Dependency Parsing of Hungarian: Baseline Results and Challenges

Richárd Farkas[1], **Veronika Vincze**[2], **Helmut Schmid**[1]
[1]Institute for Natural Language Processing, University of Stuttgart
{farkas,schmid}@ims.uni-stuttgart.de
[2]Research Group on Artificial Intelligence, Hungarian Academy of Sciences
vinczev@inf.u-szeged.hu

Abstract

Hungarian is a stereotype of morphologically rich and non-configurational languages. Here, we introduce results on dependency parsing of Hungarian that employ a 80K, multi-domain, fully manually annotated corpus, the Szeged Dependency Treebank. We show that the results achieved by state-of-the-art data-driven parsers on Hungarian and English (which is at the other end of the configurational-non-configurational spectrum) are quite similar to each other in terms of attachment scores. We reveal the reasons for this and present a systematic and comparative linguistically motivated error analysis on both languages. This analysis highlights that addressing the language-specific phenomena is required for a further remarkable error reduction.

1 Introduction

From the viewpoint of syntactic parsing, the languages of the world are usually categorized according to their level of configurationality. At one end, there is English, a strongly configurational language while Hungarian is at the other end of the spectrum. It has very few fixed structures at the sentence level. Leaving aside the issue of the internal structure of NPs, most sentence-level syntactic information in Hungarian is conveyed by morphology, not by configuration (É. Kiss, 2002).

A large part of the methodology for syntactic parsing has been developed for English. However, parsing non-configurational and less configurational languages requires different techniques. In this study, we present results on Hungarian dependency parsing and we investigate this general issue in the case of English and Hungarian.

We employed three state-of-the-art data-driven parsers (Nivre et al., 2004; McDonald et al., 2005; Bohnet, 2010), which achieved (un)labeled attachment scores on Hungarian not so different from the corresponding English scores (and even higher on certain domains/subcorpora). Our investigations show that the feature representation used by the data-driven parsers is so rich that they can – without any modification – effectively learn a reasonable model for non-configurational languages as well.

We also conducted a systematic and comparative error analysis of the system's outputs for Hungarian and English. This analysis highlights the challenges of parsing Hungarian and suggests that the further improvement of parsers requires special handling of language-specific phenomena. We believe that some of our findings can be relevant for intermediate languages on the configurational-non-configurational spectrum.

2 Chief Characteristics of the Hungarian Morphosyntax

Hungarian is an **agglutinative language**, which means that a word can have hundreds of word forms due to inflectional or derivational affixation. A lot of grammatical information is encoded in morphology and Hungarian is a stereotype of morphologically rich languages. The Hungarian **word order is free** in the sense that the positions of the subject, the object and the verb are not fixed within the sentence, but word order is related to information structure, e.g. new (or emphatic) information (the focus) always precedes the verb

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 55–65,
Avignon, France, April 23 - 27 2012.

and old information (the topic) precedes the focus position. Thus, the position relative to the verb has no predictive force as regards the syntactic function of the given argument: while in English, the noun phrase before the verb is most typically the subject, in Hungarian, it is the focus of the sentence, which itself can be the subject, object or any other argument (É. Kiss, 2002).

The grammatical function of words is determined by **case suffixes** as in *gyerek* "child" – *gyereknek* (child-DAT) "for (a/the) child". Hungarian nouns can have about 20 cases[1] which mark the relationship between the head and its arguments and adjuncts. Although there are postpositions in Hungarian, case suffixes can also express relations that are expressed by prepositions in English.

Verbs are inflected for person and number and the definiteness of the object. Since conjugational information is sufficient to deduce the pronominal subject or object, they are typically omitted from the sentence: *Várlak* (wait-1SG2OBJ) "I am waiting for you". This pro-drop feature of Hungarian leads to the fact that there are several **clauses without an overt subject or object**.

Another peculiarity of Hungarian is that the third person singular present tense indicative form of the copula is phonologically empty, i.e. there are apparently verbless sentences in Hungarian: *A ház nagy* (the house big) "The house is big". However, in other tenses or moods, the copula is present as in *A ház nagy lesz* (the house big will.be) "The house will be big".

There are two **possessive constructions** in Hungarian. First, the possessive relation is only marked on the possessed noun (in contrast, it is marked only on the possessor in English): *a fiú kutyája* (the boy dog-POSS) "the boy's dog". Second, both the possessor and the possessed bear a possessive marker: *a fiúnak a kutyája* (the boy-DAT the dog-POSS) "the boy's dog". In the latter case, the possessor and the possessed may not be adjacent within the sentence as in *A fiúnak látta a kutyáját* (the boy-DAT see-PAST3SGOBJ the dog-POSS-ACC) "He saw the boy's dog", which results in a non-projective syntactic tree. Note that in the first case, the form of the possessor coincides with that of a nominative noun while in the second case, it coincides with a dative noun.

According to these facts, a Hungarian parser must rely much more on morphological analysis than e.g. an English one since in Hungarian it is morphemes that mostly encode morphosyntactic information. One of the consequences of this is that Hungarian sentences are shorter in terms of word numbers than English ones. Based on the word counts of the Hungarian–English parallel corpus Hunglish (Varga et al., 2005), an English sentence contains 20.5% more words than its Hungarian equivalent. These extra words in English are most frequently prepositions, pronominal subjects or objects, whose parent and dependency label are relatively easy to identify (compared to other word classes). This train of thought indicates that the cross-lingual comparison of final parser scores should be conducted very carefully.

3 Related work

We decided to focus on **dependency parsing** in this study as it is a superior framework for non-configurational languages. It has gained interest in natural language processing recently because the representation itself does not require the words inside of constituents to be consecutive and it naturally represent discontinuous constructions, which are frequent in languages where grammatical relations are often signaled by morphology instead of word order (McDonald and Nivre, 2011). The two main efficient approaches for dependency parsing are the graph-based and the transition-based parsers. The graph-based models look for the highest scoring directed spanning tree in the complete graph whose nodes are the words of the sentence in question. They solve the machine learning problem of finding the optimal scoring function of subgraphs (Eisner, 1996; McDonald et al., 2005). The transition-based approaches parse a sentence in a single left-to-right pass over the words. The next transition in these systems is predicted by a classifier that is based on history-related features (Kudo and Matsumoto, 2002; Nivre et al., 2004).

Although the available **treebanks for Hungarian** are relatively big (82K sentences) and fully manually annotated, the studies on parsing Hungarian are rather limited. The Szeged (Constituency) Treebank (Csendes et al., 2005) con-

[1] Hungarian grammars and morphological coding systems do not agree on the exact number of cases, some rare suffixes are treated as derivational suffixes in one grammar and as case suffixes in others; see e.g. Farkas et al. (2010).

sists of six domains – namely, short business news, newspaper, law, literature, compositions and informatics – and it is manually annotated for the possible alternatives of words' morphological analyses, the disambiguated analysis and constituency trees. We are aware of only two articles on phrase-structure parsers which were trained and evaluated on this corpus (Barta et al., 2005; Iván et al., 2007) and there are a few studies on hand-crafted parsers reporting results on small own corpora (Babarczy et al., 2005; Prószéky et al., 2004).

The Szeged Dependency Treebank (Vincze et al., 2010) was constructed by first automatically converting the phrase-structure trees into dependency trees, then each of them was manually investigated and corrected. We note that the dependency treebank contains more information than the constituency one as linguistic phenomena (like discontinuous structures) were not annotated in the former corpus, but were added to the dependency treebank. To the best of our knowledge no parser results have been published on this corpus. Both corpora are available at `www.inf.u-szeged.hu/rgai/SzegedTreebank`.

The multilingual track of the CoNLL-2007 Shared Task (Nivre et al., 2007) addressed also the task of dependency parsing of Hungarian. The Hungarian corpus used for the shared task consists of automatically converted dependency trees from the Szeged Constituency Treebank. Several issues of the automatic conversion tool were reconsidered before the manual annotation of the Szeged Dependency Treebank was launched and the annotation guidelines contained instructions related to linguistic phenomena which could not be converted from the constituency representation – for a detailed discussion, see Vincze et al. (2010). Hence the annotation schemata of the CoNLL-2007 Hungarian corpus and the Szeged Dependency Treebank are rather different and the final scores reported for the former are not directly comparable with our reported scores here (see Section 5).

4 The Szeged Dependency Treebank

We utilize the Szeged Dependency Treebank (Vincze et al., 2010) as the basis of our experiments for Hungarian dependency parsing. It contains 82,000 sentences, 1.2 million words and 250,000 punctuation marks from six domains. The annotation employs 16 coarse grained POS tags, 95 morphological feature values and 29 dependency labels. 19.6% of the sentences in the corpus contain non-projective edges and 1.8% of the edges are non-projective[2], which is almost 5 times more frequent than in English and is the same as the Czech non-projectivity level (Buchholz and Marsi, 2006). Here we discuss two annotation principles along with our modifications in the dataset for this study which strongly influence the parsers' accuracies.

Named Entities (NEs) were treated as one token in the Szeged Dependency Treebank. Assuming a perfect phrase recogniser on the whitespace tokenised input for them is quite unrealistic. Thus we decided to split them into tokens for this study. The new tokens automatically got a *proper noun with default morphological features* morphological analysis except for the last token – the head of the phrase –, which inherited the morphological analysis of the original multiword unit (which can contain various grammatical information). This resulted in an N N N N POS sequence for *Kovács és társa kft.* "Smith and Co. Ltd." which would be annotated as N C N N in the Penn Treebank. Moreover, we did not annotate any internal structure of Named Entities. We consider the last word of multiword named entities as the head because of morphological reasons (the last word of multiword units gets inflected in Hungarian) and all the previous elements are attached to the succeeding word, i.e. the penultimate word is attached to the last word, the antepenultimate word to the penultimate one etc. The reasons for these considerations are that we believe that there are no downstream applications which can exploit the information of the internal structures of Named Entities and we imagine a pipeline where a Named Entity Recogniser precedes the parsing step.

Empty copula: In the verbless clauses (predicative nouns or adjectives) the Szeged Dependency Treebank introduces virtual nodes (16,000 items in the corpus). This solution means that a similar tree structure is ascribed to the same sentence in the present third person singular and all the other tenses / persons. A further argument for the use of a virtual node is that the virtual node is always present at the syntactic level

[2]Using the transitive closure definition of Nivre and Nilsson (2005).

corpus		Malt		MST		Mate	
		ULA	LAS	ULA	LAS	ULA	LAS
Hungarian	dev	88.3 (89.9)	85.7 (87.9)	86.9 (88.5)	80.9 (82.9)	89.7 (91.1)	86.8 (89.0)
	test	88.7 (90.2)	86.1 (88.2)	87.5 (89.0)	81.6 (83.5)	90.1 (91.5)	87.2 (89.4)
English	dev	87.8 (89.1)	84.5 (86.1)	89.4 (91.2)	86.1 (87.7)	91.6 (92.7)	88.5 (90.0)
	test	88.8 (89.9)	86.2 (87.6)	90.7 (91.8)	87.7 (89.2)	92.6 (93.4)	90.3 (91.5)

Table 1: Results achieved by the three parsers on the (full) Hungarian (Szeged Dependency Treebank) and English (CoNLL-2009) datasets. The scores in brackets are achieved with gold-standard POS tagging.

since it is overt in all the other forms, tenses and moods of the verb. Still, the state-of-the-art dependency parsers cannot handle virtual nodes. For this study, we followed the solution of the Prague Dependency Treebank (Hajič et al., 2000) and virtual nodes were removed from the gold standard annotation and all of their dependents were attached to the head of the original virtual node and they were given a dedicated edge label (Exd).

Dataset splits: We formed training, development and test sets from the corpus where each set consists of texts from each of the domains. We paid attention to the issue that a document should not be separated into different datasets because it could result in a situation where a part of the test document was seen in the training dataset (which is unrealistic because of unknown words, style and frequently used grammatical structures). As the fiction subcorpus consists of three books and the law subcorpus consists of two rules, we took half of one of the documents for the test and development sets and used the other part(s) for training there. This principle was followed at our cross-fold-validation experiments as well except for the law subcorpus. We applied 3 folds for cross-validation for the fiction subcorpus, otherwise we used 10 folds (splitting at documentary boundaries would yield a training fold consisting of just 3000 sentences).[3]

5 Experiments

We carried out experiments using three state-of-the-art parsers on the Szeged Dependency Treebank (Vincze et al., 2010) and on the English datasets of the CoNLL-2009 Shared Task (Hajič et al., 2009).

Tools: We employed a finite state automata-based **morphological analyser** constructed from the morphdb.hu lexical resource (Trón et al., 2006) and we used the MSD-style morphological code system of the Szeged TreeBank (Alexin et al., 2003). The output of the morphological analyser is a set of possible lemma–morphological analysis pairs. This set of possible morphological analyses for a word form is then used as possible alternatives – instead of open and closed tag sets – in a standard sequential POS tagger. Here, we applied the Conditional Random Fields-based Stanford POS tagger (Toutanova et al., 2003) and carried out 5-fold-cross POS training/tagging inside the subcorpora.[4] For the English experiments we used the predicted POS tags provided for the CoNLL-2009 shared task (Hajič et al., 2009).

As the **dependency parser** we employed three state-of-the-art data-driven parsers, a transition-based parser (Malt) and two graph-based parsers (MST and Mate parsers). The Malt parser (Nivre et al., 2004) is a transition-based system, which uses an arc-eager system along with support vector machines to learn the scoring function for transitions and which uses greedy, deterministic one-best search at parsing time. As one of the graph-based parsers, we employed the MST parser (McDonald et al., 2005) with a second-order feature decoder. It uses an approximate exhaustive search for unlabeled parsing, then a separate arc label classifier is applied to label each arc. The Mate parser (Bohnet, 2010) is an efficient second order dependency parser that models the interaction between siblings as well as grandchildren (Carreras, 2007). Its decoder works on labeled edges, i.e. it uses a single-step approach for obtaining labeled dependency trees. Mate uses a rich and

[3]Both the training/development/test and the cross-validation splits are available at `www.inf.u-szeged.hu/rgai/SzegedTreebank`.

[4]The JAVA implementation of the morphological analyser and the slightly modified POS tagger along with trained models are available at `http://www.inf.u-szeged.hu/rgai/magyarlanc`.

corpus	#sent.	length	CPOS	DPOS	ULA	all ULA	LAS	all LAS
newspaper	9189	21.6	97.2	96.5	88.0 (90.0)	+0.8	84.7 (87.5)	+1.0
short business	8616	23.6	98.0	97.7	93.8 (94.8)	+0.3	91.9 (93.4)	+0.4
fiction	9279	12.6	96.9	95.8	87.7 (89.4)	-0.5	83.7 (86.2)	-0.3
law	8347	27.3	98.3	98.1	90.6 (90.7)	+0.2	88.9 (89.0)	+0.2
computer	8653	21.9	96.4	95.8	91.3 (92.8)	-1.2	88.9 (91.2)	-1.6
composition	22248	13.7	96.7	95.6	92.7 (93.9)	+0.3	88.9 (91.0)	+0.3

Table 2: Domain results achieved by the Mate parser in cross-validation settings. The scores in brackets are achieved with gold-standard POS tagging. The 'all' columns contain the added value of extending the training sets with each of the five out-domain subcorpora.

well-engineered feature set and it is enhanced by a Hash Kernel, which leads to higher accuracy.

Evaluation metrics: We apply the Labeled Attachment Score (LAS) and Unlabeled Attachment Score (ULA), taking into account punctuation as well for evaluating dependency parsers and the accuracy on the main POS tags (CPOS) and a fine-grained morphological accuracy (DPOS) for evaluating the POS tagger. In the latter, the analysis is regarded as correct if the main POS tag and each of the morphological features of the token in question are correct.

Results: Table 1 shows the results got by the parsers on the whole Hungarian corpora and on the English datasets. The most important point is that scores are not different from the English scores (although they are not directly comparable). To understand the reasons for this, we manually investigated the set of firing **features** with the highest weights in the Mate parser. Although the assessment of individual feature contributions to a particular decoder decision is not straightforward, we observed that features encoding configurational information (i.e. the direction or length of an edge, the words or POS tag sequences/sets between the governor and the dependent) were frequently among the highest weighted features in English but were extremely rare in Hungarian. For instance, one of the top weighted features for a subject dependency in English was the '*there is no word between the head and the dependent*' feature while this never occurred among the top features in Hungarian.

As a control experiment, we trained the Mate parser only having access to the gold-standard POS tag sequences of the sentences, i.e. we switched off the lexicalization and detailed morphological information. The goal of this experiment was to gain an insight into the performance of the parsers which can only access configurational information. These parsers achieved worse results than the full parsers by 6.8 ULA, 20.3 LAS and 2.9 ULA, 6.4 LAS on the development sets of Hungarian and English, respectively. As expected, Hungarian suffers much more when the parser has to learn from configurational information only, especially when grammatical functions have to be predicted (LAS). Despite this, the results of Table 1 show that the parsers can practically eliminate this gap by learning from morphological features (and lexicalization). This means that the data-driven parsers employing a very rich feature set can learn a model which effectively captures the dependency structures using feature weights which are radically different from the ones used for English.

Another cause of the relatively high scores is that the **CPOS accuracy scores** on Hungarian and English are almost equal: 97.2 and 97.3, respectively. This also explains the small difference between the results got by gold-standard and predicted POS tags. Moreover, the parser can also exploit the morphological features as input in Hungarian.

The Mate parser outperformed the other two parsers on each of the four datasets. Comparing the two graph-based parsers Mate and MST, the gap between them was twice as big in LAS than in ULA in Hungarian, which demonstrates that the **one-step approach looking for the maximum labeled spanning tree** is more suitable for Hungarian than the two-step arc labeling approach of MST. This probably holds for other morphologically rich languages too as the decoder can exploit information from the labels of decoded arcs. Based on these results, we decided to use only Mate for our further experiments.

Table 2 provides an insight into the effect of **domain differences** on POS tagging and parsing scores. There is a noticeable difference between the "newspaper" and the "short business news" corpora. Although these domains seem to be close to each other at the first glance (both are news), they have different characteristics. On the one hand, short business news is a very narrow domain consisting of 2-3 sentence long financial short reports. It frequently uses the same grammatical structures (like "Stock indexes rose X percent at the Y Stock on Wednesday") and the lexicon is also limited. On the other hand, the newspaper subcorpus consists of full journal articles covering various domains and it has a fancy journalist style.

The effect of extending the training dataset with out-of-domain parses is not convincing. In spite of the ten times bigger training datasets, there are two subcorpora where they just harmed the parser, and the improvement on other subcorpora is less than 1 percent. This demonstrates well the domain-dependence of parsing.

The parser and the POS tagger react to domain difficulties in a similar way, according to the first four rows of Table 2. This observation holds for the scores of the parsers working with gold-standard POS tags, which suggests that domain difficulties harm POS tagging and parsing as well. Regarding the two last subcorpora, the compositions consist of very short and usually simple sentences and the training corpora are twice as big compared with other subcorpora. Both factors are probably the reasons for the good parsing performance. In the computer corpus, there are many English terms which are manually tagged with an "unknown" tag. They could not be accurately predicted by the POS tagger but the parser could predict their syntactic role.

Table 2 also tells us that the difference between CPOS and DPOS is usually less than 1 percent. This experimentally supports that the **ambiguity among alternative morphological analyses** is mostly present at the POS-level and the morphological features are efficiently identified by our morphological analyser. The most frequent morphological features which cannot be disambiguated at the word level are related to suffixes with multiple functions or the word itself cannot be unambiguously segmented into morphemes. Although the number of such ambiguous cases is low, they form important features for the parser, thus we will focus on the more accurate handling of these cases in future work.

Comparison to CoNLL-2007 results: The best performing participant of the CoNLL-2007 Shared Task (Nivre et al., 2007) achieved an ULA of 83.6 and LAS of 80.3 (Hall et al., 2007) on the Hungarian corpus. The difference between the top performing English and Hungarian systems were 8.14 ULA and 9.3 LAS. The results reported in 2007 were significantly lower and the gap between English and Hungarian is higher than our current values. To locate the sources of difference we carried out other experiments with Mate on the CoNLL-2007 dataset using the gold-standard POS tags (the shared task used gold-standard POS tags for evaluation).

First we trained and evaluated Mate on the original CoNLL-2007 datasets, where it achieved ULA 84.3 and LAS 80.0. Then we used the sentences of the CoNLL-2007 datasets but with the new, manual annotation. Here, Mate achieved ULA 88.6 and LAS 85.5, which means that the modified annotation schema and the less erroneous/noisy annotation caused an improvement of ULA 4.3 and LAS 5.5. The annotation schema changed a lot: coordination had to be corrected manually since it is treated differently after conversion, moreover, the internal structure of adjectival/participial phrases was not marked in the original constituency treebank, so it was also added manually (Vincze et al., 2010). The improvement in the labeled attachment score is probably due to the reduction of the label set (from 49 to 29 labels), which step was justified by the fact that some morphosyntactic information was doubly coded in the case of nouns (e.g. *házzal* (house-INS) "with the/a house") in the original CoNLL-2007 dataset – first, by their morphological case (Cas=ins) and second, by their dependency label (INS).

Lastly, as the CoNLL-2007 sentences came from the newspaper subcorpus, we can compare these scores with the ULA 90.0 and LAS 87.5 of Table 2. The ULA 1.5 and LAS 2.0 differences are the result of the bigger training corpus (9189 sentences on average compared to 6390 in the CoNLL-2007 dataset).

Hungarian			English		
	label	attachment		label	attachment
virtual nodes	31.5%	39.5%	multiword NEs	15.2%	17.6%
conjunctions and negation	–	11.2%	PP-attachment	–	15.9%
noun attachment	–	9.6%	non-canonical word order	6.4%	6.5%
more than 1 premodifier	–	5.1%	misplaced clause	–	9.7%
coordination	13.5%	16.5%	coordination	8.5%	12.5%
mislabeled adverb	16.3%	–	mislabeled adverb	40.1%	–
annotation errors	10.7%	6.8%	annotation errors	9.7%	8.5%
other	28.0%	11.3%	other	20.1%	29.3%
TOTAL	100%	100%	TOTAL	100%	100%

Table 3: The most frequent corpus-specific and general attachment and labeling error categories (based on a manual investigation of 200–200 erroneous sentences).

6 A Systematic Error Analysis

In order to discover specialties and challenges of Hungarian dependency parsing, we conducted an error analysis of parsed texts from the newspaper domain both in English and Hungarian. 200 randomly selected erroneous sentences from the output of Mate were investigated in both languages and we categorized the errors on the basis of the linguistic phenomenon responsible for the errors – for instance, when an error occurred because of the incorrect identification of a multiword Named Entity containing a conjunction, we treated it as a Named Entity error instead of a conjunction error –, i.e. our goal was to reveal the real linguistic sources of errors rather than deducing from automatically countable attachment/labeling statistics.

We used the parses based on gold-standard POS tagging for this analysis as our goal was to identify the challenges of parsing independently of the challenges of POS tagging. The error categories are summarized in Table 3 along with their relative contribution to attachment and labeling errors. This table contains the categories with over 5% relative frequency.[5]

The 200 sentences contained 429/319 and 353/330 attachment/labeling errors in Hungarian and English, respectively. In Hungarian, attachment errors outnumber label errors to a great extent whereas in English, their distribution is basically the same. This might be attributed to the higher level of non-projectivity (see Section 4) and to the more fine-grained label set of the English dataset (36 against 29 labels in English and Hungarian, respectively).

Virtual nodes: In Hungarian, the most common source of parsing errors was virtual nodes. As there are quite a lot of verbless clauses in Hungarian (see Section 2 on sentences without copula), it might be difficult to figure out the proper dependency relations within the sentence, since the verb plays the central role in the sentence, cf. Tesnière (1959). Our parser was not efficient in identifying the structure of such sentences, probably due to the lack of information for data-driven parsers (each edge is labeled as Exd while they have similar features to ordinary edges). We also note that the output of the current system with Exd labels does not contain too much information for downstream applications of parsing. The appropriate handling of virtual nodes is an important direction for future work.

Noun attachment: In Hungarian, the nominal arguments of infinitives and participles were frequently erroneously attached to the main verb. Take the following sentence: *A Horn-kabinet idején jól bevált módszerhez próbálnak meg visszatérni* (the Horn-government time-3SGPOSS-SUP well tried method-ALL try-3PL PREVERB return-INF) "They are trying to return to the well-tried method of the Horn government". In this sentence, *a Horn-kabinet idején* "during the Horn government" is a modifier of the past participle *bevált* "well-tried", however, it is attached to the main verb *próbálnak* "they are trying" by the parser. Moreover, *módszerhez* "to the method" is an argument of the infinitive *visszatérni* "to return", but the parser links it to the main

[5]The full tables are available at `www.inf.u-szeged.hu/rgai/SzegedTreebank`.

verb. In free word order languages, the order of the arguments of the infinitive and the main verb may get mixed, which is called scrambling (Ross, 1986). This is not a common source of error in English as arguments cannot scramble.

Article attachment: In Hungarian, if there is an article before a prenominal modifier, it can belong to the head noun and to the modifier as well. In *a szoba ajtaja* (the room door-3SGPOSS) "the door of the room" the article belongs to the modifier but when the prenominal modifier cannot have an article (e.g. *a februárban induló projekt* (the February-INE starting project) "the project starting in February"), it is attached to the head noun (i.e. to *projekt* "project"). It was not always clear for the parser which parent to select for the article. In contrast, these cases are not problematic in English since the modifier typically follows the head and thus each article precedes its head noun.

Conjunctions or **negation words** – most typically the words *is* "too", *csak* "only/just" and *nem/sem* "not" – were much more frequently attached to the wrong node in Hungarian than in English. In Hungarian, they are ambiguous between being adverbs and conjunctions and it is mostly their conjunctive uses which are problematic from the viewpoint of parsing. On the other hand, these words have an important role in marking the information structure of the sentence: they are usually attached to the element in focus position, and if there is no focus, they are attached to the verb. However, sentences with or without focus can have similar word order but their stress pattern is different. Dependency parsers obviously cannot recognize stress patterns, hence conjunctions and negation words are sometimes erroneously attached to the verb in Hungarian.

English sentences with **non-canonical word order** (e.g. questions) were often incorrectly parsed, e.g. the noun following the main verb is the object in sentences like *Replied a salesman: 'Exactly.'*, where it is the subject that follows the verb for stylistic reasons. However, in Hungarian, morphological information is of help in such sentences, as it is not the position relative to the verb but the case suffix that determines the grammatical role of the noun.

In English, high or low **PP-attachment** was responsible for many parsing ambiguities: most typically, the prepositional complement which follows the head was attached to the verb instead of the noun or vice versa. In contrast, Hungarian is a head-after-dependent language, which means that dependents most often occur before the head. Furthermore, there are no prepositions in Hungarian, and grammatical relations encoded by prepositions in English are conveyed by suffixes or postpositions. Thus, if there is a modifier before the nominal head, it requires the presence of a participle as in: *Felvette a kirakatban levő ruhát* (take.on-PAST3SGOBJ the shop.window-INE being dress-ACC) "She put on the dress in the shop window". The English sentence is ambiguous (either the event happens in the shop window or the dress was originally in the shop window) while the Hungarian has only the latter meaning.[6]

General dependency parsing difficulties: There were certain structures that led to typical label and/or attachment errors in both languages. The most frequent one among them is **coordination**. However, it should be mentioned that syntactic ambiguities are often problematic even for humans to disambiguate without contextual or background semantic knowledge.

In the case of label errors, the relation between the given node and its parent was labeled incorrectly. In both English and Hungarian, one of the most common errors of this type was **mislabeled adverbs** and adverbial phrases, e.g. locative adverbs were labeled as ADV/MODE. However, the frequency rate of this error type is much higher in English than in Hungarian, which may be related to the fact that in the English corpus, there is a much more balanced distribution of adverbial labels than in the Hungarian one (where the categories MODE and TLOCY are responsible for 90% of the occurrences). Assigning the most frequent label of the training dataset to each adverb yields an accuracy of 82% in English and 93% in Hungarian, which suggests that there is a higher level of ambiguity for English adverbial phrases. For instance, the preposition *by* may introduce an adverbial modifier of manner (MNR) in *by creating a bill* and the agent in a passive sentence (LGS). Thus, labeling adverbs seems to be a more

[6]However, there exists a head-before-dependent version of the sentence (*Felvette a ruhát a kirakatban*), whose preferred reading is "She was in the shop window while dressing up", that is, the modifier belongs to the verb.

difficult task in English.[7]

Clauses were also often mislabeled in both languages, most typically when there was no overt conjunction between clauses. Another source of error was when **more than one modifier** occurred before a noun (5.1% and 4.2% of attachment errors in Hungarian and in English): in these cases, the first modifier could belong to the noun (*a brown Japanese car*) or to the second modifier (*a brown haired girl*).

Multiword Named Entities: As we mentioned in Section 4, members of multiword Named Entities had a proper noun POS-tag and an NE label in our dataset. Hence when parsing is based on gold standard POS-tags, their recognition is almost perfect while it is a frequent source or errors in the CoNLL-2009 corpus. We investigated the parse of our 200 sentences with predicted POS tags at NEs and found that this introduces several errors (about 5% of both attachment and labeling errors) in Hungarian. On the other hand, the results are only slightly worse in English, i.e. identifying the inner structure of NEs does not depend on whether the parser builds on gold standard or predicted POS-tags since function words like conjunctions or prepositions – which mark grammatical relations – are tagged in the same way in both cases. The relative frequency of this error type is much higher in English even when the Hungarian parser does not have access to the gold proper noun POS tags. The reason for this is simple: in the Penn Treebank the correct internal structure of the NEs has to be identified beyond the "phrase boundaries" while in Hungarian their members just form a chain.

Annotation errors: We note that our analysis took into account only sentences which contained at least one parsing error and we crawled only the dependencies where the gold standard annotation and the output of the parser did not match. Hence, the frequency of annotation errors is probably higher than we found (about 1% of the entire set of dependencies) during our investigation as there could be annotation errors in the "error-free" sentences and also in the investigated sentences where the parser agrees with that error.

[7]We would nevertheless like to point out that adverbial labels have a highly semantic nature, i.e. it could be argued that it is not the syntactic parser that should identify them but a semantic processor.

7 Conclusions

We showed that state-of-the-art dependency parsers achieve similar results – in terms of attachment scores – on Hungarian and English. Although the results with this comparison should be taken with a pinch of salt – as sentence lengths (and information encoded in single words) differ, domain differences and annotation schema divergences are uncatchable – we conclude that parsing Hungarian is just as hard a task as parsing English. We argued that this is due to the relatively good POS tagging accuracy (which is a consequence of the low ambiguity of alternative morphological analyses of a sentence and the good coverage of the morphological analyser) and the fact that data-driven dependency parsers employ a rich feature representation which enables them to learn different kinds of feature weight profiles.

We also discussed the domain differences among the subcorpora of the Szeged Dependency Treebank and their effect on parsing results. Our results support that there can be higher differences in parsing scores among domains in one language than among corpora from a similar domain but different languages (which again marks pitfalls of inter-language comparison of parsing scores).

Our systematic error analysis showed that handling the virtual nodes (mostly empty copula) is a frequent source of errors. We identified several phenomena which are not typically listed as Hungarian syntax-specific features but are challenging for the current data-driven parsers, however, they are not problematic in English (like the attachment of conjunctions and negation words and the attachment problem of nouns and articles). We concluded – based on our quantitative analysis – that a further notable error reduction is only achievable if distinctive attention is paid to these language-specific phenomena.

We intend to investigate the problem of virtual nodes in dependency parsing in more depth and to implement new feature templates for the Hungarian-specific challenges as future work.

Acknowledgments

This work was supported in part by the Deutsche Forschungsgemeinschaft grant SFB 732 and the NIH grant (project codename MASZEKER) of the Hungarian government.

References

Zoltán Alexin, János Csirik, Tibor Gyimóthy, Károly Bibok, Csaba Hatvani, Gábor Prószéky, and László Tihanyi. 2003. Annotated Hungarian National Corpus. In *Proceedings of the EACL*, pages 53–56.

Anna Babarczy, Bálint Gábor, Gábor Hamp, and András Rung. 2005. Hunpars: a rule-based sentence parser for Hungarian. In *Proceedings of the 6th International Symposium on Computational Intelligence*.

Csongor Barta, Dóra Csendes, János Csirik, András Hócza, András Kocsor, and Kornél Kovács. 2005. Learning syntactic tree patterns from a balanced Hungarian natural language database, the Szeged Treebank. In *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 225 – 231.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164.

Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 957–961.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged Treebank. In *TSD*, pages 123–131.

Katalin É. Kiss. 2002. *The Syntax of Hungarian*. Cambridge University Press, Cambridge.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: an exploration. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 340–345.

Richárd Farkas, Dániel Szeredi, Dániel Varga, and Veronika Vincze. 2010. MSD-KR harmonizáció a Szeged Treebank 2.5-ben [Harmonizing MSD and KR codes in the Szeged Treebank 2.5]. In *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 349–353.

Jan Hajič, Alena Böhmová, Eva Hajičová, and Barbora Vidová-Hladká. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Amsterdam:Kluwer.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18.

Johan Hall, Jens Nilsson, Joakim Nivre, Gülsen Eryigit, Beáta Megyesi, Mattias Nilsson, and Markus Saers. 2007. Single Malt or Blended? A Study in Multilingual Parser Optimization. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 933–939.

Szilárd Iván, Róbert Ormándi, and András Kocsor. 2007. Magyar mondatok SVM alapú szintaxis elemzése [SVM-based syntactic parsing of Hungarian sentences]. In *V. Magyar Számítógépes Nyelvészeti Konferencia*, pages 281–283.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–7.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37:197–230.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-Projective Dependency Parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-Based Dependency Parsing. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 49–56.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Gábor Prószéky, László Tihanyi, and Gábor L. Ugray. 2004. Moose: A Robust High-Performance Parser and Generator. In *Proceedings of the 9th Workshop of the European Association for Machine Translation*.

John R. Ross. 1986. *Infinite syntax!* ABLEX, Norwood, NJ.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 173–180.

Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Eszter Simon, and Péter Vajda. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC '06)*.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP*, pages 590–596.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.