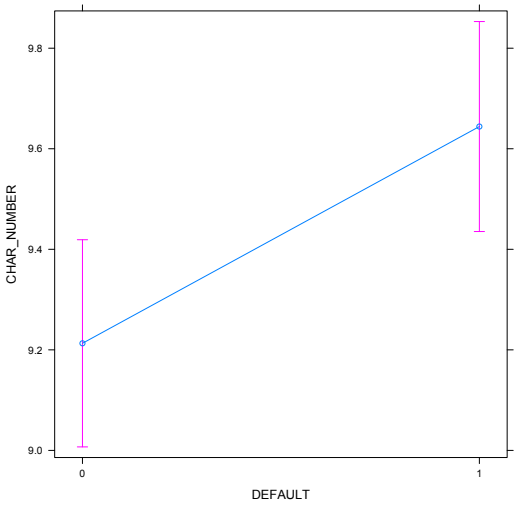
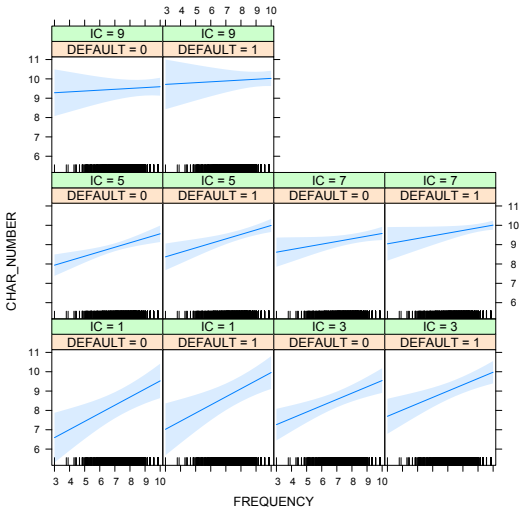


Figure 12: Mixed-effects fit for UD_Russian-SynTagRus



siderable number of lemmas, which may suggest that this opposition has not been crucial in determining word form (and therefore word length).

References

R. H. Baayen, D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Christian Bentz and Ramon FerreriCancho. 2016. Zipfs law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

Priva U Cohen. 2010. Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 90–98.

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

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

Lars Johanson. 2000. Viewpoint operators in European languages. In Östen Dahl, editor, *Tense and Aspect in the Languages of Europe*, pages 27–187. Mouton de Gruyter, Berlin.

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

Bastian Persohn. In press. Aspectuality in Bantu: On the limits of Vendlers categories. *Linguistic Discovery*.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2010. Word length are optimized for efficient communication. *PNAS*, pages 3526–3529.

Steven Thomas Piantadosi, Harry Joel Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 108 (9), pages 3526–3529.

Hans-Jürgen Sasse. 2006. Aspect and Aktionsart. In E. K. Brown, editor, *Encyclopedia of language and linguistics*, volume 1, pages 535–538. Elsevier, Boston.

Alan Timberlake. 2007. Aspect, tense, mood. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, volume 3. Grammatical Categories and the Lexicon, pages 292–293. Cambridge University Press, Cambridge.

Viveka Velupillai. 2012. *Zero coding in tense-aspect systems of creole languages*. John Benjamins Publishing Company, Amsterdam.

George Zipf. 1935. *The psycho-biology of language*. Houghton Mifflin, Boston.