# What can linguists learn from some simple statistics on annotated treebanks

Jiří Mírovský and Eva Hajičová

Charles University in Prague
Faculty of Mathematics and Physics
E-mail: mirovsky,hajicova@ufal.mff.cuni.cz

**Abstract**

The goal of the present contribution is rather modest: to collect simple statistics carried out on different layers of the annotation scenario of the Prague Dependency Treebank (PDT; [1]) in order to illustrate their usefulness for linguistic research, either by supporting existing hypotheses or suggesting new research questions or new explanations of the existing ones. For this purpose, we have collected the data from the already published papers on PDT (quoted at the relevant places), adding some more recent results and drawing some more general consequences relevant for Czech grammar writers.

## 1 Frequency of occurrences of particular phenomena

### 1.1 Non-projectivity of word order

Projectivity of dependency trees representing the syntactic structure of sentences has been and still is a frequently discussed property of the trees as this property offers a possible restriction on syntactic representations. It is well known that word order in Czech is not in principle guided by grammatical rules, so that it might be expected that the instances of non-projectivities in Czech might not be frequent. A detailed analysis of non-projective constructions in Czech is given in [13]. His statistical data are based on the PDT analytical (surface structure) level comprising 73,088 non-empty sentences and 1,255,590 words (incl. punctuation marks). There are 16,920 sentences (23.2%) in the collection that contain at least one non-projectivity (i.e. including at least one node in a non-projective position). However, from the point of view of the total number of nodes in the analyzed collection, there were only 23,691 (1.9%) nodes hanging in a non-projective way. As the PDT annotation is carried out both at the surface syntactic as well as at the underlying syntactic level, it was possible to compare the two levels. The statistical findings indicate that 71.47% of non-projectivities stem from special properties of the surface syntactic level: function words separated from the lexical words they

are associated with and analytic verb forms (50.54%), split constructions such as phrasemes and noun groups (2.46%), placement of particles "outside" the sentence (17%), grammatical restrictions on surface word order (1.47%). It seems then plausible to work with the assumption that the underlying, tectogrammatical level can be characterized as projective. Moreover, the statistical data have indicated that the main cause of non-projectivities is the information structure of the sentence (e.g. in the case of split noun groups). Even here more detailed classification of the statistical data give us some guidance (see [5]).

## 1.2 Information structure annotation of the Czech corpus (TFA)

In the theoretical account of topic-focus articulation (TFA) within the framework of the Functional Generative Description, the dichotomy of topic (what is the sentence about) and focus (what it says about the topic) is understood as based on the primary notion of contextual boundness. Every node of the tectogrammatical dependency tree carries an index of contextual boundness: a node can be either contextually bound (*t*, or, in case of contrast, *c*) or non-bound (*f*). For the identification of the dichotomy of topic and focus on the basis of contextual boundness, a rather strong hypothesis was formulated, namely that the topic-focus distinction can be made depending on the status of the main verb (i.e. the root) of the sentence and its immediate dependents.

To test this hypothesis, an implementation of the algorithm was applied to the whole PDT data. The results reported in detail in [4] can be summarized as follows: focus consisting of a contextually non-bound verb and its contextually non-bound subtrees occurred in 85.7%; focus consisting only of the contextually non-bound elements depending on the contextually bound verb together with the subtrees depending on them: 8.58%. There occurred about 4.47% of special cases and an ambiguous partition was found in 1.14% of cases. No focus was identified in 0.11% of cases.

The results indicate that a clear division of the sentence into topic and focus according to the hypothesized rules has been achieved in 94.28% of sentences to which the procedure has been applied; the real problem of the algorithm then rests with the case of ambiguous partition (1.14%) and cases where no focus was recognized (0.11%). The results achieved by the automatic procedure were then compared to the judgements of Czech speakers ([14]). The annotators were instructed to mark – according to their intuition – every single word in the sentence as belonging to topic or focus and, at the same time, they were supposed to mark which part of the sentence they understand as topic and which part as focus. It is interesting to note that the annotators' agreement in the assignments of individual words in the sentences to topic or to focus was much higher (about 75% in both the three and six parallel analyses compared to 36% of the assignments of the topic and focus as a whole) than the assignments of the topic-focus boundary.

The work on this step is still in progress. It is a matter of course that the variability of manual solutions must be taken into considerations; we are aware of

the fact that while we get only a single, unambiguous result from the automatic procedure, more ways of interpretation could be correct.

The empirical study of Czech texts has led to the assumption that the ordering of the elements in the focus part of the sentence is primarily given by the type of the complementation of the verb. A hypothesis called systemic ordering of the elements in the focus of the sentence was formulated and empirically tested pairwise (i.e. successively for two of the complementation types) and supported also by several psycholinguistic experiments. Though the hypothesis was based on the examination of hundreds of examples, the material of the PDT offers a far richer material. The statistical findings support the following assumptions: (a) the sentential character of a complementation is a very important factor in that there is a tendency of a contextually non-bound element expressed by a clause to follow the non-sentential element, (b) the influence of the form of the complementation: e.g. the assumed order Manner – Patient is more frequent if the complementation of Manner is expressed by an adverb and the complementation of Patient by a nominal group; also the outer form of the Actor plays an important role: if the Actor is expressed by infinitive, Patient precedes Actor, while the hypothesized order Actor – Patient is attested if both complementations are expressed by nominal groups; (c) with some pairs, such as Patient and Means, there was a balance between the frequency of the two possible orders, which may indicate that for some particular complementations more than a single complementation occupy one position on the scale ([10]).

In some cases the decisions of the annotators are not the only possible ones and this fact has to be taken into consideration when drawing conclusions. This observation is confirmed also by the data on annotators' agreement/disagreement, see also [12].

## 1.3 Annotation of discourse relations

The discourse annotation in PDT 3.0 was based on a narrowly specified category of language expressions commonly known as connectives. However, it soon has become clear that such an annotation would miss some important discourse relations that are expressed by other means. The importance of this broader view is supported by the comparison of the number of relations expressed by connectives and those expressed by some alternative way (called AltLexes):

| | all | intra-sentential | inter-sentential |
|---|---|---|---|
| AltLex: | 726 | 272 (2.1%) | 454 (7.7%) |
| connective: | 17,983 | 12,523 (97.9%) | 5,460 (92.3%) |
| total: | 18,709 | 12,795 (100%) | 5,914 (100%) |

The numbers indicate that AltLexes express mostly inter-sentential discourse relations. Among them, they form almost 8% of all explicitly expressed relations, which makes them an indispensable part of the analysis of discourse (see [11]).

The largest proportion of occurrences within a single (complex) sentence is documented for the relations of purpose (100%), condition (99%), and disjunctive alternative (95%). These relations only rarely occur between two independent sentences (0, 1, 5%, respectively). On the basis of these observations, a preliminary hypothesis can be formulated that the semantic content expressed by the arguments of the above relations are more closely bound together than with the other relations. Also the relatively high position of conjunction (81%) is surprising as one would expect a more balanced distribution, perhaps similar to that found with opposition (43%).

The measuring of the ratio between the number of sentences and the number of discourse relations in individual genres has led to the observation ([8]) that in the PDT journalistic data, explicit connectives are most frequently used in genres with a high degree of subjectivity, i.e. where opinions, desires, evaluations, beliefs etc. are expressed. With the exception of sport, the first eight positions are represented by genres in which a certain degree of subjectivity often plays an important role, while the "objective" genres gathered consistently lower in the connective frequency scale. On the other hand, program or captions are typical in containing only a minimum of connectives since they are either very short (captions) or they are often represented by verbless phrases only (both genres).

## 2 Annotators' agreement

One of the interesting issues that can be observed when following the data on annotators' agreement as categorized according to the linguistic levels of description is the increasing number of disagreements if one proceeds from the POS or morphological level (which is the closest one to the outer linguistic form) to the level of underlying syntax and discourse.

*Morphology:* Agreement in PDT on choosing the correct morphological tag (5 thousand different tags): 97% ([3]). For German – in Negra (54 tags): 98.57% ([2]).

*Surface syntax:* No numbers for PDT; in Negra: (F-measure) for the unlabelled structural annotation: 92.43%, and for the labelled structural annotation (labelled nodes with 25 phrase types and labelled edges with 45 grammatical functions): 88.53% ([2]).

*Deep syntax (tectogrammatics):* In PDT, the agreement on establishing the correct dependency between pairs of nodes was 91%. The agreement on assigning the correct type to the dependency relation (67 possible values of the tectogrammatical functor) was 84% ([6]).

*Topic-focus articulation:* The agreement on assigning the correct value to individual nodes in the annotation of contextual boundness (i.e. the assignment of the values 'contextually bound' or 'contextually non-bound') was 82% ([12]).

*Discourse phenomena:* The agreement on the recognition of a discourse relation (connective-based F1-measure) was 83%. The agreement on the recognition

of a textual coreference or a bridging anaphora (chain-based F1-measure) was 72% and 46%, respectively. The agreement on the type of the relations in cases where the annotators recognized the same relation (a simple ratio) was 77% (Cohen's $\kappa$ 71%) for discourse, 90% (Cohen's $\kappa$ 73%) for textual coreference, and 92% (Cohen's $\kappa$ 89%) for bridging anaphora ([9]). Sometimes even a small amount of annotated data can reveal important facts. In a small probe of annotating implicit discourse relations, the task proved to be highly challenging – the annotator's agreement on setting the type of implicit discourse relation between adjacent sentences was less than 60%.

The numbers of agreement for the different tasks cannot be directly compared (as they measure different phenomena, use different methods of evaluation and sometimes annotate different (type of) data), however, they seem to support the hypothesis that the deeper we go in the abstraction of the language description, the more difficult it is to achieve high values of the inter-annotator agreement. The above data also support the view (doubted by some linguists in the past) that it is easier to assign the structure (in other terms, the relation of dependency: the status of the governor and that of the dependent) than the value (type) of the dependency relations. This observation is also supported by the data on the Prague Czech-English Dependency Treebank (PCEDT) where the agreement on establishing the correct dependency between pairs of nodes was 88% while the agreement on assigning the correct type to the dependency relation was 85.5% ([7]).

## 3 Conclusion

We have collected some observations related to different layers of corpus annotation to demonstrate that even simple frequency data may give a linguist an important guidance for his/her deeper analysis of different linguistic phenomena. The prescribed length of the paper has allowed us just to summarize these observations; a more detailed statistics as well as analysis of the data can be found in the papers referred to.

## Acknowledgements

We gratefully acknowledge support from the Grant Agency of the Czech Republic (project n. P406/12/0658). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2010013).

## References

[1] Bejček, E., Hajičová, E., Hajič, J. et al. (2013) *Prague Dependency Treebank 3.0*. Data/software, Charles University in Prague, Czech Republic.

[2] Brants, T. (2000) Inter-Annotator Agreement for a German Newspaper Corpus. In: *Proceedings of the Second LREC*, Athens, Greece.

[3] Hajič, J. (2005) Complex corpus annotation: The Prague dependency treebank. In *Insight into the Slovak and Czech Corpus Linguistics 2005*, 54.

[4] Hajičová, E., Havelka, J., Veselá, K. (2005) Corpus Evidence of Contextual Boundness and Focus. In: *Proceedings of the Corpus Linguistics Conference Series*, U. of Birmingham, pp. 1–9.

[5] Hajičová, E., Havelka, J., Sgall, P. et al. (2004) Issues of Projectivity in the Prague Dependency Treebank. In *The Prague Bulletin of Mathematical Linguistics, 81*, Charles University in Prague, pp. 5–22.

[6] Hajičová, E., Pajas, P., Veselá, K. (2002) Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations. In *The Prague Bulletin of Mathematical Linguistics, 77*, Charles University in Prague, pp. 5–18.

[7] Mikulová, M., Štěpánek, J. (2010) Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank. In: *Proceedings of the 7th LREC*, Valletta, Malta, pp. 1836–1839.

[8] Poláková, L., Jínová, P., Mírovský, J. (2014) Genres in the Prague Discourse Treebank. In: *Proceedings of the 9th LREC*, Reykjavík, Iceland, pp. 1320–1326.

[9] Poláková, L., Mírovský, J., Nedoluzhko, A. et al. (2013) Introducing the Prague Discourse Treebank 1.0. In: *Proceedings of the 6th IJCNLP*, Nagoya, Japan, pp. 91–99.

[10] Rysová, K. (2013) *On Word Order from the Communicative Point of View*. PhD Thesis at Faculty of Arts, Charles University in Prague.

[11] Rysová, M. (2012) Alternative Lexicalizations of Discourse Connectives in Czech. In: *Proceedings of the 8th LREC*, İstanbul, Turkey, pp. 2800–2807.

[12] Veselá, K., Havelka, J., Hajičová, E. (2004) Annotators' Agreement: The Case of Topic-Focus Articulation. In: *Proceedings of the 4th LREC*, Lisboa, Portugal, pp. 2191–2194.

[13] Zeman, D. (2004) *Parsing with a Statistical Dependency Model*. PhD thesis, Univerzita Karlova v Praze, Praha.

[14] Zikánová, Š., Týnovský, M., Havelka, J. (2007) Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. In: *The Prague Bulletin of Mathematical Linguistics, 87*, Charles University in Prague, pp. 61–70.