PAWS: A Multi-lingual Parallel Treebank with Anaphoric Relations

Anna Nedoluzhko and **Michal Novák**
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
{nedoluzko,mnovak}@ufal.mff.cuni.cz

Maciej Ogrodniczuk
Polish Academy of Sciences
Institute of Computer Science
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

We present PAWS, a multi-lingual parallel treebank with coreference annotation. It consists of English texts from the Wall Street Journal translated into Czech, Russian and Polish. In addition, the texts are syntactically parsed and word-aligned. PAWS is based on PCEDT 2.0 and continues the tradition of multilingual treebanks with coreference annotation. The paper focuses on the coreference annotation in PAWS and its language-specific differences. PAWS offers linguistic material that can be further leveraged in cross-lingual studies, especially on coreference.

1 Introduction

In recent years, we have witnessed a rise in multilingual approaches to both theoretical and computational linguistics. Coreferential and anaphoric relations are no exception. For instance, the CoNLL 2012 Shared Task (Pradhan et al., 2012) has focused on modeling coreference in three different languages, making use of the data from the OntoNotes corpus (Weischedel et al., 2013). Since then, several other multilingual parallel corpora annotated with referential relations were produced (see Section 2). In this work, we go even further. We present the PAWS treebank, a multi-lingual parallel treebank annotated with full-fledged coreference relations. Its current release consists of texts in four languages: English, Czech, Russian and Polish.

A decision to build such treebank has multiple motivations, mostly related to cross-lingual studies of coreference relations.

First, construction of such corpus tests applicability of a particular annotation schema for other languages. The project of Universal Dependencies[1] has shown that efforts devoted to seeking a language-universal syntactic and morphological representation may open up a space for novel research within the field. Concerning coreference, a single annotation schema has been applied to English, Chinese and Arabic already in OntoNotes 5.0 (Weischedel et al., 2013) and on parallel English-German-Russian texts by Grishina and Stede (2015).

Second, from a perspective of theoretical linguistics, a cross-lingual view on particular linguistic phenomena may give us more information than a monolingual view. The present work focuses on three Slavic languages, which despite their apparent closeness exhibit considerable differences in phenomena related to coreference, e.g. a degree of using pro-drops, or diverse usage of reflexive pronouns. With our corpus such phenomena can be directly compared across languages. This work thus follows on the comparative analysis that has been previously conducted on coreferential expressions in English and Czech (Novák and Nedoluzhko, 2015) and reflexive possessives in English, Czech and Russian (Nedoluzhko et al., 2016a).

Last but not least, a new coreference-annotated parallel corpus may drive a research on cross-lingual automatic approaches related to coreference. It includes coreference projection (Postolache et al., 2006; Grishina and Stede, 2017) and bilingually-informed coreference resolution (Mitkov and Barbu, 2003; Novák and Žabokrtský, 2014). Unlike ParCor 1.0 (Guillou et al., 2014), PAWS is not tailored to machine translation experiments. Nevertheless, its parallel nature suggests that it can also be leveraged for these purposes.

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

[1] http://universaldependencies.org

Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference, pages 68–76
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

face. Therefore, we adopted the style based on the theory of Functional Generative Description (Sgall et al., 1986), first used for Czech in Prague Dependency Treebank 2.0 (Hajič et al., 2006) and for Czech and English in 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*. It consists of dependency trees containing both overt as well as important elided content words. Presence of elided words makes it possible to represent coreferential relations even for dropped pronouns.

To facilitate cross-lingual studies, we equip the treebank with word alignment links between all the language pairs. Since these links are annotated on the tectogrammatical layer, they also cover the reconstructed zeros. Most of the alignment links were collected automatically. However, for selected types of coreferential expressions, we labeled the alignment links also manually.

Figure 1 illustrates the annotation of a sample sentence in all 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. Whereas the solid 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 Corpora

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 (Pęzik, 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 (Hajič et al., 2012, PCEDT) and its coreferential extension (Nedoluzhko et al., 2016b, PCEDT 2.0 Coref). ParCor 1.0 (Guillou et al., 2014) 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 coreference 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 PAWS Data and Its Rich Annotation

This paper presents the PAWS treebank, which stands for *P*arallel *A*naphoric *W*all *S*treet Journal. In its current version it comprises parallel texts in English, Czech, Russian, and Polish.

English texts with their Czech translations were extracted from Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012). Namely, the data consist of 50 documents from sections `wsj1900-49`. The English texts originally come from the Wall Street Journal section of Penn Treebank (Marcus et al., 1999).

Russian and Polish texts were translated from English by one native speaker for each of the target languages. The translations were revised and corrected by the translators again, if necessary. Basic statistics of the collected texts is shown in the upper part of Table 1.

All the texts were annotated with rich linguistic information stratified into two layers of dependency trees – the surface and deep syntax (tectogrammatical) layer. Whereas the English and Czech annotation was copied from the PCEDT without any change, we produced the Russian and Polish annotation entirely within this project.

In PCEDT, English surface syntax trees had been built by transforming manually annotated constituency trees in Penn Treebank. On the other hand, Czech surface syntax trees had been created automatically by tools available in the multi-purpose NLP framework Treex (Popel and Žabokrtský, 2010). Both the English and Czech tectogrammatical layer had been annotated manu-

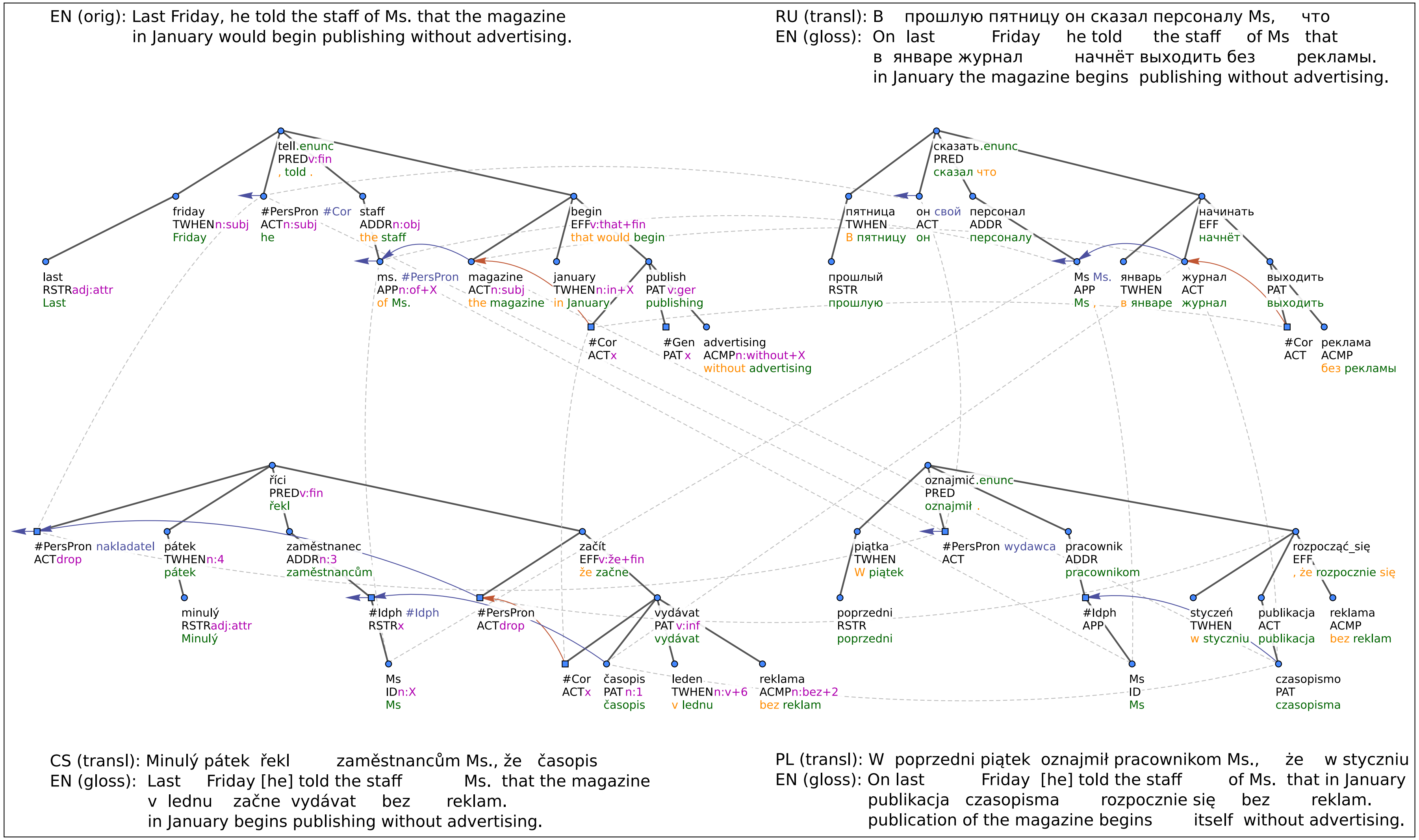

EN (orig): Last Friday, he told the staff of Ms. that the magazine
in January would begin publishing without advertising.
RU (transl): В прошлую пятницу он сказал персоналу Ms, что
EN (gloss): On last Friday he told the staff of Ms that
в январе журнал начнёт выходить без рекламы.
in January the magazine begins publishing without advertising.
tell.enunc PREDv:fin , told .
last RSTRadj:attr Last
friday TWHENn:subj Friday
#PersPron #Cor ACTn:subj he
staff ADDRn:obj the staff
begin EFFv:that+fin that would begin
ms. #PersPron APPn:of+X of Ms.
magazine ACTn:subj the magazine
january TWHENn:in+X in January
publish PAT v:ger publishing
#Cor ACTx
#Gen PAT x
advertising ACMPn:without+X without advertising
сказать.enunc PRED сказал что
пятница TWHEN В пятницу
он свой ACT он
персонал ADDR персоналу
начинать EFF начнёт
прошлый RSTR прошлую
Ms Ms. APP Ms ,
январь TWHEN в январе
журнал ACT журнал
выходить PAT выходить
#Cor ACT
реклама ACMP без рекламы
říci PREDv:fin řekl
#PersPron nakladatel ACTdrop
pátek TWHENn:4 pátek
zaměstnanec ADDRn:3 zaměstnancům
začít EFFv:že+fin že začne
minulý RSTRadj:attr Minulý
#Idph #Idph RSTRx
#PersPron ACTdrop
vydávat PAT v:inf vydávat
Ms IDn:X Ms
#Cor ACTx
časopis PAT n:1 časopis
leden TWHENn:v+6 v lednu
reklama ACMPn:bez+2 bez reklam
oznajmić.enunc PRED oznajmił .
piątka TWHEN W piątek
#PersPron wydawca ACT
pracownik ADDR pracownikom
rozpocząć_się EFF , że rozpocznie się
poprzedni RSTR poprzedni
#Idph APP
styczeń TWHEN w styczniu
publikacja ACT publikacja
reklama ACMP bez reklam
Ms ID Ms
czasopismo PAT czasopisma
CS (transl): Minulý pátek řekl zaměstnancům Ms., že časopis
EN (gloss): Last Friday [he] told the staff Ms. that the magazine
v lednu začne vydávat bez reklam.
in January begins publishing without advertising.
PL (transl): W poprzedni piątek oznajmił pracownikom Ms., że w styczniu
EN (gloss): On last Friday [he] told the staff of Ms. that in January
publikacja czasopisma rozpocznie się bez reklam.
publication of the magazine begins itself without advertising.

Figure 1: Tectogrammatical representation of a sample sentence in all four available languages, visualized by the TrEd tool. For clarity, we show alignments for coreferential nodes only.

ally from scratch.

The amount of automatic processing was even larger in the Russian and Polish annotations. Treex was employed to obtain both types of trees. Joint part-of-speech tagging, morphological analysis and dependency parsing provided by the UDPipe tool (Straka et al., 2016) were the key elements to build surface syntax trees. These trees were then transformed to tectogrammatical ones by a mostly generic sequence of rule-based modifications.

In other words, the final tectogrammatical trees are simplified and not always guaranteed to be correct, especially in the following aspects:

- Lemmata. Lemmata were set automatically for Russian and Polish and they have been corrected consistently only for expressions that take (or may take) part in coreference relations. The remaining nodes have been mostly corrected during the annotation of tectogrammatical structure and coreference, if annotators discovered a mistake, but no special check has been carried out.

- Obligatory valency positions of predicates. Unlike English and Czech, for which valency lexicons[2] had been used for consistent annotation of valency modifications, we used no such lexicons for Russian and Polish.

- Semantic roles. For Russian, we manually annotated semantic roles for arguments of a predicate, temporal, locative, and causal adjuncts etc. The annotation of semantic roles followed the guidelines for annotation on the tectogrammatical level in the Prague Dependency Treebank for Czech and English (Mikulová et al., 2007), but it was simplified in some respects. For example, instead of nine temporal roles, we used only three basic ones for Russian. As for Polish, semantic roles have not been annotated yet; we plan to add them in future development of the corpus.

- Ellipses. Whereas the English and Czech ellipses had been added by the rules used for the Prague Dependency Treebank[3], the inventory of reconstructed ellipsis types in Russian and Polish was narrowed. It includes only the cases necessary for coreference annotation.

- Identification structures. For example, in the sample sentence in Figure 1, the name of the magazine (*Ms.*) is marked as an identification structure (with a special governing node *#Idph*) in English and Czech. However, this is not the case of Russian and Polish, where the tectogrammatical structure is more simple.

4 Annotation of Coreference in PAWS

The coreference annotation of PAWS has been conducted manually according to Prague coreference annotation style (Nedoluzhko et al., 2016b).[4] It takes place on the tectogrammatical layer to allow for marking zero anaphora. The annotation covers the cases of grammatical (syntactic) and textual coreference.

The **grammatical coreference** typically occurs within a single sentence, the antecedent is expected to be derived on the basis of grammar rules of a given language. These are the cases of relative and reflexive pronouns, verbs of control, coreference of arguments hidden in reciprocal constructions (***Peter***$_i$ *and* ***Mary***$_j$ *kissed* ***Ø***$_{i+j}$). and coreference with verbal modifications that have dual dependency (*John saw* ***Mary*** *[****Ø*** *run around the lake]*). All the cases of grammatical coreference have been systematically annotated for English and Czech (Nedoluzhko et al., 2016b). For Russian and Polish, grammatical coreference annotation has been consistently provided for the cases of relative and reflexive pronouns. Coreference of arguments of verbs of control and coreference in reciprocal constructions have been manually annotated for Russian but only partly for Polish. However, this task is not especially urgent for our planned comparative analysis of coreferential expressions. In all four analyzed languages, the controllees of the arguments of control verbs, second arguments in reciprocal constructions and arguments in constructions with dual dependencies are unexpressed, thus the results of the comparison will be mostly trivial. For example in Figure 1, the unexpressed controllee is reconstructed as the first argument of the verb *publish* and its counterparts in Czech and Russian (see the dependent

[2] Lexicons PDT-Vallex (Hajič et al., 2003) and Engvallex (Urešová, 2012) for Czech and English, respectively.

[3] Described in more details in (Mikulová, 2014).

[4] This paper also describes comparative analysis of our approach with coreference annotation in the OntoNotes.

node with the lemma *#Cor* under the node *publish*).[5] It is controlled by the first argument of the verb *begin* (Czech: *začít*, Russian: начинать) and it cannot be explicitly expressed in either of the languages.

By **textual coreference**, arguments are not realized by grammatical means alone, but also via context. Within this type, we annotate the following relations:

- Pronominal coreference of personal, possessive and demonstrative pronouns (e.g., *Mary – she – her*).
- Coreference with textual ellipsis, for example coreference of zero subjects in pro-drop languages. This is the case of the unexpressed subject *he* in the Czech and Polish translations of the main clause *he told the staff of Ms.* in the running example (see Figure 1). In such cases, the special node *#PersPron* is added to the tectogrammatical tree and the coreference relation to the antecedent in the previous context is annotated (as shown in the figure). Interestingly, in the dependent clause of this sentence, the subject is dropped only in Czech and it is not cross-lingually coreferential with the expressions at the same position in the other languages (In Czech, it is coreferential with the subject of the main clause *he*; in English and Russian, this is the magazine; in Polish, this is the publication (*publikacja*)).
- Nominal textual coreference in case when the anaphoric expression is a full nominal group (noun with or without modifications) coreferring with an antecedent in the preceding context. In the running example, such relation is held between *magazine* (Polish: *czasopismo*, Czech: *časopis*, Russian: журнал), the name of this magazine *Ms.* in the same sentence and an antecedent in one of the previous sentences.
- Anaphoric reference of local and temporal adverbs (*there*, *then*, etc.).
- Textual reference to multiple antecedents (so-called *split antecedent*). In this case, there are (technically) two coreference links of a special type, pointing to the split parts of the antecedent.

In the same way as for the other coreference-annotated corpora with Prague-style annotation, the textual coreference is marked in case of anaphoric references to events (so-called *abstract anaphora*), i.e. anaphoric references to verbal groups, clauses, sentences and larger textual segments (Nedoluzhko and Lapshinova-Koltunski, 2016). If the antecedent does not exceed one sentence, it is annotated in the same way as other coreference types, the root of the verbal phrase being the antecedent of a pronominal element.

If 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 context, a special relation with no explicitly marked antecedent is annotated.

We also specifically mark presence of *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 cases of temporal and local deixis (*this year, this country*), deixis with pronominal adverbs (*here*), as well as exophoric reference to the whole text.

In accordance with the Prague coreference annotation tradition, textual coreference is marked up to the length of 20 sentences.

For more detailed description and examples of the applied coreference annotation scheme, see (Nedoluzhko et al., 2016b).

5 Statistics and Observations

The bottom part of Table 1 shows the statistics of coreference-related annotation in PAWS. Here are the main observations:

1. **The number of tectogrammatical nodes in Czech is larger** than in the three remaining languages. This could be caused either by the translator’s style or by some language-specific features of Czech. The answer to this question requires further comparison (first of all to other translated and non-translated texts) but manual analysis of the texts shows a strong tendency in Czech to use finite subordinated clauses instead of non-finite infinitive or gerundial clauses in English, Polish and Russian. Finite constructions are naturally longer than infinite ones.

[5] In Polish, the sentence has a different syntactic structure, so the argument cannot be reconstructed.

	English	Czech	Russian	Polish
Sentences	1,078			
Tokens	26,149	25,697	25,704	25,763
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. overt	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

Table 1: Statistics of the data and its coreference-related annotation.

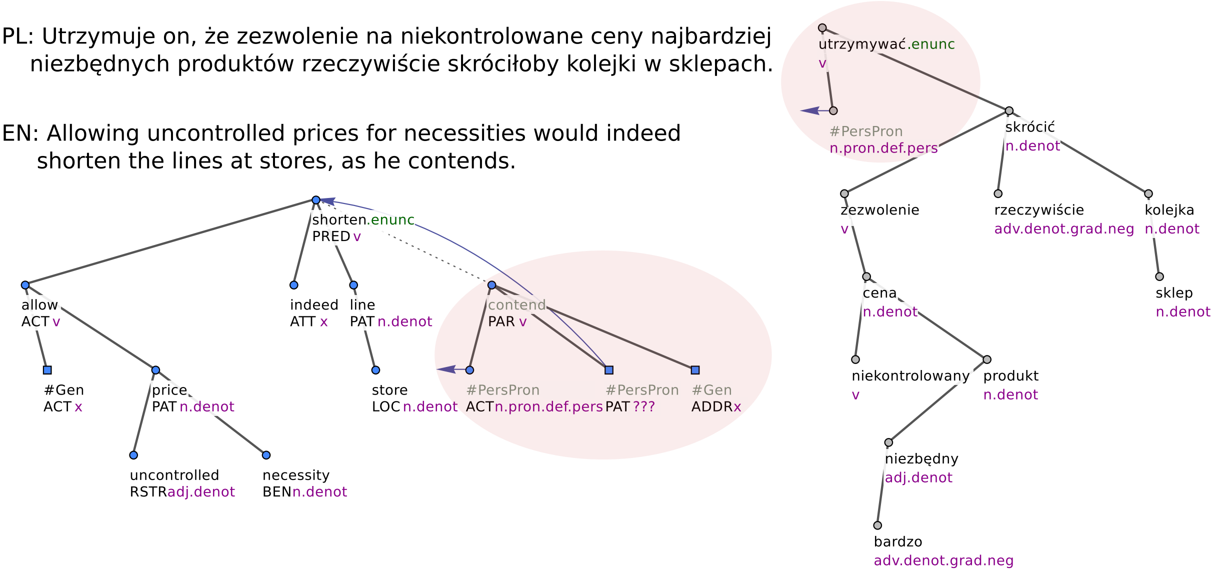
2. **The number of coreferring nodes in Polish is smaller** than in the three remaining languages. The explanation for this substantial difference is in the simplification of the tectogrammatical annotation for Polish. To keep the annotation consistency for different kind of complicated syntactic structures, the tectogrammatical annotation rules for Czech, English and Russian are very sophisticated. For example, for verbs of speech (e.g., *say*, *claim*, *contend*), the valency position of the verbal content has been reconstructed in the tectogrammatical tree (according to verbal valency lexicons for these languages), even if it is not explicitly expressed in the corresponding clause. See Figure 2, where two obligatory valencies are reconstructed for English, but not for Polish.

3. On the other hand, **the biggest number of coreferring nodes is in Czech**. This correlates with the greater amount of tectogrammatical nodes as well as to the fact that Czech uses personal constructions with overt and unexpressed pronouns more frequently. Besides, this high number reflects especially detailed manual annotation of tectogrammatical level, by which the omitted valency positions have been reconstructed also by a large part of deverbatives, which was not the case for other languages.

4. **The number of grammatical coreference relations is the largest in Russian**. 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 (see Section 4). As for the large number for Russian, it can be partially explained by a large number of infinitive constructions, where unexpressed subjects are controlled by the actants of their governing control verbs by means of grammatical coreference.

5. **Overt textual pronominal coreference**. This point is especially interesting, as it shows the different degree of pro-drop qualities of English, Czech, Polish and Russian. As observed from the table, overt textual pronominal coreference is most frequent in English. 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 analyzed 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 slightly more pro-drop than Russian (Kibrik, 2011).

6. Another observation supported by the brief inspection of Table 1 is that **coreference is more frequently realized by nominal groups in Russian** than in the other languages. This observation requires further analysis. This could be a translation effect that should be however proved by compar-



PL: Utrzymuje on, że zezwolenie na niekontrolowane ceny najbardziej niezbędnych produktów rzeczywiście skróciłoby kolejki w sklepach.

EN: Allowing uncontrolled prices for necessities would indeed shorten the lines at stores, as he contends.

Figure 2: Different tectogrammatical representation of the English sentence and its Polish translation. The English sentence contains additional unexpressed coreferential node.

ison with other translations. On the other hand, the annotation effect is also possible. In some cases, especially in cases of nominal coreference, a coreference relation may be ambiguous (so-called near-identity (Recasens et al., 2011)) and it is up to the annotator, whether to annotate it or not. In such cases, the use of anaphoric markers can influence the annotator's decision: In case of explicit anaphoric reference, the relation is more likely to be annotated.

6 Word Alignment

The annotated texts are equipped with word alignment between each pair of the languages, both on the surface and deep syntax representations. Alignment links were collected by running GIZA++ (Och and Ney, 2000) on a union of the texts in question and a large number of additional parallel texts. The additional parallel texts were collected using the OPUS project (Tiedemann, 2012) and their size was roughly 15 million sentence pairs for each language pair. The word alignment was then projected to the tectogrammatical layer and complemented with alignment for reconstructed nodes using syntax-based heuristics.

For selected types of coreferential expressions, we annotated their cross-lingual counterparts also manually. Particularly, we marked alignment of English, Czech and Russian pronouns and zeros to their counterparts in each of these three languages.[6] Polish is not covered by manual alignment, yet.

7 Availability

PAWS is freely available for non-commercial research and educational purposes. It can be downloaded from the Lindat/Clarin repository.[7] The treebank is released in the following file formats:

Plain text format. The texts with inline annotation of coreferential mentions. This format also contains reconstructed ellipses, which can be easily removed by running a script that we provide in the release.

Treex XML format. The internal format of PAWS contains the entire annotation. Documents in this format can be viewed using the TrEd tool.

CoNLL 2012 format. This format was used for the CoNLL 2012 Shared Task in coreference resolution. As this format allows for representing surface words only, it does not include all annotated mentions and anaphoric links, especially for prodrop languages.

8 Conclusion

In this work, we introduced the PAWS treebank: a multi-lingual parallel treebank with manual annotation of coreferential relations and cross-lingual

[6] It extends the annotation of English-Czech alignment already provided in PCEDT 2.0 Coref.

[7] `http://hdl.handle.net/11234/1-2683`

alignment between selected types of coreferential expressions. The treebank currently comprises English texts and its Czech, Russian and Polish translations.

We have primarily built PAWS for future analysis on difference between the languages in terms of how they express coreference. Nevertheless, due to its extensive annotation of syntax, semantic roles, coreference relations and alignment it may serve as a basis for many different linguistic studies. Cross-lingual analysis of any phenomena can bring a deeper insight and allow for its better understanding than if each of the languages was analyzed in isolation.

9 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). We also thank three anonymous reviewers for their instructive comments.

References

Yulia Grishina and Manfred Stede. 2015. Knowledge-Lean Projection of Coreference Chains across Languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 14–22, Beijing, China. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multi-Source Annotation Projection of Coreference Chains: Assessing Strategies and Testing Opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 41–50, Valencia, Spain. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 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)*, pages 3191–3198, Reykjavik, Iceland. European Language Resources Association.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. 2006. Prague Dependency Treebank 2.0. Philadelphia, USA. Linguistic Data Consortium.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Andrej A. Kibrik. 2011. *Reference in Discourse*. Oxford University Press, Oxford, United Kingdom.

Marek Laziński and Magdalena Kuratczyk. 2016. The University of Warsaw Polish-Russian Parallel Corpus. In *Polish-Language Parallel Corpora*, pages 83–95. Instytut Lingwistyki Stosowanej UW, Warsaw, Poland.

Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Penn Treebank 3. Philadelphia, USA. Linguistic Data Consortium.

Marie Mikulová. 2014. Semantic Representation of Ellipsis in the Prague Dependency Treebanks. In *Proceedings of the Twenty-Sixth Conference on Computational Linguistics and Speech Processing ROCLING XXVI (2014)*, pages 125–138, Taipei, Taiwan. Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Ševčíková, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2007. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Technical Report 3.1. ÚFAL, Charles University. Prague, Czech Republic.

Ruslan Mitkov and Catalina Barbu. 2003. Using Bilingual Corpora to Improve Pronoun Resolution. *Languages in contrast*, 4(2).

Anna Nedoluzhko, Anna Schwarz (Khoroshkina), and Michal Novák. 2016a. Possessives in Parallel English-Czech-Russian Texts. *Computational Linguistics and Intellectual Technologies*, (15):483–497.

Anna Nedoluzhko and Ekaterina Lapshinova-Koltunski. 2016. Abstract Coreference in a Multilingual Perspective: A View on Czech and German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 47–52, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 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.

Michal Novák and Anna Nedoluzhko. 2015. Correspondences between Czech and English Coreferential Expressions. *Discours: Revue de linguistique, psycholinguistique et informatique.*, 16:1–41.

Michal Novák and Zdeněk Žabokrtský. 2014. Cross-lingual Coreference Resolution of Pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petr Pajas and Jan Štěpánek. 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.

Piotr Pęzik. 2016. Exploring Phraseological Equivalence with Paralela. In *Polish-Language Parallel Corpora*, pages 67–81. Instytut Lingwistyki Stosowanej UW, Warsaw.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, pages 293–304, Berlin, Heidelberg. Springer-Verlag.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring Coreference Chains through Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. European Language Resources Association.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, Non-identity, and Near-identity: Addressing the Complexity of Coreference. *Lingua*, 121:1138–1152.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, Netherlands.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Paris, France. European Language Resources Association.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Paris, France. European Language Resources Association.

Zdeňka Urešová. 2012. Building the PDT-VALLEX Valency Lexicon. In *Proceedings of the fifth Corpus Linguistics Conference*, pages 1–18, Liverpool, UK. University of Liverpool.

František Čermák and Alexandr Rosen. 2012. The Case of InterCorp, a Multilingual Parallel Corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Ruprecht von Waldenfels. 2006. Compiling a Parallel Corpus of Slavic Languages. Text Strategies, Tools and the Question of Lemmatization in Alignment. *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, 9:123–138.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0. Philadelphia, USA. Linguistic Data Consortium.