Signals of Attribution in the Prague Dependency Treebank

Lucie Poláková, Pavlína Jínová and Jiří Mírovský

Faculty of Mathematics and Physics
Charles University in Prague
E-mail: {polakova|jinova|mirovsky}@ufal.mff.cuni.cz

Abstract

The paper aims at mining a richly annotated treebank for features relevant in automatic annotation/detection of attribution – ascription of text contents to agents who expressed them. We find three such features, implement an automatic procedure to detect attribution relations in our data and evaluate its results.

1 Introduction

In discourse-oriented linguistic research, attribution, or the ascription of text contents to the agents (sources) who expressed them, has become an important component of analysis, e.g. in the Penn Discourse Treebank [8], or it even developed to independent annotation projects, cf. Pareti [6].

Attribution relations (ARs) can be signaled with a range of language means. Mostly, it is clauses containing verbs of saying and thinking, but also further, non-verbal attribution phrases, compare Example 1 with two contents attributed to somebody else than the author. The example contains a prepositional signal *according to Kalina* and a clausal (verbal) signal *he remarks*.[1]

(1) A special category is the bank's award for the best Czech recording. **According to Kalina**, *this is an insurance for the case that the domestic production fails in all other categories.* However, *that did not happen*, **he remarks**.

In the Prague Dependency Treebank (PDT, Czech journalistic texts, [2]), annotation of discourse relations was first introduced in 2013 [7] but with no annotation of attribution so far. Before that, a complex manual analysis on three levels of description (morphology, surface and underlying syntax = tectogrammatics, [5])

[1] Attributed contents in Example 1 are highlighted in italics, attribution cues in bold.

had been carried out. Some of these annotated features appeared to be of great advantage for annotating intra-sentential discourse information.

The aim of this paper is twofold: i) to detect which attributes from the rich PDT annotation (or their combination) capture signals of attribution relations, and ii) to evaluate the reliability of these signals for an automatic annotation of this phenomenon. This is quite a natural next step towards a complex description of discourse relations. We are aware that the given task is partly dependent only on semantic and pragmatic features and as such cannot be fully automatized, our goal is therefore rather in determining how far we can facilitate the task by relying on already available information.

2 Preparatory Analysis

2.1 Method

Similarly as Pareti [6], we recognize an attribution relation as consisting of three main elements: the source of the attribution (agent expressing the contents), the attributed content and the cue – typically an attribution verb, less frequently also prepositions, adverbs, punctuation marks etc.

The research in this paper is targeted for future assignment of attribution primarily to discourse arguments and relations, thus it addresses mainly the possibilities of the identification of the cue.[2] Also, verbal cues that only introduce a sentence constituent, as in *He announced the break of contacts with the rebels.* are not targeted here, as non-clausal sentence constituents alone are not annotated as discourse arguments in the PDT so far.

To obtain a view of possible signals of attribution in the PDT, a manual inspection or random data samples was conducted, resulting in a list of signals which was afterwards further analyzed. Basically, morphological, syntactic and lexical features were encountered besides features connected with the text structure. Five attributes of the tectogrammatical layer seemed to represent the core of attribution signals; for each of them, 50 random occurrences in the corpus were examined to estimate their reliability for an automatic annotation. Three most promising attributes from these five are described in detail in Section 2.2 below. Other signals, assessed as either less distinctive or too rare for our purposes, are not further addressed in this paper.

2.2 Tracked Signals: Reported Speech and Verbs of Saying

Reported speech in the tectogrammatical representation is marked with the attribute *is_dsp_root* – **the root of a direct speech**. The goal of introducing this attribute was originally to mark syntactically unanchored reported contents (i.e. a reported

[2] So far, it does not concern the identification of sources, and only partly investigates the attributed contents, although the analyzed attributes in the PDT mostly also directly point to these two attribution elements.

speech not representing an obligatory modification of a governing verb of saying). The attribute is nevertheless assigned also to syntactically incorporated reported contents, both those graphically signaled by quotation marks and those without them.

However, in some cases, the *is_dsp_root* attribute is marked inconsistently, as it was not the main focus of the tectogrammatical annotation. That is where **valency frames** (syntactico-semantic roles) of the verbs can significantly help. In the valency lexicon of Czech verbs Vallex ([3], [4]), whose electronic version can be linked to the verbs in the corpus (see below in 3.1), each verb belongs to a certain semantic class (like motion, perception, change). For our experiment, we selected the semantic class of communication, as it intersects the best with verbs of saying. Verbs of thinking (the class of mental action in Vallex) were left out from the experiment in this phase. The list of verbs of communication (in Vallex 2.7) comprises 431 verbal frames,[3] 391 of which are relevant for the analysis of attribution. The combination of a unique verbal frame ID and a desired valency frame constellation is a promising way to detect both attribution cues and contents. Also, in this way, irrelevant meanings of polysemous verbs can be sorted out, as they have different valency frames.

Syntactically unanchored reported speech appears in Czech typically in cases where an introductory verb does not open a valency position for the content of saying (no direct object possible) or the position of a direct object is taken by another expression, cf. the expression *utkání* [*match*] in Example 2. Such structures in the PDT annotation are interpreted as if a verb of saying was missing. It is therefore represented by **a newly established node with the *t-lemma* substitute *#EmpVerb*** (empty verb) in the position of a non-obligatory verb complement [5, p. 421ff].[4] In Example 2, the whole reported content *I managed to win important rallies, Hyo arranged for the mistakes* is rooted in a generated *#EmpVerb* node representing approximately the (missing) verb *saying*. At the same time, this empty verb node is in the position of verbal complement (the COMPL functor) with dual dependency both on the verb *zhodnotila* [*evaluated*] and the noun *Novotná*.

(2) *Dařilo se mi vyhrávat důležité výměny, o chyby se postarala Hyová, zhodnotila ani ne hodinové utkání Novotná.*

[*I managed to win important rallies, Hyo arranged for the mistakes, Novotná evaluated the not even one hour lasting match.*]

It can be considered a reliable signal for attribution, with the added value of directly pointing at the source – it is always the entity in the position of the secondary parent of the complement (*Novotná* in Example 2).

[3] one verb lemma can have several different frames

[4] referred to as #EmpVerb.COMPL in Table 1 below

3 Automatic Detection of Attribution

3.1 Experiment Setting

For testing the theoretical analysis from the previous section, we have implemented an automatic procedure for detection of ARs in the PDT data. The selected features are, again:

- is_dsp_root – reported speech signaled by a dedicated attribute
- Vallex – usage of one of selected verbs of saying, extracted from the semantic class of communication from the valency lexicon Vallex
- #EmpVerb.COMPL – syntactically unanchored direct speech represented by a generated empty verb node in the position of verbal complement

To detect verbs of saying, we used the annotation of semantic classes in the valency lexicon Vallex, as described above in Section 2.2. However, information from Vallex about verb frames and their membership in the semantic class of communication could not be used directly. Verbs in the PDT data are not linked to Vallex but instead to so-called PDT-Vallex, where there is no annotation of semantic classes. Unfortunately, these two lexicons are not compatible in a straightforward way. For transforming the information about semantic classes from Vallex to PDT-Vallex, we used an automatic alignment of these two lexicons created by Bejček [1].

The automatic procedure for the attribution detection was tested on a selection of 15 manually evaluated documents from the PDT, comprising in total of 563 sentences. In an attempt to avoid documents with contents attributed only to the author of the text, the documents were selected based on different proportions of occurrences of the attribute *is_dsp_root*, three documents did not contain any occurrence of this attribute at all.

3.2 Results

Table 1 shows numbers of hits of individual or combined features of the automatic procedure in the manually evaluated data and, for comparison, also in 9/10 of the whole PDT data. A “hit” means a position in the data where the procedure detected one or more signals of attribution, that means, where it found at least one signal that the text span is attributed to some other source than the author. Sentences attributed only to the author of the text were ignored in the manual evaluation, or, in other words, a zero hit of the procedure in such a sentence did not count as a positive result. If there were several signals of attribution for the same text span (typically a clause), we count it as one hit in the respective row of the table. It means that, for example, in the manually evaluated data *is_dsp_root* was detected 68 times as the only signal of attribution and 48 times together with a verb of communication.

In the manually evaluated data, the automatic procedure correctly identified 137 out of 182 attribution relations, and incorrectly marked 3 relations. This means that the precision was 98%, recall 75%, and F1-measure 85%.

Feature(s)	In manual evaluation	In 9/10 of the PDT
is_dsp_root	68	1,693
Vallex + is_dsp_root	48	1,022
Vallex	10	1,324
#EmpVerb.COMPL	7	84
#EmpVerb.COMPL + is_dsp_root	5	71
Vallex + #EmpVerb.COMPL + is_dsp_root	1	16
Vallex + #EmpVerb.COMPL	1	10
total number of hits	140	4,220
total number of sentences	563	43,955

Table 1: Numbers of hits of individual features in the manually checked data and in 9/10 of the whole PDT data.

The high precision of the automatic procedure is an encouraging result and, considering that we have at this moment implemented only three signals of ARs, we consider the recall and the F1-measure figures also quite satisfactory.

3.3 Analysis of the Results

As Table 1 shows, the *is_dsp_root* attribute is the most reliable signal for identification of the reported contents among the implemented attributes. It correctly identified, as a single signal or in combination, 122 out of 182 ARs present in our data. This attribute moreover precisely delimits the reported content (the t-node with this attribute and its subtree) and points to the cue (if any present). Using valency frames from Vallex is more complicated due to its potential false positivity (see below). We were able to correctly detect 57 cue verbs in 182 ARs, however, it should be noted that not all ARs have a verbal cue. The effectiveness of this feature could be increased by finer rules regarding the individual frames. *#EmpVerb.COMPL* is a very precise signal of ARs, but, at the same time, it is quite rare. There are only 181 occurrences of these structures in the 9/10 of the PDT data. But, this signal is linguistically interesting in one respect – it can show which verbs outside the core of *verba dicendi* also can introduce attributed contents. We came across Czech verbs roughly corresponding to English *to join in, to conclude, to repeat, to give up, to praise, to react, to be delighted* and so on.

From the 45 undetected attribution relations, more than a half (25) were cases of a reported speech without any introductory verb. Such sentences mostly appear in a longer sequence of uninterrupted direct speech. The verb of saying is usually used only once for such a sequence. In 19 of these cases, attributing the content to somebody else than the author would be nevertheless possible by tracking the use of first person singular or plural (which is typical in our data – mostly news interviews). The remaining 6 cases could be identified as reported speech only thanks to thematic progressions and semantics.[5] Further, 9 undetected ARs were marked lexically with *podle* [*according to*] phrase, *prý* [*reportedly*], and *údajně* [*allegedly*]. In 4 cases, the procedure did not identify a verb of saying because its valency frame did not match any frame in our Vallex-originated list. In the remaining cases, the content of saying was expressed only through a demonstrative pronoun, and so the verbal cue and the content appeared in different sentences. Finally, the procedure so far failed to recognize parenthetical attributive structures with reverse syntactic order of the type *as he claims*.

There were three false positive hits in the manually evaluated texts. Although this is a small number, the individual cases point at two systematic problems of the procedure. First, it is the identified verbs of saying uttered by the author himself about himself, including certain fixed connections like *lépe řečeno* [*or rather,* lit. *better said*]. Second, it is some non-speaking meanings of some polysemous verbs. Most of the irrelevant frames were sorted out by the semantic class in Vallex, but some can remain, cf. the meaning of the verb *potvrdily* [*confirmed*] in Example 3.

(3) *Vítkovice potvrdily výhrou 2:0 nad Uherským Brodem, že budou patřit k nejvážnějším kandidátům na postup.*

[*Vítkovice confirmed by winning 2:0 over Uherský Brod that it will belong to the most serious candidates for the advance.*]

4 Discussion and Conclusions

Despite the complexity of detecting ARs in a text, we believe to have shown with our experiment that this task can be significantly facilitated if reliable syntactic annotation is at one's disposal. A crucial role also plays an available electronic lexicon of verbs with their syntactico-semantic roles (valency lexicon). Being that far, only implementation of three strongest features suffices to achieve very high precision and a fair recall. The procedure can be easily enhanced by adding further, rather primitive features like switching the category of person, lexical cues (according to + proper names, allegedly) etc. The proposed procedure is useful for any Czech treebank with tectogrammatical analysis (with a necessary decrease in performance in case of solely automatic parsing). On the other hand, the use of the valency lexicon makes it language-dependent. Also, for the time being, the analysis and the automatic procedure does not concern verbs of thinking that are, in our opinion, even trickier in expressing attribution relations than verbs of saying. We plan to address this issue in future experiments. For our research, which focuses on assigning attribution to already annotated discourse relations and arguments, the proposed experiment is a promising start. Manual evaluation of the results revealed very well the nature of cases where the procedure fails, which is a valuable linguistic feedback for understanding attribution and its principles.

[5] There were also two cases in our sample data where it could not be decided at all to whom they should be attributed. These cases were excluded from the evaluation.

Acknowledgements

We gratefully acknowledge support from the Ministry of Education of the Czech Republic (projects LM2010013 (LINDAT/CLARIN) and LH14011) and the Czech Science Foundation (project P406/12/0658). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project.

References

[1] Bejček E. (2015). *Automatické propojování lexikografických zdrojů a korpusových dat.* [*Automatic linking of lexicographic sources and corpus data.*] Ph.D. thesis, Charles University in Prague, Faculty of Mathematics and Physics.

[2] Bejček E., E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek and Š. Zikánová (2013). *Prague Dependency Treebank 3.0.* Data/software, Univerzita Karlova v Praze, MFF, ÚFAL, Prague.

[3] Lopatková M. (2008). Valence a její formální popis. Vybrané aspekty budování slovníku VALLEX. [Valency and Its Formal Description. Selected Aspects in Development of Valency Lexicon.] In: *Proceedings of Malý informatický seminář (MIS 2008).* Praha: Matfyzpress, pp. 58–88.

[4] Lopatková M., V. Kettnerová, E. Bejček, K. Skwarska and Z. Žabokrtský (2012). *VALLEX 2.6.* Data/software, ÚFAL MFF UK, http://ufal.mff.cuni.cz/vallex/2.6/. (The newest publicly available version is Vallex 2.7: http://ufal.mff.cuni.cz/vallex/2.7/.)

[5] Mikulová M., A Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá and Z. Žabokrtský (2006). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual*, Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague.

[6] Pareti, S. (2015). Annotating Attribution Relations across Languages and Genres. In: *Proceedings of the Eleventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, London, UK.

[7] Poláková L., J. Mírovský, A. Nedoluzhko, P. Jínová, Š. Zikánová and E. Hajičová (2013). Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing,* pp. 91—99, Nagoya, Japan.

[8] Prasad R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi and B. Webber (2008). The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008,* pp. 2961–2968, Marrakech, Morocco.