Predicting default and non-default aspectual coding: Impact and density of information features*†

Michael Richter and Tariq Yousef[1]

[1] Leipzig University, Natural Language Processing Group
mprrichter@gmail.com, tariq@informatik.uni-leipzig.de

Abstract. This paper presents a study on the automatic classification of default and non-default codings for aspect-marked verbs in six Slavic and one Baltic language. As classifier a Support Vector Machine (SVM) and as verbal features *Shannon Information* (SI) and *Average Information Content* (IC) have been utilised. In all languages high accuracy of the classification has been achieved. In addition, we found indications for the validity of the *Uniform Information Density principle* within SI and IC.

Keywords: Verb aspect, coding, information content.

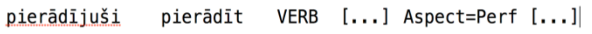
1 Introduction

The first aim of the present study is to test whether default and non-default coding of aspect-marked verbs in the six Slavic languages Bulgarian, Old Church Slavonic, Polish, Slovak, Slovenian and Ukranian and, in addition, the Baltic language Latvian can be automatically classified by two verbal information features that is, (i) *Average Information Content* (henceforth 'IC') ([1], [2]), and (ii) *Shannon Information* (henceforth 'SI', [3]). The aim and the choice of the two information features are motivated by Shannon's *source coding theorem* [3] on the interaction of information, coding and length of signs within binary alphabets. We formulate the following research question: can Shannon's theorem be transferred to natural languages and does coding of aspect marked verbs interact with the information that they carry? As classifier for the binary classification task that is, the classification of aspect-marked verbs into default- and non-default classes, we employed a Support Vector Machine (henceforth SVM, [4]). The choice of the test set of languages is motivated by the overt marking of aspect on

* Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number: 357550571.

† This work 'Predicting default and non-default aspectual coding: Impact and density of information features ' is the extended version of an abstract under the same title presented at KONVENS 2019 and at the time of the release of the proceedings of NL4AI, published at *Preliminary Proceedings of the 15th conference of Natural Language Processing (KONVENS 2019): Kaleidoscope Abstracts*, 275 – 277 (2019) by Michael Richter and Tariq Yousef. (https://creativecommons.org/licenses/by-nc-sa/4.0/)

verbs in these languages. As data resource we exploited Universal Dependency Treebanks in CoNNL-U format (https://universaldependencies.org) because verbal aspect is encoded in these corpora, as exemplified for the Latvian verb *pierādīt* 'prove' in figure 1. The token *pierādījuši* 'proven' carries perfective aspect:



```
pierādījuši    pierādīt    VERB   [...] Aspect=Perf [...]
```

Fig. 1. Corpus entry of the Latvian verb *pierādīt* 'prove' with aspect information.

What does default and non-default coding mean? Our point of departure is that verbs have a dominant aspect category and that this category can be determined by frequency distributions: default forms will occur more frequently than non-default forms. Take as an example the Polish verb *spotkać* 'meet'. This verb form has the default aspect 'perfective' while the verb form *spotykać* carries imperfective aspect and is thus non-default coded. The *Form Frequency Correspondence Principle* (henceforth FFC, [5]) is based on this default /non-default-dichotomy. FFC says that default-coded words (in general) tend to be shorter than non-default-coded words and - according to Zipf's *principle of least effort* [6] – longer words carry more information than shorter ones (otherwise the greater length, that is, the higher effort, would be uneconomic).

The second aim of the study is to test whether the *Uniform Information Density* – hypothesis (henceforth UIDh [7], [8], [9], [10]) holds within the features IC and SI of the target verbs. This is a novel interpretation of UIDh. The hypothesis says that the amount of information within messages should cross linguistically be uniform and there should neither be extreme peaks nor extreme troughs in the stream of information in order to facilitate language processing and comprehension. Our research question is: Are there extreme information peaks and troughs within a single linguistic unit which might make the procession of that unit difficult?

According to UIDh, the variances in information density in the languages in the focus of this study should not be far apart. In its original form, UIDh is applied to discrete signs carrying individual information. We, however, apply UIDh to two different information values of a *single* sign. UIDh is formulated within the framework of *Surprisal theory*: the difficulty of processing signs of natural language is proportional to its informativity in context [11] and signs must not be too informative in order to be processable. [12] states that surprisal is a measure of reranking cost: facing an unexpected word in the sentence, a (human) sentence processor has revise his or her incremental expectations that is, a "shift in the resource allocation (equivalently, in the conditional probability distribution over interpretations)" is required [12]. In this study we test the prediction whether SI and IC have a uniform information density (UID) that is, the information values should not have high variances [13] and tend towards zero.

2 Related work

Although the interaction of IC and coding has, to our best knowledge, has not yet been studied for natural language, the interaction of IC and length of words is the topic in a couple of studies. [1] brought to light that IC is a strong predictor of phone deletion in English. [2] showed for ten Indo-European languages that IC, estimated from bigram-, trigram-, and 4-gram-contexts of the target words, is a better predictor of word length than frequency. [2] ascribe the attested correlation of word length and information content to the principle of UID: the amount of information over time must be constant, and it follows that longer word forms must be more informative than short ones.

[14] investigated for Arabic, Chinese, English, Finnish, German, Hindi, Persian, Russian and Spanish, whether the length of words can be better predicted by IC, when it is estimated from syntactic dependents rather than from unstructured contexts of target words. Her finding was that words that convey more IC to their contexts tend to be longer. The study of [15] yielded a controversial result: for 30 languages in focus, the lengths of aspect-coded verbs could be better predicted by unigrams than by syntactic contexts.

The validity of UIDh has been tested so far only for distinct linguistic units: [12] and [16] found – in order to test UIDh - a positive correlation between surprisal and difficulty of signs, which was operationalized by measuring reading times: surprising words in sentences need more time to be read. [9] showed in their study on omission of the relative pronoun in English relative clauses (RC), that if *that* is expected and thus low informative, it tends to be omitted. However, in cases of unexpected and high informative RC, *that* is not omitted: The use of the relativiser signals to the human processor that a relative sentence follows, and thus reduces the amount of surprisal and information. Using the example of article omission in German, [17] demonstrated, that UID depends on whether information is determined by terminal symbols or by POS tags and that POS tags provide a better basis for explaining article-omission.

3 Method

3.1 Data

Data resources are the corpora ‘bg_btb-ud-train.csv’ (Bulgarian), 'cu-ud-train.csv’ (Old Church Slavonic), ‘pl_lfg-ud-train.csv’ (Polish), ‘sk_snk-ud-train.csv’ (Slovak), ‘sl_ssj-ud-train.csv’ (Slovenian), ‘uk_iu-ud-train.csv’ (Ukrainian) and ‘lv_lvtb-ud-train.csv’ (Latvian), from the Universal Dependency Treebank, version 2.3 (https://universaldependencies.org). All aspect-marked verbs were extracted. The number of the resulting verb forms for each language is displayed in table 1.

Table 1. The number of verb forms in the test set of languages.

language	number of verb forms
Bulgarian	13,714
Latvian	17,046
Old Church Slavonic	9,575
Polish	17,199
Slovak	11,749
Slovenian	11,629

Ukrainian	9,789

3.2 Classifier and features

We employed a Support Vector Machine binary classifier with a radial basis function kernel [4] which utilises as features IC and SI . The aim was to classify the data (aspect marked verbs) into two categories, default (0) and non-default (1). We used 80% of the data set to train the model, and the rest to assess the quality of the classifier. The estimation of IC is given in (1), it is the average amount of information, that a verb form conveys within all of its contexts. :

$$IC = E(-log_2(P(W = w \,|C = c_i))) \qquad (1)$$

IC is the expectation value of the negative log of conditional probability of a verb form *w* (marked with imperfective or with perfective aspect) given contexts *C*. As contexts, we took bigrams, i. e. *lexical surprisal* ([11], [16]), to both directions of the target verbs since a study of [15] disclosed that target verbs convey the highest amount of information in this context window. In (2), the estimation of SI is given [3]. SI is the information of each individual verb form *w* in its contexts:

$$SI = -log_2\big(P(W = w|\ context)\big) \qquad (2)$$

3.3 Default and non-default forms

For each verb, the default and non-default aspect was determined. We reduced aspect oppositions to the binary imperfective-perfective distinction and subsumed the habitual and progressive aspects under the imperfective and the resultative aspect under the perfective aspect, respectively. Verb forms in the prospective aspect have been ignored, since its value is not clear with respect to the imperfective and perfective opposition. We checked for every verb the number of occurrences in perfective and imperfective aspect, and took the difference of both occurrences. The more frequent aspect forms were taken as default aspect of the respective verb lemma.

The differences were normalized, and ten thresholds between [.09:1] were set as differences between default and non-default. The threshold '1' was omitted a priori, since it captures cases of verbs occurring only in one aspect form that is, either solely perfective or solely imperfective aspect.

4 Results

We focused on the thresholds in the interval [.19, .59] on the normalised threshold-scale, in order to ensure a sufficient number of default and nondefault encodings for the training of the SVM-classifier. The thresholds of the interval [.59, .99] provided a too small number of non-default aspect coded verb forms. At the lowest threshold value, i.e. .19, the frequencies of default and non-default coded verbs differ only slightly and both groups are almost equally distributed. In table 2, the range of accuracy values within the interval [.19, .59] for the seven languages in focus are given (left accuracy values for threshold .19, the right values for threshold .59):

Table 2. Range of classification accuracy for the seven languages in our study.

language	accuracy (%)
Bulgarian	99.5 – 99.8
Old Church Slavonic	94.3 – 97.8
Polish	99.7 – 99.9
Slovak	99. 5 – 99.6
Slovenian	100 – 100
Ukrainian	99.1 – 100
Latvian	98.3 – 99.5

It comes to light that the accuracy is almost independent of the threshold and thus of the frequency distribution: even with an almost equal distribution of default and non-default aspect frequencies that is, with threshold .19, almost perfect accuracy values are achieved. In order to estimate UID, we used (3). More precisely, we utilised *global information density* UID_{GLOBAL} which is the variance within information values [13]: id_i is the information density of SI and IC of a single verb form, and μ is the mean of *id*:

$$UID_{GLOBAL} = -E(\sum_{i=1}^{N}(id_i - \mu))^2 \qquad (3)$$

Applying (3) to our test set of languages, an identical pattern in all languages comes to light: the variance of information within IC and SI is small and the majority of variance values tends to be close to zero (note that UID_{GLOBAL} values are negative by definition). As illustration, UID_{GLOBAL} of Polish, Slovenian and Latvian are given in figure 2:

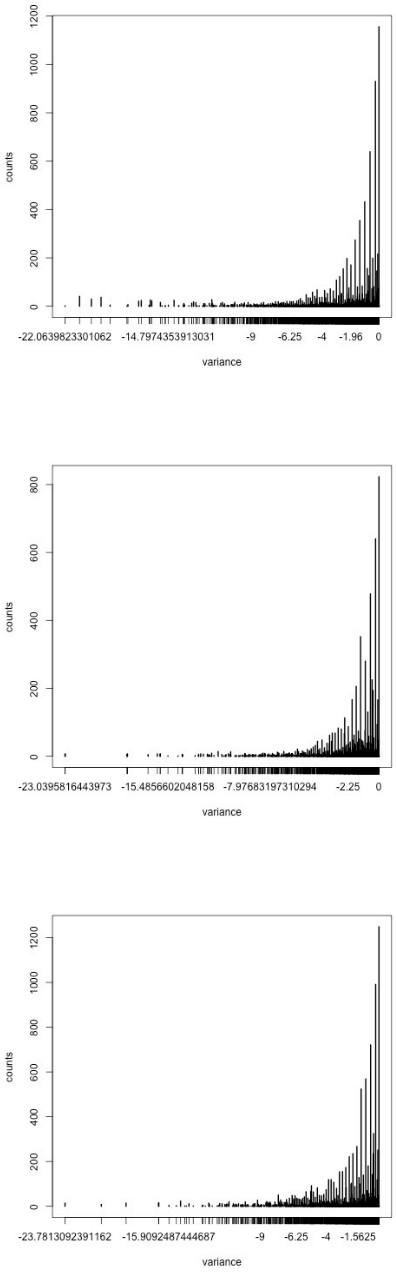


Fig. 2. UID$_{glob}$ in Polish, Slovenian and Latvian.

5 Conclusion

A classification with high accuracy of default / non-default coding of verbs could be achieved with a SVM classifier and the features SI and IC. As Shannon's source coding theorem predicts, we found interaction of aspectual coding and information: Our study provides evidence that non-default coded verb forms are more informative than default forms. Almost identical accuracy has been achieved with all tested threshold values, and we take this finding as an indication of a – in the average – constant amount of information of IC and SI.

With regard to the second aim, our study disclosed that UIDh holds within the features IC and SI. The variation within the two features tends to be close to zero in all languages in our test set and our prediction turns out to be correct: both features convey an uniform stream of information throughout the forms of the seven languages in focus. This ensures that information does not become, in the words of [9], "dangerously high". The question arises whether UID can be consciously regulated in SI and IC, i.e. whether it is a conscious linguistic behavior. If, for example, a speaker plans to use an unexpected and therefore informative word form, he or she could at the same time decide to use that form in expected contexts which cause not much surprisal. Whether regulation of SI and IC is a conscious linguistic behavior is a question that requires future work in the form of psycholinguistic experiments. A practical application of this study is POS-tagging in languages with fuzzy distinction between word classes such as Tagalog. This is based on our hypothesis that default / non-default-coding correlates with word classes for instance with the noun / verb-distinction. According to this hypothesis, the word class of default form of a lemma could differ from the word class of a non-default form.

References

1. Cohen Priva, U.: Using information content to predict phone deletion. In: Proceedings of the 27th West Coast Conference on Formal Linguistics, pp. 90 – 98 (2008).
2. Piantadosi, S. T., Tily, H., Gibson, E: Word lengths are optimized for efficient communication. PNAS, 108(9), 3526 – 3529 (2011).
3. Shannon, C. E., Weaver, W.: A mathematical theory of communication. The Bell System Technical Journal 27 (1948).
4. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features (1998).
   Retrieved from http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf.
5. Haspelmath, M, Calude, A., Spagnol, M., Narrog, H., Bamyaci, E.: Coding causal noncausal verb alternations: A form–frequency correspondence explanation. Journal of Linguistics, 50(3), 587 – 625 (2014).
6. Zipf, G. K.: Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press (1949).
7. Genzel, D., Charniak, E.: Entropy rate constancy in text. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp.199 – 206 (2002).

8. Aylett , M., Turk, A.: The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech, 47(1), 31 – 56 (2004).
9. Levy, R., Jaeger, T. F.: Speakers optimize information density through syntactic reduction. In: Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS) (2007).
10. Jaeger, T. F.: Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61 (1), 23 – 62 (2010).
11. Hale, J.: A probabilistic Earley parser as a psycholinguistic model. In: Proceedings of NAACL, pp. 1 – 8 (2001).
12. Levy, R.: Memory and Surprisal in Human Sentence Comprehension. In: van Gompel, R. (ed.) Sentence Processing, pp. 78 – 114. Psychology Press, Hove (2013).
13. Collins, M. X.: Information density and dependency length as complementary cognitive models. Journal of Psycholinguistic Research, 43(5), 651 – 681 (2014).
14. Levchina, N.: 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) (2017).
15. Richter, M., Kyogoku. Y., Kölbl. M.: Interaction of Information Content and Frequency as predictors of verbs' lengths. In: Abramowicz , W., Corchuelo, R. (eds.) Business Information System. 22nd International Conference, BIS 2019, Seville, Spain, June 26–28, 2019, Proceedings, Part I (Lecture Notes in Business Information Processing 353), pp. 271 – 282. Springer (2019).
16. Levy. R.: Expectation-based syntactic comprehension. Cognition, 106: 1126–1177 (2008).
17. Horch, E., Reich, I.: 2016. On “Article Omission” in German and the “Uniform Information Density Hypothesis”. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), pp.125 – 127 (2016).

.