Evaluation of Universal Dependency parsers for Hungarian

Evelin Ács[1], Gábor Recski[2]

[1] Department of Theoretical Linguistics
Eötvös Loránd University
[2] Department of Automation and Applied Informatics
Budapest University of Technology and Economics

Abstract. We present the results of manual error analysis performed on sample output of three recent dependency parser systems that achieved top scores at the CoNLL 2017 Shared Task: *Multilingual Parsing From Raw Text To Universal Dependencies*. We separate error classes that are sensitive to the technicalities of UD annotation and evaluation principles, such as the treatment of punctuation and coordination, from core issues of syntactic analysis, such as identifying the main predicate in a sentence and resolving structural ambiguity.

1 Introduction

This paper evaluates the performance of three state-of-the-art neural dependency parsers on Hungarian data. The 2017 CoNLL Shared Task on multilingual dependency parsing [1] provided an opportunity to test multiple language-independent parsers on data released as part of the Universal Dependencies v2.0 corpus. Quantitative analysis has been performed as part of the shared task, here we perform a qualitative analysis on a small subset of the Hungarian test data to identify the most important causes of parser errors. We briefly describe the Universal Dependencies project, the shared task, and some related work on the parsing of Hungarian in Section 2. Section 3 presents the results of our manual evaluation and Section 4 draws some conclusions.

2 Background

2.1 Universal Dependencies

Dependency parsing has recently become one of the most investigated technologies in natural language processing, dependency structures are among the most common types of syntactic representations used by downstream NLP applications. In an effort to develop a cross-linguistically consistent annotation system, the Universal Dependencies (UD) project[1] brought together hundreds

[1] http://universaldependencies.org/

of researchers and has resulted in the publication of over 100 UD treebanks in 60 languages (as of version 2.1, released in November 2017) [2].

As a response to the heightened interest in UD and dependency parsing, the 2017 edition of the Conference on Natural Language Learing (CoNLL) organized a shared task on *"Multilingual Parsing from Raw Text to Universal Dependencies"* [1]. The training data – based on version 2.0 of the Universal Dependency dataset – consisted of 64 treebanks for 45 languages. Test treebanks contained at least 10,000 words for each language in the training set and an additional 4 surprise languages.

33 research groups submitted solutions to the task, their systems were ranked based on the macro-average of labeled attachment F-scores (LAS) achieved on each language. LAS matches require that a dependency is assigned to the correct pair of tokens in a sentence and with the correct label. In contrast, unlabeled attachment score (UAS) is more lenient in that it disregards edge labels. A third metric commonly used to evaluate dependency parsers is Content-word Labeled Attachment Score (CLAS), which only considers relation between content words and not function words or punctuation.

In Section 3 we shall analyze errors made by the top three parsers in the competition. The Stanford [3] and C2L2 (Cornell) [4] teams submitted neural parsers that use LSTMs for representing input sentences; both of these systems leverage character-level representations to handle languages with rich morphologies. The Stuttgart IMS team's solution [5] uses CRFs for POS/morphological tagging and a neural tagger for predicting supertags. Overall scores and scores for Hungarian data achieved by each of these three systems is presented in Table 1. Note that the gap between these three systems and the next teams is quite large so that Stanford, C2L2, and IMS are the top three systems based on any of the metrics presented here, and in particular for the Hungarian data.

	Overall			Hungarian		
	LAS	CLAS	UAS	LAS	CLAS	UAS
UnstableParser (Stanford)	76.30	72.57	81.30	77.56	76.08	82.35
C2L2 (Cornell)	75.00	70.90	80.32	76.55	74.36	82.07
IMS (Stuttgart)	74.42	70.18	79.90	73.55	70.87	79.90

Table 1. LAS, CLAS and UAS scores of all three parsers

2.2 Dependency parsing of Hungarian

The Hungarian section of the Universal Dependencies dataset has been created using the Szeged Dependency Treebank [6], challenges of the conversion process are described in [7]. A manual error analysis similar to ours has been performed on Hungarian data before: [8] inspects 200 sentences from the output of Bohnet's parser [9] trained on the Szeged Dependency Treebank. A meaningful comparison of our analysis and theirs is not possible due to the differences between the two tasks: most error classes are specific to the respective annotation systems.

3 Evaluation

We inspected manually the analyses given by each of the three parsers on the first 50 sentences of the Hungarian test data. We grouped errors both by the types of dependency relations they involved and by the types of errors, i.e. the way in which the parsers misinterpreted the structure of a phrase, a clause, or an entire sentence. The number of erroneous edges in each output is similar in all three outputs: the Stanford data contained 208, C2L2 245, and IMS 261. Table 2 lists the top errors by edge type.

UnstableParser		C2L2		IMS	
punct	43	punct	46	punct	44
cc	13	cc	17	cc	17
det	11	det	16	det	11
advmod	9	conj	6	advmod	11
amod	7	conj-nmod	5	amod	7
conj	7	cc-advmod	5	amod-conj	6

Table 2. Types of erroneous edges

As we shall also see when grouping errors by their possible cause, punctuation is the single largest error class for each of the three systems. It has been questioned whether edges in a dependency graph that connect punctuation symbols to some word in the sentence are relevant to dependency structure, in fact the UD community is currently experimenting with the CLAS score as a means to disregard these edges when evaluating dependency parsers [1, p.7]. The `cc` relation is also ignored by CLAS scoring: it is responsible for connecting conjuncts such as *és* ('and'), *de* ('but'), etc. to some other word in the sentence.

3.1 Error types

We shall now describe the most common classes of errors, based on a close observation of each misinterpreted sentence. Besides punctuation and conjuncts we shall discuss 4 additional problem classes that are each responsible for between 2 and 7% of all observed errors (see Table 3 for counts).

Root elements In nearly a fifth of all sentences observed, parsers assigned the `root` dependency to the wrong word, i.e. they failed to identify the main predicate of the sentence. These errors are worthy of attention not only because of their frequency but because they are usually responsible for several further erroneous edges – if the parser misses the main predicate, it is likely to miss relations of each of its dependents. An example of this phenomenon is shown in Figures 1 and 2, which show the gold and erroneous dependency analyses of the sentence in (1).

	UnstableParser	C2L2	IMS
punct, cc	59	63	61
root	9 (15)	9 (17)	10 (19)
conj	9	9	7
modifier POS	8	5	13
structural ambiguity	6 (8)	6 (8)	5 (7)

Table 3. Number of occurrences of each error type (number of edges affected, if different)

(1) – *Azért* *nem* *lehetett* *olyan* *rossz* *közelről* *élvezni* *a*
– Because not be-CAN-PAST so bad near-DEL enjoy-INF the
nehézsúly *Lewis-Holyfield-csúcsrangadóját!*
heavy-weight Lewis-Holyfield-faceoff-ACC!

It can't have been that bad, enjoying the Lewis-Holyfield faceoff from so close!

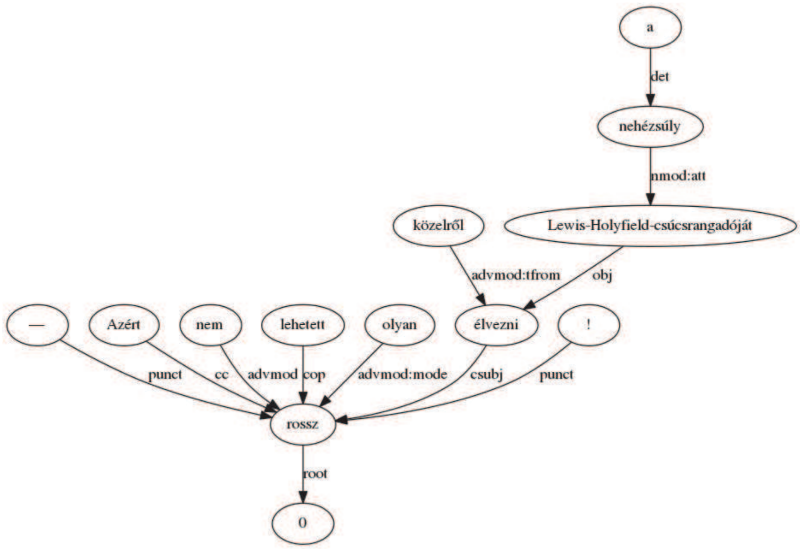

Fig. 1: Gold analysis of (1)

Coordination Another group of errors involves coordinating conjunctions. In UD, conjunctions are treated asymmetrically: one of the coordinated elements is considered the head of the conjunction and others are connected only to this element (via the `conj` relation) but not to any other word in the sentence. Parser

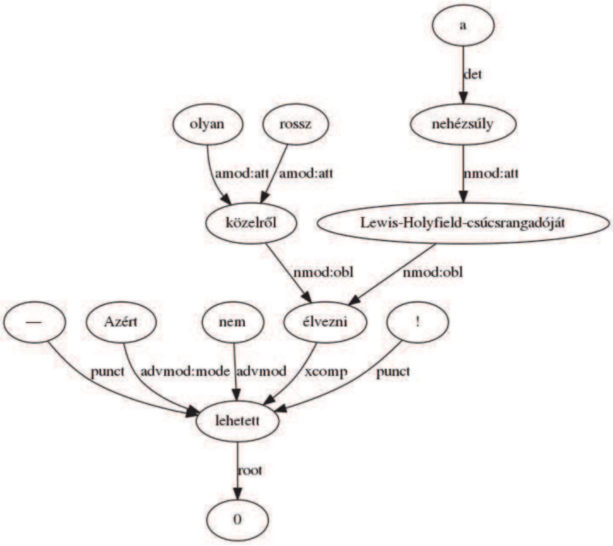

Fig. 2: Incorrect analysis of (1) by the IMS parser

errors occur when these non-head elements of a conjunction are also connected to other words. These erroneous relations can be justified, since they reflect dependencies that actually hold between some word and *each* element of a coordinated structure – nevertheless this treatment goes against UD conventions. An example is shown in Figure 4, a partial analysis of the sentence in (2).

(2) *Ezek is leginkább csak a januári napokban, amikorra*
these too mainly just the January-ATT day-PL-INE, when-SUBL
a fa már kiszáradt és egy csillagszóró is lángba
the tree already dry-out-PAST and a sparkler too flame-INE
boríthatja – mondta az alezredes.
cover-DEF – say-PAST the colonel.

But only in the days of January, when the tree is dry and a sparkler might burn it down – said the colonel.

Modifiers The UD relations `nmod` and `amod` represent the dependencies between a noun and its nominal or adjectival modifier, respectively. Similarly, the `advmod` relation connects adverbs to predicates or modifiers. A large portion of errors were caused by parsers mixing the above three labels on edges that were otherwise correctly identified, i.e. they connected the modifiers to the right word. Since the distinction between `nmod`, `amod`, and `advmod` is based entirely on the part-of-speech (POS) categories of dependents, one may expect that each of these errors are direct results of POS-tagging mistakes. In fact, out of 26 such

Fig. 3: Partial analysis of (2)

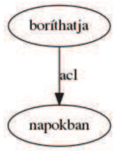

Fig. 4: Partial analysis of (2) by both the C2L2 and the Stanford parser

errors in the three datasets (Stanford: 8, C2L2: 5, IMS: 13), only 14 (4, 2, 8) are in line with the above assumption: the output contains an incorrect POS-tag for the modifier word and the dependency label reflects the same mistake (an example is shown in Table 4). In the remaining 12 cases dependency labels were assigned incorrectly despite a correct POS-tag. In 4 cases, however, 2 made by the Stanford system and 2 by IMS, one may argue that the incorrect dependency labels are actually justified, while gold labels are a result of annotators' compliance with gold POS-tags that are linguistically questionable. An example is shown in Table 5.

	az	*Y2K*	*problémát*
	the	Y2K	problem-ACC
gold POS	DET	NOUN	NOUN
gold dependency	det	nmod	
IMS POS	DET	ADJ	NOUN
IMS dependency	det	amod	

Table 4. Gold and IMS analyses of a noun phrase

Structural ambiguity The final error group involves sentences that are structurally ambiguous and whose parses are consistent with a different constituent structure than the one reflected by the gold dependency annotation. The ambiguous phrase of one such sentence is shown in (3), with English paraphrases for both possible readings. The two dependency structures are shown in Figure 5 and Figure 6.

	türelmetlenül	újra	tárcsáz
	impatient-ESS	again	dial
	'dials again impatiently'		
gold POS	ADJ	ADV	VERB
gold dependency	amod	advmod	
IMS POS	ADJ	ADV	VERB
IMS dependency	advmod	advmod	

Table 5. Gold and IMS analyses of a noun phrase

(3) *a Péterfy kórház sürgősségi belgyógyászati és klinikai toxikológiai*
the Péterfy hospital emergency internal-medicine and clinical toxicology
osztálya
department-POSS

The department of emergency internal medicine and clinical toxicology
The emergency department of internal medicine and clinical toxicology

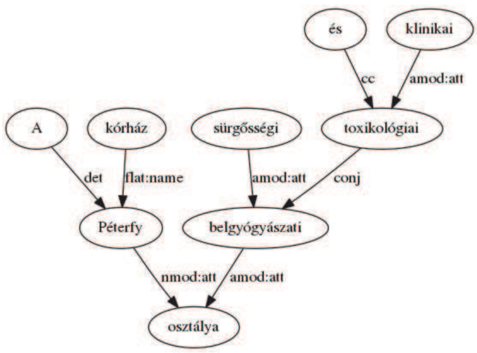

Fig. 5: Gold analysis of (3).

3.2 Comparison

[8] includes an error analysis on the output of Bohnet's parser trained on Hungarian data from the Szeged Dependency Treebank. Their method is similar to ours and involves manual inspection of 200 parser errors on the news section of the Szeged dataset.

4 Conclusion

We have presented the results of manual error analysis of three dependency parsers on a small sample of Hungarian data. We have identified several error

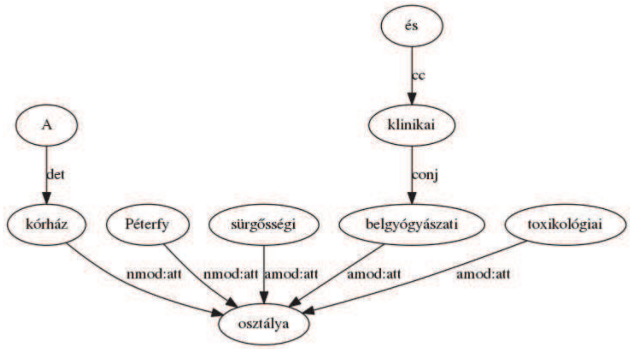

Fig. 6: UnstableParser's analysis of (3).

classes that are in some ways technical: those concerning punctuations and conjuncts have little relevance to the dependency structure of content words and underline the necessity of alternative evaluation metrics like CLAS, while those involving coordinating conjunctions introduce edges that may be justifiable and might challenge UD's current treatment of coordination. Modifier relations have brought to light errors in POS-tagging and some possible inconsistencies in the gold standard data. Finally, we have seen examples of structural ambiguity, which remains one of the most challenging problems in syntactic analysis.

References

1. Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C.D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H.F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., Li, J.: CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 1–19
2. Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M.J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bengoetxea, K., Bhat, R.A., Bick, E., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G.G.A., Cetin, S., Chalub, F., Choi, J., Cinková, S., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.C., de Paiva, V., Diaz de Ilarraza, A., Dirix, P., Dobrovoljc, K., Dozat, T.,

Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà My, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Laippala, V., Lambertino, L., Lando, T., Lee, J., Lê Hong, P., Lenci, A., Lertpradit, S., Leung, H., Li, C.Y., Li, J., Li, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyen Thi, L., Nguyen Thi Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Östling, R., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalnia, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Rama, T., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Romanenko, M., Rosa, R., Rovati, D., Sagot, B., Saleh, S., Samardžić, T., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Washington, J.N., Wirén, M., Wong, T.s., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhu, H.: Universal dependencies 2.1 (2017) LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
3. Dozat, T., Qi, P., Manning, C.D.: Stanford's graph-based neural dependency parser at the conll 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 20–30
4. Shi, T., Wu, F.G., Chen, X., Cheng, Y.: Combining global models for parsing universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 31–39
5. Björkelund, A., Falenska, A., Yu, X., Kuhn, J.: IMS at the CoNLL 2017 UD Shared Task: CRFs and Perceptrons Meet Neural Networks. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, Association for Computational Linguistics (2017) 40–51
6. Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
7. Vincze, V., Simkó, K.I., Szántó, Z., Farkas, R.: Universal dependencies and morphology for hungarian-and on the price of universality, Association for Computational Linguistics (2017)

8. Farkas, R., Vincze, V., Schmid, H.: Dependency parsing of hungarian: Baseline results and challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2012) 55–65
9. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, Coling 2010 Organizing Committee (2010) 89–97