Through Low-Cost Annotation to Reliable Parsing Evaluation

Marek Grác, Miloš Jakubíček, Vojtěch Kovář
{xgrac,xjakub,xkovar3}@fi.muni.cz

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

Abstract. In this paper, we present an application-driven low-cost concept of building a multi-purpose language resource for Czech which is based on currently available results of previous work by various research teams active in the area of natural language processing. We particularly focus on the first phase which consists in extracting noun phrases from a morphologically annotated corpus and providing a simple and easy-to-use application for verifying them. For the extraction task, three Czech parsers have been accommodated and evaluated. Finally we discuss the currently achieved results in the context of ongoing work and show that they lead to consistent and reliable results.

Keywords: noun phrases, parsing, parser evaluation, annotation, inter-annotator agreement

1 Introduction

Language resources for natural language processing are very important for development as well as improvement of existing natural language processing (NLP) tools. Situation for different European languages ranges from very good to almost non-existent. In the worst case there are almost no resources and we face the problem of creating them in a cheap and fast way while providing high quality results. An important aspect of building language resources is that one should always keep in mind the usability of the resource for particular applications. This approach, known as *application-driven development*, can already in early development stages prevent major design flaws which might not be automatically recoverable later on and could limit the usefulness of the resulting work.

In this approach, we present a language resource – an annotated corpus – as an intermediate step for improving many NLP applications (some of them even before full completion of the corpus). This kind of development helps us to create data which does not necessarily attempt to cover all nuances of natural language but strictly concentrates on particular tasks we want to solve.

Czech language is one of the most described European languages and there are several NLP teams that compete to build optimal language resources and provide cutting-edge applications on top of them. Our main goal was to create a new language resource based partially on existing data. We chose to use morphologically annotated corpus DESAM (Pala et al., 1997), that was manually checked by annotators. On top of this corpus we want to build a multi-layer annotation.

2 The Annotation Project

2.1 General Overview

The project aims to provide a multi-purpose annotated corpus for enhancing several NLP applications. In the first phase we want to perform computer aided syntactic annotation of noun phrases (that will be described further on). As the next step those noun phrases will be matched to the complex valency frames available in the Czech verb valency lexicon VerbaLex (Horák and Pala, 2007). Among other benefits, valency frames in this lexicon contain semantic information in the form of so called *semantic roles* as present in the Czech WordNet (Pala and Smrž, 2004). This can give us a mapping from noun phrases to semantic roles which can further serve as a base for a shallow corpus-based ontology. Such annotated resource can be further used in learning and testing other NLP applications, such as anaphora resolution, word sense disambiguation and others.

As already mentioned, the whole process of the annotation is computer-aided, i. e. with extensive exploitation of available NLP tools and resources for Czech. Thus, as side effects of the annotation, we will gain improvements of these tools and resources by investigating differences between their output and the annotation and fixing errors on the right places.

2.2 Annotation Principles

Annotation of language resources is a task which needs to be prepared precisely otherwise we end up with data that does not meet our quality standard. Our previous experience with annotators (mostly students from language department) gives us some hope that they can be trained to do simple tasks. Most of the errors they previously made were not caused by missing linguistic insight or knowledge but rather by lack of strict standards that would guarantee high inter-annotator agreement. Such strict standards are unfortunately very difficult to set for complex tasks like syntactic annotation. We can offer only a limited set of examples but there are always a number of cases when annotators are not able to give convincing conclusions.

We assume that an annotation standard is usually an attempt to approximate several mutually exclusive and contradictory constraints:

1. **completeness**: the annotation should provide complete linguistic insight into the particular area;
2. **consistency**: the annotation should be consistent, i. e. same or similar language phenomena should be handled in same or similar ways;
3. **usability**: the annotation should enable straightforward usage in the intended applications;
4. **simplicity**: the annotation should be as simple as possible to make high inter-annotator agreement achievable.

In our experience most language resources try to find a trade-off among the constraints by prioritizing them in the order given above. They prefer completeness over consistency, and both of them over simplicity.

Following the YAGNI[1], KISS[2] and “worse is better” (Gabriel, 1991) principles, we are strongly convinced that the reverse order of those constraints represents a much better priority list to be met when building a language resource. Thus, our priorities are:

- **simplicity**: so that annotators do not err too often;
- **usability**: so that the usage of the resource will be straightforward;
- **consistency**: following from simplicity;
- **completeness**: just in case everything is simple, usable and consistent.

The decision was made to create a work-flow for resource building that will conform to this “worse is better” concept. Annotation process was divided into several consequential phases where we obtain usable data immediately after each phase. The division into several parts of monotonous work has to be handled with annotators carefully. Our solution is to change their annotation task every week if possible to minimize negative effects of performing repetitive monotonous work (see (Fisherl, 1993) for details on this topic). This way their monthly work can be distributed e. g. between discovering noun phrases and their valency mapping.

[1] http://en.wikipedia.org/wiki/You_ain't_gonna_need_it

[2] http://en.wikipedia.org/wiki/KISS_principle

Also, creativity is something what is not expected in this project (according to *simplicity*). We wish to constrain the annotators as much as possible with a simple annotation scheme since limiting creativity increases inter-annotator agreement and (therefore) also *consistency*.

2.3 Noun Phrases Annotation

The first phase of annotation is to identify noun phrases in corpora. We soon realized that we have to distinguish between two basic types of noun phrases. Their precise description can be found in the next section; now a definition by example is enough for us. We will borrow one of the famous examples – 'I saw a man with a telescope'. There are two short (minimal) noun phrases 'a man' and 'a telescope' and we do not attempt to distinguish their relations. We could find a maximal noun phrase ('a man with a telescope') but maximal noun phrases are determined semantically, the inter-annotator agreement for them was too low and expert was contacted too often. Minimal noun phrases can be described more precisely and agreement between annotators is much higher.

Identifying a minimal noun phrase is still a task which brings in too much creativity. To speed up the process we decided to change it from manual to computer aided. Annotators will identify only those noun phrases which will be found by one of the used syntactic analyser. For each noun phrase only one solution can be chosen (see related screenshot in Figure 1). Possible answers are 'correct', 'incorrect' and 'not sure with answer'. Additionally, sentence can be tagged as 'complete' when all correct noun phrases are tagged. 'Complete' sentence can contain numerous of incorrect noun phrases. As we mentioned only noun phrases found by syntactic analysers are taken into account, otherwise the annotation process would be too slow and costly. The disadvantage of this approach will be compensated by manual post-processing of problematic parts only (the recall of the parsers on detecting noun phrases was not measured so far but it is estimated to be over 90 percent).

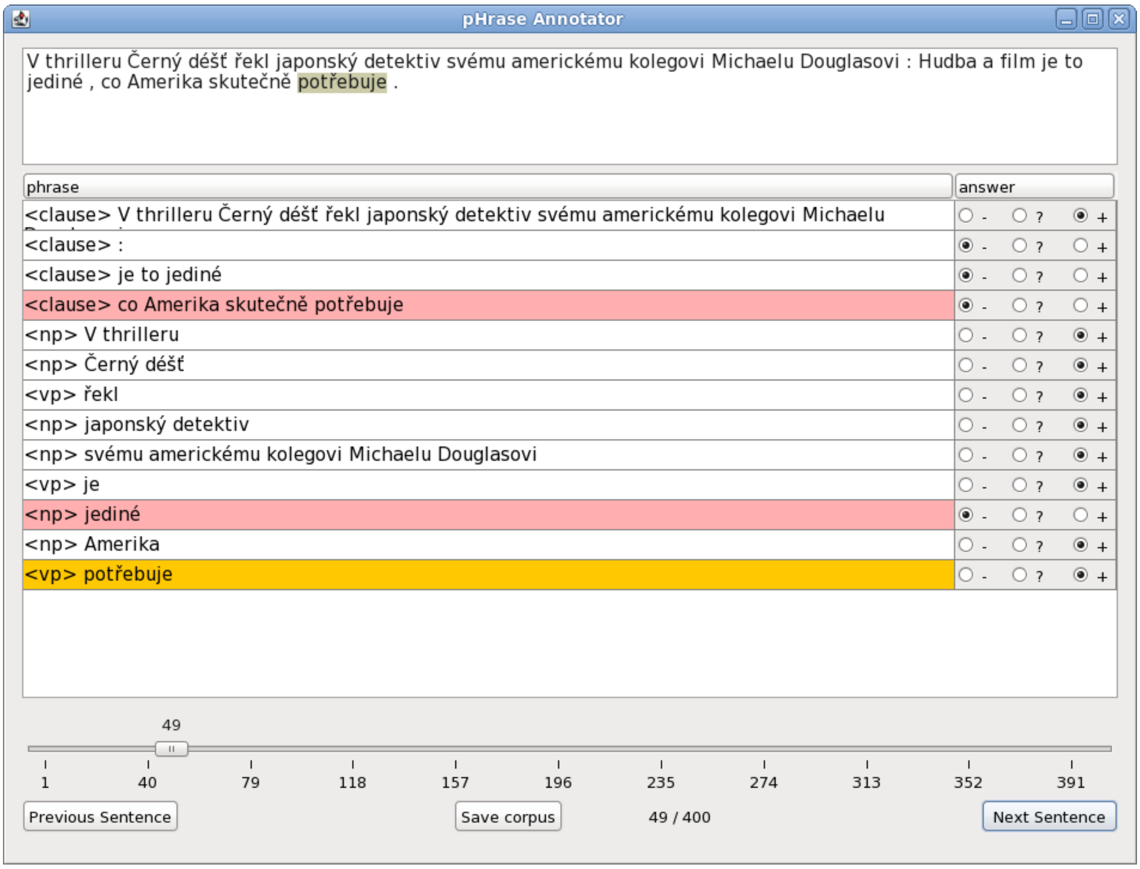
The described process makes annotators very effective and they can tag usually around one hundred sentences per hour. This means we can identify noun phrases in corpora containing one million tokens (e. g. DESAM (Pala et al., 1997), MAK (Jazykovedný ústav L'. Štúra SAV, 2009)) in 3 man-months (without cross-validation which can multiply that amount of time).

Technical implementation of corpora and tools is based on the NITE NXT (Soria et al., 2002) toolkit developed by several universities. This framework is originally developed for use with multimodal corpora but can be also very usable in case we work with various levels of annotation and their relations. On the basis of this framework we have created a set of utilities for import and export of our native formats and set of GUIs for annotators. Thanks to the object oriented approach and the XML format it is very easy to support other import or export formats varying across applications.

3 On Parsing Evaluation

Parsing evaluation is a very actual problem in the NLP field. A common way to estimate a parser quality is comparing its output (usually in form of syntactic trees) to a gold standard data available in a syntactically annotated corpus (typically treebank). This can lead to two problems:

- Missing resources: there can be no quality treebank for the language or formalism in use so that there is no direct way to estimate the parser quality.
- Erroneous resources: the process of syntactic annotation is usually very complex and requires extensive annotation instructions (see e. g. (Hajič et al., 2005)). It is very difficult for the annotators to sustain attention to hundreds of the annotation rules which leads to errors in the annotation.
- Insufficient evaluation metrics: we consider mostly used evaluations metrics (Parseval (Harrison et al., 1991), Leaf-Ancestor Assessment (Sampson, 2000), dependency precision) to be



V thrilleru Černý déšť řekl japonský detektiv svému americkému kolegovi Michaelu Douglasovi : Hudba a film je to jediné , co Amerika skutečně potřebuje .

phrase	answer
<clause> V thrilleru Černý déšť řekl japonský detektiv svému americkému kolegovi Michaelu	○ - ○ ? ● +
<clause> :	● - ○ ? ○ +
<clause> je to jediné	● - ○ ? ○ +
<clause> co Amerika skutečně potřebuje	● - ○ ? ○ +
<np> V thrilleru	○ - ○ ? ● +
<np> Černý déšť	○ - ○ ? ● +
<vp> řekl	○ - ○ ? ● +
<np> japonský detektiv	○ - ○ ? ● +
<np> svému americkému kolegovi Michaelu Douglasovi	○ - ○ ? ● +
<vp> je	○ - ○ ? ● +
<np> jediné	● - ○ ? ○ +
<np> Amerika	○ - ○ ? ● +
<vp> potřebuje	○ - ○ ? ● +

Figure 1: Annotation application screenshot

generally insufficient since they do not address usefulness for practical applications – while it is questionable whether it is possible at all to create a better one.

Another way to parsing evaluation are several application-driven evaluation techniques that have been recently proposed (e. g. (Miyao et al., 2009)). In this case, we measure how parsing can improve results of practically defined problems, such as information extraction (e. g. mentioned above), morphological disambiguation (Jakubíček et al., 2009), punctuation correction (Jakubíček and Horák, 2010) etc. This approach eliminates the mentioned problems of treebanks but such measures are typically not general enough (e. g. a parser that performs morphological disambiguation very well may not be suitable for information extraction). However, combining more such application-driven tests, which is our direction, could give us more general results that are suitable for comparison of different parsers with different internal formalisms, even more general and predicative than results of comparing output trees to the gold standard ones in the treebanks.

Within the context of the project described above we have proposed a new concept for parser evaluation and comparison based on measuring precision of noun and prepositional phrases (further referred just as "N/P phrases") detection. The measuring can be done in both automatic (in case we have a quality source of N/P phrases) and manual way. We argue that measuring noun phrases is more predicative and robust than measuring similarity of parsing trees because

- the N/P phrases structure is one of the basic and most important layers in the natural language syntax and it is of high importance for practical applications to know which parsers are able to solve this problem well;
- they can be very precisely defined and therefore there is no need for an extensive manual for deciding their correctness. This should guarantee minimal number of errors in annotation and high inter-annotator agreement (comparing e. g. to treebank annotation).

4 Annotation and Evaluation

4.1 N/P Phrases

As mentioned previously, we distinguish between maximal (long) and minimal (short) phrases. N/P phrase is defined as a constituent in a correct parsing tree that has noun or preposition as its head. Only the top such constituents are considered, i. e. if there is an N/P phrase containing another N/P phrase, only the former one is a maximal phrase. The semantics of such phrases is usually complement (argument, valency) or adjunct (modifier) of the verb.

To eliminate the frequent ambiguity of PP-attachment that often humans are not able to decide (e. g. as it can be found in the mentioned sentence 'I saw a man with a telescope'), we decided to split the maximal phrases on prepositions and particles. Also, the relative clauses that can be part of an N/P phrase are stripped out and we process them separately. By this procedure, we obtain the partial phrases that we use in the evaluation and annotation process.

This definition of N/P phrase follows our idea of making annotators' decisions as straightforward and unambiguous as possible which improves both precision and efficiency of the annotation.

4.2 Involved Parsers

The N/P phrases were drafted from three different parsers for Czech that we have had access to. In the following, we will shortly introduce them.

The *synt parser* (Horák et al., 2007) has been developed at the NLP Centre at the Faculty of Informatics, Masaryk University in Brno (FI MU). It is based on a large context-free grammar with contextual constraints with a head-driven chart parser as the backbone algorithm. For N/P phrases extraction we exploited its phrase-detection algorithm (Jakubíček et al., 2009).

Parser	Testing set	Precision
PDT manual annotation	PDT d-test	97.5 %
Collins	PDT d-test	93.1 %
Synt	DESAM	72.4 %
SET	DESAM	85.7 %

Table 1: Results of the annotation – precision on noun phrases

# of sentences	IAA	Cohen's κ	κ_{MAX}
1,200	0.86 %	0.61	0.85

Table 2: Inter-annotator agreement evaluation

The *SET parser* (Kovář et al., 2009) comes from the NLP Centre at FI MU as well. This parser is based on pattern matching linking rules and performs output in the form of phrases as well as in several other formats including dependency and constituent trees.

The *Collins' parser adapted for PDT* (Hajič et al., 1999) is a representative of dependency parsers that are in development at the Institute of Formal and Applied Linguistics in Prague. It uses the Prague Dependency Treebank (PDT, (Hajič, 2004)) as the data for training and testing and it is limited to its format, i. e. the only output it can provide is in the form of dependency trees. For this reason we have implemented an algorithm for extracting N/P phrases from dependency trees that is based on identifying constituents in dependency trees according to morphological categories of their heads and pruning them to obtain minimal noun phrases.

4.3 Data Sets

As mentioned above, our primary source of data is the DESAM corpus (Pala et al., 1997). This corpus has manual morphological annotation and its size is around one million tokens. Our aim is to extend the markup of this corpus by annotation of noun phrases and their semantic classes.

For the purpose of parser comparison and evaluation, we decided to involve the PDT into the process. This way we can estimate the relation between precision on our N/P phrases and the dependency precision. Also, by manual annotation we can discover serious errors in the PDT annotation. For these purposes, we have used 300 sentences randomly selected from the *d-test* part of the corpus.

4.4 Results

In Table 1, we can see the results of the annotation procedure and in Table 2 the inter-annotator agreement evaluation (given as plain agreement and Cohen's κ coefficient (Cohen, 1960)). In all cases, the manual morphological tagging was used as the input for the parsers, no automatic disambiguation took place. For both data sets, the N/P phrases from 300 randomly selected sentences were extracted and annotated for their correctness. The percentage of correct N/P phrases was computed to obtain the precision value as present in Table 1. By this computer-aided annotation, we are not able to find out the recall values until we have complete N/P phrases annotation.

The results show that the PDT annotation is quite precise with regard to minimal phrases. The Collins' parser wins in phrases detection precision – although on different data set, which is due to different morphological tagging, we believe that the results are more or less comparable. The disadvantage of the Collins' parser is that it cannot be further improved, because of its statistical nature, according to differences between annotation and its outputs. The two other parsers are rule-based and their rules can be easily fixed so that there is a very good chance that their results will be better in the future.

Finally, we can observe that the phrase detection precision is typically higher than the dependency precision of the parsers (Kovář et al., 2009; Holan and Žabokrtský, 2006). In other words (and according to our expectations), the parsers perform better in phrases detection than in building dependency trees.

5 Conclusions

In the paper, we have presented a concept of application-driven development of a multi-purpose linguistic resource. We have described the first phase of the project which consists of computer-aided annotation of minimal noun and prepositional phrases and explained how this annotation can be exploited in parser evaluation and comparison. We also displayed some preliminary results of the annotation with regard to parsers accuracy.

In the future, the project will continue with annotation of a larger amount of phrases and matching them to the verb valency frames and their semantic roles. Along with that, our linguistic tools and resources will be improved by fixing errors discovered in the process of annotation.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and by the Czech Science Foundation under the project P401/10/0792. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nr. 248307.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement*, volume 20, pages 37–46.

Fisherl, C. (1993). Boredom at work: A neglected concept. *Human Relations*, 46(3):395.

Gabriel, R. P. (1991). Lisp: Good news, bad news, how to win big. *AI Expert*, 6:30–39.

Hajič, J. (2004). Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Hajič, J., Collins, M., Ramshaw, L., and Tillmann, C. (1999). A Statistical Parser for Czech. In *Proceedings ACL'99*, Maryland, USA.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Štěpánek, J., Pajas, P., and Kárník, J. (2005). Anotace na analytické rovině – Návod pro anotátory. `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer`.

Harrison, P., Abney, S., Black, E., Flickinger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, R., Marcus, M., Santorini, B., and Strzalkowski, T. (1991). Evaluating syntax performance of parser/grammars of English. In Neal, J. G. and Walter, S. M., editors, *Natural Language Processing Systems Evaluation Workshop: Final Technical Report RL-TR-91-362*, pages 71–77, Griffiss Air Force Base, NY. Rome Laboratory.

Holan, T. and Žabokrtský, Z. (2006). Combining Czech Dependency Parsers. In *Lecture Notes in Artificial Intelligence, Proceedings of TSD 2006*, pages 95–102, Brno, Czech Republic. Springer Verlag.

Horák, A., Holan, T., Kadlec, V., and Kovář, V. (2007). Dependency and Phrasal Parsers of the Czech Language: A Comparison. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2007*, pages 76–84, Plzeň, Czech Republic. Springer-Verlag.

Horák, A. and Pala, K. (2007). Building a large lexicon of complex valency frames. In *Proceedings of the FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages*, pages 31–38, Lund University, Sweden. Tartu, Estonia.

Jakubíček, M. and Horák, A. (2010). Punctuation detection with full syntactic parsing. In *Research in Computing Science, Special Issue: Natural Language Processing and its Applications*, pages 335–343, Mexico.

Jakubíček, M., Horák, A., and Kovář, V. (2009). Mining phrases from syntactic analysis. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue 2009*, pages 124–130, Plzeň, Czech Republic. Springer-Verlag.

Jazykovedný ústav L'. Štúra SAV (2009). Slovenský národný korpus – r-mak-3.0.

Kovář, V., Horák, A., and Jakubíček, M. (2009). Syntactic analysis as pattern matching: The SET parsing system. In *Proceedings of the 4th Language & Technology Conference*, pages 100–104, Poznañ, Poland.

Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T., and Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 25(3):394–400.

Pala, K., Rychlý, P., and Smrž, P. (1997). DESAM — annotated corpus for Czech. In *Proceedings of SOFSEM'97*, pages 523–530. Springer-Verlag. Lecture Notes in Computer Science 1338.

Pala, K. and Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(2-3):79–88.

Sampson, G. (2000). A Proposal for Improving the Measurement of Parse Accuracy. *International Journal of Corpus Linguistics*, 5(01):53–68.

Soria, C., Bernsen, N. O., Cadee, N., Carletta, J., Dybkjaer, L., Evert, S., Heid, U., Isard, A., Kolodnytsky, M., Lauer, C.and Lezius, W., Noldus, L., Pirrelli, V., and Reithinger, N. (2002). Advanced tools for the study of natural interactivity. In *Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.