Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

ANALYSIS OF COREFERENTIAL EXPRESSIONS IN PAWS (ENGLISH-CZECH-RUSSIAN-POLISH PARALLEL TREEBANK WITH ANAPHORIC RELATIONS)

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz)[1],
Novák M. (mnovak@ufal.mff.cuni.cz)[1],
Ogrodniczuk M. (maciej.ogrodniczuk@ipipan.waw.pl)[2]

[1]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic; [2]Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland

In this paper, we decribe the coreference annotation on a multi-lingual parallel treebank (PAWS), a portion of Wall Street Journal translated into Czech, Russian and Polish which continues the tradition of multilingual treebanks with coreference annotation. The paper focuses on language-specific differences. We analyse syntactic structures concerning anaphoric relations in the languages under analysis, such as personal and impersonal constructions in polypredicative constructions and pro-drop qualities.

Keywords: parallel corpus, multilingual, coreference, Czech, English, Russian, Polish

АНАЛИЗ КОРЕФЕРЕНТНЫХ СВЯЗЕЙ В КОРПУСЕ PAWS (АНГЛО-ЧЕШСКО-РУССКО-ПОЛЬСКИЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС ЗАВИСИМОСТЕЙ С АНАФОРИЧЕСКОЙ РАЗМЕТКОЙ)

Недолужко А. Ю. (nedoluzko@ufal.mff.cuni.cz)[1],
Новак М. (mnovak@ufal.mff.cuni.cz)[1],
Огродничук М. (maciej.ogrodniczuk@ipipan.waw.pl)[2]

[1]Карлов университет, Прага, Чехия;
[2]Польская академия наук, Варшава, Польша

1. Introduction

In recent years, there appeared a number of multi-lingual parallel corpora annotated with referential relations. One of such corpora is the PAWS treebank, which stands for ***P***arallel ***A***naphoric ***W***all ***S***treet Journal. PAWS is a multi-lingual parallel treebank annotated with coreference relations [Nedoluzhko et al., 2018], it is freely available for non-commercial research and educational purposes[1]. Its current release consists of texts in four languages: English (original) and translations into Czech, Russian and Polish.

The aim of this paper is a contrastive analysis of how coreference relations are expressed in particular languages, based on the data from this treebank. The analysis is approached directly by contrasting aligned coreferential expressions in the languages, as it was already done for various expressions in English and Czech [Novák and Nedoluzhko, 2015] and reflexive possessives in English, Czech and Russian [Nedoluzhko et al., 2016a].

As the proposed treebank currently consists of three Slavic languages, it may serve as a valuable source for linguistic research on this language family. However, the translation factor should be taken into account. We deal with the translations from English into Slavic languages, so the direct calques between closely related Czech, Polish and Russian are not possible. On the other hand, translators treat the texts differently: Some of them stay closer to the texts, others try to primarily transfer the meaning, applying the mechanisms of explicitation and implicitation [Blum-Kulka, 1986]. Taking into account the relatively small dataset, the comparison of the resulting structures does not give statistically valuable results, although it gives a number of interesting observations.

The main feature of PAWS is its manual annotation of coreferential relations in all included languages. As two of the languages (Czech and Polish) extensively use zero subjects, we could miss a lot of valuable information if we annotated coreference only

[1] It can be downloaded from the Lindat/Clarin repository (http://hdl.handle.net/11234/1-2683).

on surface. Therefore, we adopted the style based on the theory of Functional Generative Description [Sgall et al., 1986], first used for Czech in the Prague Dependency Treebank 2.0 [Hajič et al., 2006] and for Czech and English in the Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2012]. In this style, coreference and other anaphoric relations are annotated on the layer of deep syntax called *tectogrammatical layer* which consists of dependency trees containing both explicitly expressed as well as important elided content words. Presence of elided words makes it possible to represent coreferential relations for dropped pronouns as well as for elided noun phrases in some specific syntactic constructions.

To facilitate the cross-lingual analysis, we equip the treebank with word alignment links between all nodes in all languages under analysis, including the reconstructed zeros. **Figure 1** (at the end of the paper) illustrates the annotation of a sample sentence in all four languages, as visualized by the TrEd tool [Pajas and Štěpánek, 2008]. Every sentence is represented as a dependency tree, with squared nodes representing the expressions elided on surface (cf. **`#Cor`** in the English sentence, **`#PersPron`** in the Czech and Polish sentences, etc.). The solid blue and red arrows correspond to coreferential links, word alignment is marked by dashed lines between the nodes in the trees (for clarity, the figure shows only alignment of coreferential expressions).

2. Related work

Our work relates to all multilingual parallel corpora with linguistic annotation, especially those for Slavic languages. ParaSol: A Parallel Corpus of Slavic and other languages [Waldenfels, 2006] is an aligned corpus of translated and original belletristic texts featuring automatic morphosyntactic annotations. The latest version comprises more than 30 languages. InterCorp [Čermák and Rosen, 2012] is another large multi-lingual parallel synchronic corpus with Czech as a pivot language, i.e. every text has its Czech version. It features part-of-speech tagging and lemmatization. The Polish-Russian Parallel Corpus [Laziński and Kuratczyk, 2016] features morphosyntactic description yet both sides differ as far as disambiguation is concerned (present in Polish, absent in Russian part). Paralela [Pezik, 2016] is a translation-based Polish-English corpus based on publicly available multilingual text collections and open-source parallel corpora featuring morphosyntactic annotation.

PAWS is also one of a few corpora annotated with coreference relations. Its English and Czech part directly corresponds to a subset of the Prague Czech-English Dependency Treebank 2.0 [$1] and its coreferential extension [Nedoluzhko et al., 2016b, PCEDT 2.0 Coref] and the Russian part corresponds to the PCEDT-R corpus [Nedoluzhko et al., 2016a], where the texts had been translated into Russian and aligned to Czech and English but they had not been annotated with coreferential relations there. ParCor 1.0 [$1] also belongs to this category. It is a German-English parallel corpus consisting of more than 8,000 sentences. Unlike PAWS, which has annotation of full corerence chains, only pronominal coreference is annotated in ParCor. On the other hand, texts in the corpus come from different genres, which is not the case in PAWS.

3. Data and Basic Statistics

The English texts originally come from the Wall Street Journal section of the Penn Treebank PTB. Czech, Russian and Polish texts have been translated by native speakers of the corresponding languages. English texts with their Czech translations have been extracted from Prague Czech-English Dependency Treebank 2.0 [Hajič et al. 2012]. The data consist of documents located in the first half of the PCEDT section 19 (wsj_1900 to wsj_1949). The basic statistics is shown in **Table 1**.

Table 1: Basic statistics for PAWS

	English	Czech	Russian	Polish
Documents	50			
Sentences	1,078			
Tokens	26,149	25,697	25,704	25,763

All texts have been annotated with rich linguistic information on dependency trees. For Czech and English, the annotation was copied from the PCEDT without any change. For Russian and Polish, the final tectogrammatical trees are slightly simplified and not always guaranteed to be correct, especially as concerns obligatory valency positions of predicates, semantic roles and some types of ellipses.[2]

4. Annotation of Coreference in PAWS

The coreference relations in PAWS have been annotated manually according to the Prague coreference annotation style [Nedoluzhko et al., 2016b]. The annotation covers the cases of grammatical (syntactic) and textual coreference.

The grammatical coreference typically occurs within a single sentence: These are the cases of relative and reflexive pronouns, verbs of control etc. By textual coreference, arguments are not realized by grammatical means alone, but also via context. Within this type, pronominal coreference of personal, possessive and demonstrative pronouns is annotated, as well as coreference with textual ellipsis, nominal textual coreference in case when the anaphoric expression is a full nominal group, anaphoric reference of local and temporal adverbs (*there, then* etc.) and textual reference to multiple antecedents (so-called *split antecedent*).

In case when an anaphoric expression refers endophorically to a discourse segment of more than one sentence, including the cases where the antecedent is understood by inference from a broader co-text, the special relation (*reference to a segment*) is annotated. This kind of relation has no explicitly marked antecedent.

We also have a specifically marked link for *exophora*, which denotes that the referent is "out" of the co-text, i.e. it is only known from the actual situation. Exophoric reference is annotated in case of temporal and local deixis (*this year, this country*), deixis with pronominal adverbs (*here*), as well as exophoric reference to the whole text.

[2] See [Nedoluzhko et al. 2018] for more details.

Table 2 shows the statistics of coreference-related annotation in PAWS.

Table 2: Coreference-related annotation in PAWS

	English	Czech	Russian	Polish
Tectogrammatical nodes	18,611	20,696	18,874	18,541
Coreferring nodes	4,210	4,403	4,254	3,371
grammatical coreference	729	528	749	294
textual pron. coref. expressed	544	213	493	206
textual pron. coref. elided	76	643	32	243
textual nominal coreference	1,361	1,496	1,610	1,568
first mentions	1,277	1,330	1,243	979
reference to split antecedents	149	149	91	65
reference to a segment	28	23	16	12
exophora	46	21	20	4

5. Contrastive analysis of coreference relations statistics

The brief inspection of Table 2 shows that there are significant differences in the numbers of relations between the languages under analysis. Some of these differences may be caused by the simplification of the tectogrammatical annotation for Polish, and partly also for Russian. For example, we observe that the number of coreferring nodes in Polish is smaller than in the three remaining languages. The reason is that we did not reconstruct all unexpressed valency positions for Polish (e.g. we didn't insert elided Addressee for the verbs of speech (such as *say*, *claim*, *contend*, etc.) which may be connected by coreference relations. Such relations are rather formal, but technically they are missing in Polish, thus reducing the total of coreferring nodes.

Other differences may reflect the varieties in the grammatical structures or different grammatical tendencies in the languages.

For example, in Table 2, we observe that the number of tectogrammatical nodes in Czech is larger than in the three remaining languages. This could be caused by the translator's style, in this case it would be the tendency of the Czech translator to larger explicitation [Blum-Kulka, 1986]. However, the manual analysis of the texts shows a strong tendency of Czech to use finite subordinated clauses instead of non-finite infinitive or gerundial clauses in English, Polish and Russian. Finite constructions are natually longer than infinite ones, so the larger number of tectogrammatiacl nodes in Czech could be also explained by this reason. Consider Example 1, where, the gerundial clause in English (*continuing a rebound from steep year-ago losses*) is naturally translated into infinite clauses in Polish and Russian, but it is transferred to a finite subordinate clause (*čímž pokračuje v zotavení z velkých loňských ztrát*) in Czech. Both in Polish and Russian, the translation with a finite subordinate clause is also possible, but, as the data show, this is not often the case: On the one hand, infinite constructions are fully acceptable in these two languages, on the other hand, gerundial constructions in English naturally trigger the similar ones in the target language. As for Czech, an infinite clause is not acceptable in this case.

Example 1:

EN: *Morrison Knudsen Corp. posted third-quarter net income of $7.9 million,* ***continuing*** *a rebound from steep year-ago losses.*

PL: *Morrison Knudsen Corp. zaksięgował dochód netto za trzeci kwartał równy 7,9 milionom dolarów,* ***kontynuując*** *odbicie po znacznych zeszłorocznych stratach.*

CZ: *Společnost Morrison Knudsen Corp. vykázala čistý zisk za třetí čtvrtletí ve výši 7.9 miliónu dolarů,* ***čímž pokračuje*** *v zotavení z velkých loňských ztrát.*

RU: *Корпорация Morrison Knudsen опубликовала данные о чистых доходах, составивших $7,9 млн или 69 центов за акцию, в третьем квартале,* ***продолжая*** *восстанавливаться после больших прошлогодних убытков.*

The prevailing personal subordinate clauses in polypredicative constructions with (both expressed and unexpressed) pronouns in Czech also correlates with the biggest number of coreferring nodes in Czech, as follows from the statistics of the PAWS coreference-related annotation in **Table 2**.

The tendency to impersonal constructions in Polish and Russian is very strong. In some cases, they even tend to be grammaticalized, as in **Example 2**, where the impersonal gerundial constructions *based / bazując / исходя из* function more like secondary prepositions[3]. In this example, the grammatical coreference of the first argument of the gerundial form is problematic, and both in Polish and Russian the use of a gerundial form conflicts grammatical rules of these languages, saying that, e.g. for Russian, an animate subject should be the prototypical coreferential antecedent for the gerund. This conflict is one of the arguments of grammaticalization.

Example 2:

EN: ***Based*** *on the number of Mesa shares* […]*, the proposed takeover would have a value of about $15.3 million.*

PL: ***Bazując*** *na pozostałej liczbie akcji Mesy* […] *proponowane przejęcie osiągnęłoby wartość około 15,3 milionów dolarów.*

RU: *Запланированное поглощение,* ***исходя из*** *количества акций Mesa* […] *имело бы стоимость почти $15,3 млн.*

Another interesting fact following from the coreference annotation statistics in Table 2 is the highest number of grammatical coreference relations in Russian[4], which can be partially explained by a large number of infinitive constructions, where unexpressed subjects are controlled by the actants of their governing control verbs

[3] In the given example, the gerundial forms in Polish (*bazując*) and Russian (*исходя из*) are very close to the English one (based). However, the syntactic construction is slightly different, so it should not be considered as a calcue.

[4] In Polish, on the contrary, it is very small. The reason for the small number in Polish is the missing annotation of the control verbs coreference.

by means of grammatical coreference. In Example 3, the infinitive clause *to employ a financial consultant to advise them* is translated with an infinitive clause into Russian, as a deverbative construction into Polish and as a subordinate clause into Czech:

Example 3:

EN: *In response to the specific offer, Gary Risley, Mesa vice president, said management will ask directors* ***to employ*** *a financial consultant to advise them*.

PL: *W odpowiedzi na szczegółową ofertę, Gary Risley, zastępca prezesa Mesy, powiedział, że zarząd poprosi dyrektorów* ***o zatrudnienie*** *konsultanta finansowego w celach doradczych*.

CZ: *Gary Risley, vicepresident společnosti Mesa, uvedl, že jako odpověď na konkrétní nabídku požádá vedení společnosti představenstvo,* ***aby použilo služeb finančního poradce***.

RU: *В ответ на конкретное предложение Гэри Рисли, вице-президент Mesa, сказал, что руководство попросит директоров* ***нанять*** *финансового советника для получения консультации*.

Interestingly, in Example 3, the infinitive construction is the only possible one in the Russian translation. In Polish, the deverbative construction (*o zatrudnienie*) can be changed to the infinitive one or to a finite subordinate clause. In Czech, the finite subordinate clause (*aby použilo služeb finančního poradce*) can be changed to either an infinitive or a deverbative clause.

The difference in corresponding numbers of coreferential nodes in Table 2 is also influenced by the frequent use of deverbatives in translations in all three Slavic languages. See Example 4, where the original finite clause is translated to deverbative clauses into Polish, Czech and Russian.[5]

Example 4:

EN: *Last week, Mesa rejected a general proposal from StatesWest* ***that the two carriers combine***.

CZ: *Minulý týden společnost Mesa odmítla základní nabídku společnosti StatesWest* ***na sloučení*** *obou přepravců*.

PL: *W zeszłym tygodniu Mesa odrzuciła ogólną propozycję StatesWest* ***dotyczącą połączenia*** *obu przewoźników*.

RU: *На прошлой неделе Mesa отклонила общее предложение от StatesWest* ***об объединении*** *двух перевозчиков*.

[5] In this case, this is rather a technical issue pointing on the fact that coreference annotation of the arguments of deverbatives is a very complicated task which was not completed consistently for none of the languages under analysis.

Finally, the point of explicitly expressed textual pronominal coreference is especially interesting, as it shows the different degree of pro-drop quialities of English, Czech, Polish and Russian. As observed from Table 2, explicitly expressed textual pronominal coreference is most frequent in English (544 cases). Indeed, in English, there is no possibility for subject omission, whereas for Slavic languages this often happens. However, the subject can be omitted in the analysed languages to a different degree. Czech is a highly pro-drop language, where anaphoric use of personal pronouns in the subject position is untypical. On the other hand, Polish and Russian show substantially lower degree of pro-drop qualities, Polish being less pro-drop than Czech, but significantly more pro-drop than Russian. Our numbers here correspond to the analysis in [Kibrik, 2011], where the distribution of pro-drop qualities in these languages is the same. The big number of elided coreferential nodes in Czech (643 relations) also supports this statement.

6. Translation factor

The comparison of the parallel sentences in the languages under analysis shows that in many cases the choice of a language expression is not given by the grammatical structure of the corresponding language, but it is triggered by the syntactic structure of the original English sentence. This factor is very important when analysing translated texts and it may potentially explain many statistical differencies. For example, Table 2 gives evidence that coreference is more frequently realized by nominal groups in Russian than in the other languages (1,610 cases). This could be a translation effect that should be however proved by comparison with other translations. The same is true about the difference in the number of tectogrammatical nodes between the languages.

Moreover, the specifity of the texts (mostly business-focused news) causes a number of calcues which make the analysis on the textual level rather problematic.

7. Conclusion

In this work, we presented the basic statistics of coreference-related annotation in the PAWS treebank, a multi-lingual parallel treebank with manual annotation of coreferential relations in English, Czech, Russian and Polish. We proposed explanations to some differences between the languages under analysis, as concerns the number of tectogrammatical nodes, coreferring expressions, grammatical coreference or pronouns. The basic reasons for these differences are (i) in the preferrable use of finite constructions in Czech and infinite constructions in English, Russian and Polish; (ii) in the different pro-drop qualities of the languages. Furthermore, the translation factor is crusial, especially given the relatively small number of the annotated sentences.

8. Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (project GA16-05394S) and the Polish National Science Centre (contract number 2014/15/B/HS2/03435). The research reported in the present contribution

has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

1. *Blum-Kulka, Sh.* (1986). Shifts of Cohesion and Coherence in Translation. J. House, Sh. Blum-Kulka (eds): Interlingual and Intercultural Communication. Tübingen: Narr, 17-35.
2. *Čermák, F. and Rosen, A.* (2012). The case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 17(3):411–427.
3. *Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B.* (2014). ParCor 1.0: A Parallel Pronoun Coreference Corpus to Support Statistical MT. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland. European Language Resources Association.
4. *Hajič J., Panevová J., Hajičová E., Sgall P., Pajas P., Štěpánek J., Havelka J., Mikulová M., Žabokrtský Z., Ševčíková-Razímová M., Urešová Z.* (2006): Prague Dependency Treebank 2.0. Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA.
5. *Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z.* (2012): Announcing Prague Czech-English Dependency Treebank 2.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), European Language Resources Association, İstanbul, Turkey, pp. 3153-3160.
6. *Kibrik, A.* (2011). Reference in Discourse. Oxford, United Kingdom.
7. *Laziński, M. and Kuratczyk, M.* (2016). The University of Warsaw Polish-Russian Parallel Corpus. In Ewa Gruszczynska et al., editors, Polish-Language Parallel Corpora, pages 83–95. Instytut Lingwistyki Stosowanej UW, Warsaw.
8. *Nedoluzhko, A., Khoroshkina, A. S., and Novák, M.* (2016a). Possessives in Parallel English-Czech-Russian Texts. Computational Linguistics and Intellectual Technologies, (15): pp. 483–497.
9. *Nedoluzhko, A., Novak, M., Cinková, S., Mikulová, M., and Mírovský, J.* (2016b). Coreference in Prague Czech-English Dependency Treebank. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 169–176, Paris, France. European Language Resources Association.
10. *Novak, M. and Nedoluzhko, A.* (2015). Correspondences between Czech and English Coreferential Expressions. Discours: Revue de linguistique, psycholinguistique et informatique., 16:1–41.
11. *Nedoluzhko, A., Novak, M., Ogrodniczuk, M.* (2018). PAWS: A Multi-lingual Parallel Treebank with Anaphoric Relations. In M. Poesio, editor, CRAC: Computational Models of Reference, Anaphora, and Coreference, co-located with NAACL 2018. USA, New Orleans, The Association for Computational Linguistics.

12. *Pajas, P. and Stěpánek, J.* (2008). Recent Advances in a Feature-rich Framework for Treebank Annotation. In Proceedings of the 22nd International Conference on Computational Linguistics—Volume 1, Stroudsburg, PA, USA. Association for Computational Linguistics.
13. *Pezik, P.* (2016). Exploring Phraseological Equivalence with Paralela. In Ewa Gruszczynska et al., editors, Polish-Language Parallel Corpora, pages 67–81. Instytut Lingwistyki Stosowanej UW, Warsaw.
14. *Sgall, P., Hajicová, E., and Panevová, J.* (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Dordrecht, Netherlands.
15. *Waldenfels, v. R.* (2006). Compiling a parallel corpus of Slavic languages. In B. Brehmer, et al., editors, Text strategies, tools and the question of lemmatization in alignment. Beitrage der Europäischen Slavistischen Linguistik (POLYSLAV 9), pages 123–138. München.

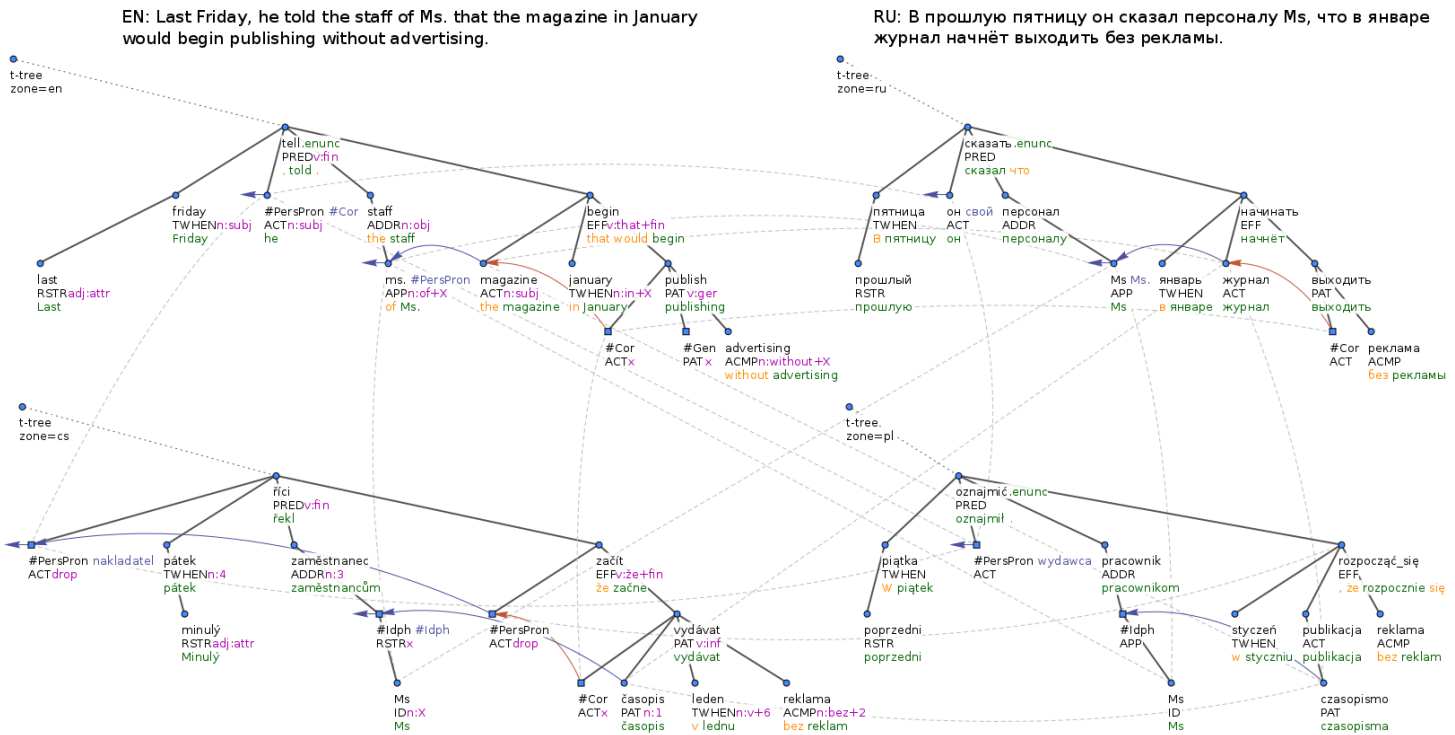

Figure 1: Tectogrammatical representation of a sample sentence in all four languages, visualized by the TrEd tool