Automated Extraction of Lexico-Syntactic Information

Ondřej Bojar

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Abstract.

This article explains the need of high quality large lexicons describing syntactic properties of a natural language in order to accomplish various natural language processing tasks. Data sources useful to collect such lexicons for Czech are described and it is documented that the available treebank data are not a sufficient source. An approach to building such lexicons with the help of an automated procedure and non-treebank data sources is outlined.

Introduction

All applications of natural language processing (NLP) require high quality data describing the language. Currently, the need is most notably observed in the phase of syntactic analysis of natural languages. The applications vary from grammar and style checking over machine translation to question answering and automatic inferring. However, the syntactic information of individual lexical items (lexico-syntactic information hereafter) is expected to be helpful for lower lever tasks, such as language modelling for speech recognition, too.

Syntactic structure of natural languages is an extremely complicated phenomenon. Having adopted the framework of dependency syntax in general (cf. Mel'čuk [1988]) and more specifically the Functional Generative Description for Czech (FGD, Sgall et al. [1986], Hajičová et al. [1998]), a syntactic analysis of a sentence is a tree with nodes corresponding to the input words[1]. Given a string of n words forming a sentence of a natural language, there are exponentially many ($O(n^n)$) structural configurations possible. Ribarov (thesis in preparation) reports an extremely low ratio between the number of structures actually seen in PDT (see below) and the number of theoretically possible structures, a sentence of 20 words has 15.6 billion configurations possible and only a few hundred structures were observed. In other words, out of the space of all possible trees, extremely few ones are actually used. Anyway, as Bojar [2004] reports, having employed certain linguistically adequate local constraints that have to be satisfied in a correct analysis, there are still exponentially many analyses available.[2]

This negative result is by no means restricted to a specific linguistic theory or formalism. All frameworks and all languages suffer the problem of too many analyses available. Statistical approaches to NLP (see section Parsers of Czech) do not explain this observation at all, they circumvent the problem by providing most common (most frequent) analyses and therefore commonly seem to perform well.

Native speakers do not seem to get in this exponential trouble. They are usually capable of choosing one single analysis (given the full context information) or of selecting a very limited number of plausible analyses, i.e. analyses that will become acceptable given some necessary contextual support. Ignoring the massive parallelism of human processing, there are undoubtedly more mutual restrictions on what is acceptable in a more or less local context than our theories, formalisms or lexicons capture so far.

Many of the restrictions come from individual words: for example, it is possible to attach a noun in nominative under a finite verb in general (as long as there is no subject present yet), but certain verbs do not allow such a modification. An automatic syntactic analyser must be equipped with lexical information in order to allow such modifications for some verbs and reject for others. Another example is given in Figure 1. The sentences 1 and 2 differ in the lexical value of the last word only. However, this difference is important enough for a native speaker to unambiguously select only the structures as indicated in Figure 1. An uninformed syntactic analyser would have to allow both structures for each of the sentences.[3]

[1] There are two separate levels defined: an auxiliary level of surface syntax where input words correspond one to one to the nodes and a deep syntactic level (called TECTOGRAMMATICAL in FGD) where only autosemantic words are present and nodes for deleted words are artificially added.

[2] A grammar induced from 5,000 training sentences assigns approximately 9 different possible heads to a node on average which leads to $O(9^n)$ analyses of a sentence of n words.

[3] Formally defined (see Panevová [1980]), the difference lies in VALENCY REQUIREMENTS of the verb *přicházet*. In the

Another area of NLP where current approaches suffer from lack of lexico-syntactic information is machine translation (MT). Available translation dictionaries (if machine-readable at all) were designed for humans and a lot of necessary information is not explicitly stated in them.

The process of building precise lexicons covering many words is very demanding.[4] Eventually, this process has to be carried out by human lexicographers, because certain distinctions such as OBLIGATORITY vs. OPTIONALITY of verbal modifications (cf. Panevová [1980]) cannot be resolved automatically. There are also specific situations (see below) where human annotation is bearable and actually easier than developing an automatic procedure. However, a lot of effort can be saved by employing an automatic preprocessor to provide lexicographers with lower amount of more relevant preprocessed data.

The following sections describe data sources and tools available to build or augment syntactic lexicons and sketch some novel algorithms of the preprocessing aimed at the two indicated tasks: 1. extending lexicons of valency information and 2. providing machine-readable translation dictionaries with syntactic information and other hints for MT systems.

Data Sources

Czech National Corpus (CNC)

Czech National Corpus (CNC, Kocek et al. [2000]) is a balanced collection of contemporary written Czech. The release labelled SYN2000 contains 100 million tokens in 1.76 million sentences. The data are automatically morphologically tagged, but no syntactic information is provided.

Prague Dependency Treebank (PDT)

Prague Dependency Treebank (PDT, Böhmová et al. [2001]) contains 1.5 million tokens in 98,000 sentences from selected texts of CNC. The annotation comprises manual morphological tags, manual surface dependency (analytic) trees and manual deep dependency (tectogrammatical) trees.

Thanks to the syntactic annotation, PDT is an excellent source of syntactic information about Czech. However, as shown in the following table, the amount of data is by no means sufficient for collecting information on individual lexical items:

After having observed	20,000	75,000	training sentences
a new lemma (i.e. word) comes every	1.6	1.8	test sentences
a new full morphological tag comes every	110	290	test sentences
a new simplified tag[5]comes every	280	870	test sentences

Another example of insufficiency of PDT is described in Bojar [2003]: There are 22,276 different Czech verbs covered in CNC. PDT covers only 5,407 of these verbs and only for a few hundreds of verbs, PDT contains more than 50 occurrences per verb.

sentence 2, the prepositional phrase *na chuť* indicates an idiomatic use of the verb. In this situation, the verb requires an ADDRESSEE and the word *hypermarketu* is the only candidate. No such requirement is present in the sentence 1.

[4] Finished lexicons, such as Skoumalová [2001] miss many of verbal frames, other lexicons such as VALLEX (Straňáková-Lopatková and Žabokrtský [2002]) are still under development and new entries are added at a relatively slow pace.

[5] The simplified tag comprises only POS, SUBPOS, CASE, NUMBER and GENDER information.

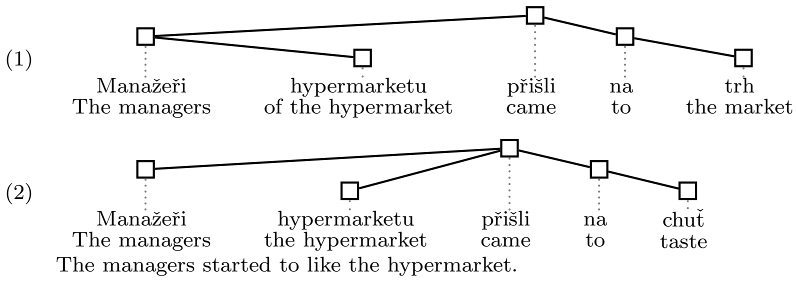


Figure 1. Lexical information is necessary to choose the correct surface syntactic structure.

Prague Czech-English Dependency Treebank (PCEDT)

PCEDT[6] is a new data source for Czech-English MT systems. The core part of PCEDT 1.0 is a Czech translation of 21,600 English sentences from the Wall Street Journal section of Penn Treebank 3 corpus (PTB, Marcus et al. [1993], LDC99T42). Sentences of the Czech translation were automatically morphologically annotated and parsed into two levels (analytical and tectogrammatical) of dependency structures, closely related to the PDT annotation scheme. The original English sentences were transformed from the Penn Treebank phrase-structure trees into dependency representations, too.

PCEDT 1.0 also comprises a parallel Czech-English corpus of plain text from Reader's Digest 1993–1996 consisting of 53,000 parallel sentences.

Machine-Readable Czech-English Translation Dictionaries

There are several machine-readable Czech-English translation dictionaries (MRDs) publicly accessible containing around two hundred thousand translation pairs. At least one of the dictionaries is available under the free GNU/FDL license. This dictionary is still under development and a recent release consisting of 115,929 translation pairs can be found on the PCEDT 1.0 CD.

However, all the MRDs were originally prepared for humans and a lot of information crucial for automatic systems is missing. Typically, the entries are represented as plain text strings (e.g. *accession to = vstoupení do* or *adjudge sb. to be guilty = uznat vinným koho*) where part of the string is the entry itself and part is some relevant syntactic information (*sb.* or *koho*). Especially in older entries, no attempt was made at an explicit distinction what is the entry and what is the additional information.

Parsers of Czech

There are three statistical parsers of Czech. See Bojar [2003] for an evaluation of parsers by Collins et al. [1999] and Zeman [2002]. A novel parser by Charniak [2000] was adapted for Czech too, but not evaluated yet. The accuracy of parsers by Zeman and Collins ranges from 70 to 83% of correctly assigned dependencies, but this corresponds to 15 to 31% correct sentences only. Clearly, the utility of these parsers with respect to the task of extracting precise lexico-syntactic information is still limited.[7]

Moreover, the parsers by Collins and Charniak are theoretically inadequate, because they are in principle incapable of producing non-projective analyses. As Holan [2003] and others report, nearly 25% of sentences in PDT are non-projective.

Methods of Automated Extraction of Lexico-Syntactic Information

Extending Monolingual Syntactic Lexicons

As documented above, available treebank data for Czech are not a sufficient source for syntactic information about individual words and the current parsers still suffer quite a high level of error rate. My approach was first outlined in Bojar [2002] and tries to solve the problem of finding verb valency frames by "picking nice examples".

Instead of a careful and thorough analysis of every tree in a treebank (because of the small amount of data in treebanks, the information in every single tree must be treated as priceless), I employ a linguistically motivated filtering technique on plain text (with morphological annotation) data to select only sentences that are easy to parse but still contain the relevant information.

Having selected nice examples, the accuracy of available parsers rises by 5 to 10 % and the output verbal modifications are more precise. (For a full report see Bojar [2003]). A similar approach can be used for valency information of other parts of speech, if the filtering is slightly modified.

Providing Translation Dictionaries with Syntactic Information

As described above, available Czech-English translation dictionaries are too shallow and usually contain pairs of word forms only. In any approach to MT, syntactic information should be utilized whenever possible in order to arrive at a better translation. See Čmejrek et al. [2003] for an example of a Czech to English MT system based on deep syntactic analysis of Czech but currently limited to word-to-word translations only due to the lack of a better dictionary.

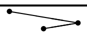
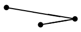
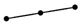
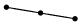
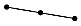
[6] http://ufal.ms.mff.cuni.cz/new/mt/PCEDT_0.95/

[7] For instance, Bojar [2003] reports that Collins assigns immediate dependents only in 55% occurrences of verbs, in other words only 55% of verb frames are observed correctly.

I propose several distinct steps to perform with source dictionaries in order to provide them with syntactic and other information useful for MT:

Morphological analysis. Very few entries in available MRDs contain any indication on part of speech and other morphological categories. An excellent morphological analyser for Czech is provided by Hajič [2001], but the disambiguation must be carried out. Automatic taggers[8] are not suited well for the task, because there is no context information in word entries. The easiest way to disambiguate the part of speech information is a semi-automatic process: single word entries are disambiguated automatically, if their form equals to exactly one of the lemmas available in the morphological analyses. Multi-word entries have to be disambiguated manually, in groups possessing the same POS ambiguity. In many situations, there is no hint available in the data itself and common knowledge has to be used. Examples are given in Figure 2.

Adding syntactic structure. Currently, there is no Czech MRD equipped with syntactic information of multi-word entries. However, the internal syntactic structure of multi-word entries is relevant for selecting the most appropriate translation, especially if there are some extra modifications present in (the middle of) the multi-word expression. The syntactic structure determines what additional modifications are allowed where. The following table illustrates different syntactic structures of Czech lexical entries consisting of a noun, an adjective and a noun.

Noun Adjective Noun	Syntactic Structure	English Translation
Komise Evropské unie		Commission of the European Community
náhrada způsobené škody		dilapidation
látkou potažené sedadlo		fabric-covered seat
poruchy způsobené přijímačem		set noise
nevolnost způsobená pohybem		kinesia

There are three options when adding the syntactic structure to multi-word expressions. First, the structure might be added manually, possibly making use of detailed morphological analysis including case, number and gender. (Again, handling all the expressions with the same morphological pattern at once will speed up the manual annotation.) Second, the structure might be searched for in PDT. With respect to the size of PDT, only the most common expressions are expected to come up. Third, automatically parsed sentences might be used to find the structure. In order to reduce the error, selected simple sentences (see above) should be used. Depending on the number and complexity of word entries, different approaches should be chosen. Manual annotation is always less error prone and preferable but not always plausible.

Adding agreement constraints. Based on the internal syntactic structure of multi-word entries, some elements in the multi-word expression have to share certain morphological properties (e.g. noun-adjective agreement in case, number and gender). When generating Czech text, using the multi-word entry in a specific syntactic construction imposes some morphological requirements not just on the head word but on other words in the multi-word expression, too. Similarly, if the required

[8] A (morphological) tagger selects single morphological tag for every word in a sentence based on the words in close neighbourhood. Cf. Hajič and Hladká [1998].

Noun and Noun/Adjective	Correct Interpretation	English Translation
husa divoká	Noun Adjective	grey goose
kniha účetní†	Noun Adjective	account book
napětí dovolené†	Noun Adjective	permissible stress
chyba měření	Noun Noun	measurement error
plán prací†	Noun Noun	schedule of operation
rozsah měření	Noun Noun	range of measurement
Numeral/Verb and Noun	**Correct Interpretation**	**English Translation**
tři prdele†	Numeral Noun	shitloads
pět švestek	Numeral Noun	one's duds*
pět chválu	Verb Noun	sing someone's praises

† These expressions allow for another interpretation, too, mostly kind of funny.
* Part of the idiom *pick up one's duds*.

Figure 2. Examples of morphological ambiguity in translation dictionaries.

agreement is not met in a Czech text, the sequence of the words is not an occurrence of the multi-word entry but a rather random collocation.

Once the syntactic structure is present in the lexicon, the agreement constraints can be given to nearly all the entries sharing the same syntactic structure. One could also think of an algorithm to provide agreement constraints without actually needing the syntactic annotation: searching for the collocations of lemmatized lexical entries in a lemmatized corpus and extracting the most common observed agreement requirements. Which of this two options is easier to follow depends on the number of lexical entries to annotate, on their complexity and on the difficulty of adding syntactic information itself.

Providing entries with examples. The utility of a dictionary would clearly rise, if the dictionary provided several examples for every lexical entry. Such lexicon would be easier to use not only by humans, but automatic (e.g. statistical) methods of MT could be trained on the examples, too. Moreover, searching for examples of lexical entries is the first step when adding frequency information to the entries. Frequency information is vital for every statistical method of NLP.

The better the syntactic annotation of the entries is, the fewer false examples are found by an automatic procedure (both when searching in a treebank or when searching in a plain corpus). Syntactic structure and/or agreement constraints of the entries allow the procedure to reject random collocations. Making use of the syntactic information, more examples might be found: the syntactic structure of lexical entries indicates, where and what extra modifications are possible. One can add some sort of wildcards in such places and search for “discontinuous” examples of the multi-word expressions, too.

Adding monolingual frequency information. Once the algorithm for searching examples in one of the languages is ready, monolingual frequency information can be added to the entries. Entries with no observation at all are likely to be mistyped or otherwise wrong. High-frequency entries should be promoted by an MT system, whenever possible.

Clearly, due to the size of PDT, frequency information for Czech entries must be searched for in a larger corpus such as CNC and a similar requirement holds for English.

Adding parallel frequency information. Parallel frequencies, i.e. frequencies of observation of the whole translation pairs, are yet better source for an MT system. The size of available parallel Czech-English corpora is still rather limited, however the Reader’s Digest corpus promises at least some meaningful results.

Matching entries with data from deep syntactic lexicons. A completely different task is to match entries from plain translation dictionaries to entries in deep syntactic lexicons. Both for Czech and English, deep syntactic lexicons are currently under development. VALLEX 1.0 (Straňáková-Lopatková and Žabokrtský [2002]) contains entries for roughly 1400 Czech verbs. FrameNet[9] provides valency information for more than 5500 English verbs, nouns and adjectives. Although the annotation schemata of VALLEX and FrameNet are not equivalent[10], a close match is possible and interesting both from the theoretical and practical point of view. By intuition, deep syntactic structure of two languages should be much closer to each other than the surface structure representation, however there are examples where the deep structure still differs. From the theoretical point of view, a deep syntactic translation dictionary reveals the unspotted differences, and from the practical point of view, the transfer phase in an MT system should be much simpler if one transfers between deep syntactic representations.

The task of matching VALLEX and FrameNet entries would have to be carried out manually (with semi-automatic tools solving frequent situations) because there is no parallel corpus with Czech side annotated with VALLEX entries and the English side with FrameNet. However, a lot of effort can be saved if one makes use of the surface syntactic information already stored in the translation dictionary. The translation dictionary will serve as a bridging link between the two deep syntactic lexicons.

[9] `http://www.icsi.berkeley.edu/framenet/`, ?

[10] FrameNet uses an unlimited number of role labels (e.g. Agent, Goal, Purpose, Victim, . . .) specific for every semantic class of verbs while VALLEX uses a restricted set of more general labels (e.g. Actor, Patient).

Conclusion

I documented the complexity of syntactic analysis and the necessity of lexico-syntactic information. Treebank data was shown to be insufficient for the task of extracting the information in a large scale. Two novel methods were proposed utilising the available information: a method for extending syntactic lexicons with the help of nice examples and a method for providing translation dictionaries with syntactic information and other hints for MT systems.

Further Research

In further research, I wish to concentrate on implementing the outlined algorithms. Most of work on picking nice examples has been implemented in Bojar [2002] already, but only preliminary design decisions were realized for the task of providing translation dictionaries with syntactic information. Another goal is to actually collect the lexicographic data using the implemented algorithms.

An important part of my forthcoming effort shall be devoted to evaluation, too. Spending time on building lexicons is useless, if I cannot prove that the lexico-syntactic information indeed contributes to the precision of NLP tasks, syntactic analysis and machine translation in particular.

Acknowledgement. This work has been partially supported by the grant GAČR 405/03/0914.

References

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers, 2001.

Ondřej Bojar. Automatická extrakce lexikálně-syntaktických údajů z korpusu (Automatic extraction of lexico-syntactic information from corpora). Master's thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2002. In Czech.

Ondřej Bojar. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120, 2003. ISSN 0032-6585.

Ondřej Bojar. Czech Syntactic Analysis Constraint-Based, XDG: One Possible Start. *Prague Bulletin of Mathematical Linguistics*, 81:43–54, 2004. ISSN 0032-6585.

Eugene Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*, pages 132–139, Seattle, Washington, USA, April 2000.

Martin Čmejrek, Jan Cuřín, and Jiří Havelka. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April 2003. ISBN 1-932432-00-0.

Michael Collins, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann. A Statistical Parser of Czech. In *Proceedings of 37th ACL Conference*, pages 505–512, University of Maryland, College Park, USA, 1999.

Jan Hajič. *Disambiguation of Rich Inflection - Computational Morphology of Czech*, volume I. Prague Karolinum, Charles University Press, 2001. 334 pp.

Jan Hajič and Barbora Hladká. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pages 483–490, Montreal, Canada, 1998.

Eva Hajičová, Barbara Partee, and Petr Sgall. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer Academic Publishers, Amsterdam, Netherlands, 1998. ISBN 0-7923-5289-0.

Tomáš Holan. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003 2003. ISBN 80-86732-22-3.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Praha, 2000. ISBN 80-85899-94-9.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330, 1993.

Igor A. Mel'čuk. *Dependency Syntax - Theory and Practice*. Albany: State University of New York Press, 1988.

Jarmila Panevová. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence]*. Academia, Prague, Czech Republic, 1980.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Hana Skoumalová. *Czech syntactic lexicon*. PhD thesis, Univerzita Karlova, Filozofická fakulta, 2001.

Markéta Straňáková-Lopatková and Zdeněk Žabokrtský. Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, volume 3, pages 949–956. ELRA, 2002.

Daniel Zeman. Can Subcategorization Help a Statistical Parser? In *Proceedings of the 19th International Conference on Computational Linguistics (Coling 2002)*, Taibei, Tchaj-wan, 2002. Zhongyang Yanjiuyuan (Academia Sinica).