Hungarian copula constructions in dependency syntax and parsing

Katalin Ilona Simkó
University of Szeged
Institute of Informatics
Department of General Linguistics
Hungary
simko@hung.u-szeged.hu

Veronika Vincze
University of Szeged
Institute of Informatics
MTA-SZTE
Research Group on Artificial Intelligence
Hungary
vinczev@inf.u-szeged.hu

Abstract

Copula constructions are problematic in the syntax of most languages. The paper describes three different dependency syntactic methods for handling copula constructions: function head, content head and complex label analysis. Furthermore, we also propose a POS-based approach to copula detection. We evaluate the impact of these approaches in computational parsing, in two parsing experiments for Hungarian.

1 Introduction

Copula constructions show some special behaviour in most human languages. In sentences with copula constructions, the sentence's predicate is not simply the main verb of the clause, but the copula verb plus a nominal predicate (in "Peter sleeps" the sentence's predicate is the verb, "sleeps", while in "Peter is tired", it is the copula verb and the nominal predicate, "is tired") . This is further complicated by the fact that the copula verb shows non-conventional behaviour in many languages: it is often not present in the surface structure for one or more slots of the verbal paradigm.

These constructions are widely studied: many approaches are available in many different syntactic frameworks, like in Den Dikken (2006), Partee (1998) and É. Kiss (2002) in constituency grammar; or Dalrymple et al. (2004) and Laczkó (2012) in LFG.

In this paper, we focus on dependency syntactic approaches. In dependency syntax, the syntactic structure's nodes are the words themselves and the tree is made up of their hierarchical relations, making both two-word predicates and the missing verbal forms cause difficulties. Should the copula, the verbal part of the predicate, be the head of the structure, parallel to most other types of constructions? And if so, how can we deal with cases where the copula is not present in the surface structure?

In this paper, three different answers to these questions are discussed: the function head analysis, where function words, such as the copula, remain the heads of the structures; the content head analysis, where the content words, in this case, the nominal part of the predicate, are the heads; and the complex label analysis, where the copula remains the head also, but the approach offers a different solution to zero copulas.

First, we give a short description of Hungarian copula constructions. Second, the three dependency syntactic frameworks are discussed in more detail. Then, we describe two experiments aiming to evaluate these frameworks in computational linguistics, specifically in dependency parsing for Hungarian, similar to Nivre et al. (2007). The first experiment compares the three previously mentioned frameworks, while the second introduces our new approach, based on differentiating the copula and existential "be" verbs on the level of POS-tagging, which can improve the performance of the content head analysis.

2 Copula constructions in Hungarian

The Hungarian verb *van* "be" behaves similarly to "be" verbs in other languages: it has two distinct uses: as an existential and as a copular verb. In the existential use, *van* behaves just as any other main, content verb: it is the only predicative element in the clause and it is always present in the surface structure. On the other hand, in the copular use *van* requires a nominal predicate, a noun or an adjective in the nominative case; copular *van* is never present in the surface structure for 3rd person, present tense, declarative clauses, but its other forms are the same as for the existential.

Below we illustrate Hungarian copula constructions with several examples, see Table 1 and Ex-

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 240-247,
Pisa, Italy, September 18-20 2017

	Existential *van*	Copular *van*
1st Sg PR	vagyok	vagyok
2nd Sg PR	vagy	vagy
3rd Sg PR	van	-
1st Pl PR	vagyunk	vagyunk
2nd Pl PR	vagytok	vagytok
3rd Pl PR	vannak	-
1st Sg PAST	voltam	voltam
2nd Sg PAST	voltál	voltál
3rd Sg PAST	volt	volt
1st Pl PAST	voltunk	voltunk
2nd Pl PAST	voltatok	voltatok
3rd Pl PAST	voltak	voltak

Table 1: Present and past tense paradigm for existential and copular *van* in Hungarian.

amples (1-4), where (1) and (2) are present and past tense existential *van* sentences, while (3) and (4) are copular. In Examples (1) and (2), *van* and *volt* are the only predicative elements of the sentence respectively. In the copular *van* sentences (3) and (4), the first, present tense sentence has the zero copula, in the surface structure only *orvos* "doctor", the nominal predicate makes up the predicative part of the sentence, while in Example (4), where the copula is overt, the nominal predicate and the copula, *orvos* "doctor" and *volt* "was" jointly make up the predicative part of the sentence.

(1) Péter a szobában van.
Peter.NOM the room.INE is.PR.3rdSG
Peter is in the room.

(2) Péter a szobában volt.
Peter.NOM the room.INE is.PAST.3rdSG
Peter was in the room.

(3) Péter orvos.
Peter.NOM doctor.NOM
Peter is a doctor.

(4) Péter orvos volt.
Peter.NOM doctor.NOM is.PAST.3rdSG
Peter was a doctor.

The copula's behaviour in Hungarian is by no means unique: for most languages, the copula shows some difference from verbs in general and zero copulas in the verbal paradigm are also relatively common (Curnow, 2000).

3 Copula constructions in dependency syntax

Copula constructions in languages like Hungarian cause two problems for dependency syntax. First, with the dual predicate (nominal + copula) it is not obvious which one should be the head of the construction: should the verbal element be the head parallel to non-copular sentences or should the nominal be the head as that element is always overt? Second, how to handle the zero copula in the syntactic structures?

In this section, three approaches are described giving different answers to the questions above: the function head approach, the content head approach and the complex label approach.

3.1 Function head approach

The function head approach to dependency syntax goes back to the foundations of Mel'čuk's (2009) framework. He proposed that the function words of the sentence should be the heads over content words; function words should be the ones setting up the basic syntactic structure of the sentence.

Mel'čuk also writes about copular constructions and the above-mentioned issues in his work and stands by the function head analysis: he proposes that in languages where the copula is only zero in certain slots of the paradigm, but overt in others, a virtual, zero verb form should be inserted into the syntactic structure. This zero copula is the head of the structure, the nominal predicate is a dependent of it. This way, we preserve a common structure for all sentences in which the inflected verb is always the head of the clause, but we violate one of the core principles of dependency syntax: surface structure words are no longer the only nodes in the tree.

The function head approach is the annotation of the Szeged Dependency Treebank (Vincze et al., 2010), the large Hungarian dependency treebank used for the experiments described in the paper. The first column of Table (2) shows the Szeged Dependency Treebank's annotation for the existential sentence (Example (1)), and the copular sentences with overt and zero copula, Examples (3) and (4). The capitalized *VAN* is the inserted virtual node in zero copula sentences that was added manually to all sentences of this type in a preprocessing step.

3.2 Content head approach

The content head approach recently gained popularity in computational dependency syntax due to the Universal Dependencies project (Nivre, 2015).

This analysis considers content words the frame of the syntactic structure: content words are the heads and function words are their dependents. This separates the copula from all other verbs, even the existential verb. As all other verbs carry content, they are heads in this analysis also, while the copula, as a function word, becomes a dependent in this analysis. This way we no longer have a common structure for all clauses, but we have an analysis that has no issues with the zero copula.

A section of the Szeged Dependency Treebank has been converted to the Universal Dependencies annotation (Vincze et al., 2015; Vincze et al., 2017). In the experiments, this treebank is used as the content head analysis. The second column of Table (2) shows the sentences in Examples (1), (3) and (4) again, this time with the content head analysis in the Szeged Universal Dependencies Treebank.

3.3 Complex label approach

The complex label approach is a computational linguistic variation of the function head analysis detailed in Seeker et al. (2012).

They keep the function words as the heads, therefore keeping the copula as the head of the copular clause, but they deal with the zero copula in a different way. The analysis does not use virtual nodes, but instead "shows" the missing copula in the dependency labels originating from where it would be inserted. As in the zero copula example for complex label in Table 2.

, the root node of the structure in the function head analysis would be a virtual *VAN* node, the subject, *Peter* would be a dependent of *VAN*. Therefore the Complex label dependency label of the subject is **ROOT-VAN-SUBJ**: the original "route" to it would be **ROOT** label to *VAN*, **SUBJ** label to *Peter*, the virtual node is removed, but the "route" is still shown. This approach gives a similar structure for all clauses with overt verbs, only distinguishing the zero copula. Due to combinations of the complex labels, the approach also uses a lot more (potentially infinite) different dependency labels in the analysis.

The Szeged Dependency Treebank has also been converted to this analysis, which will be used in the experiments. Dependency trees for Examples (1), (3) and (4) are shown again in the third column of Table (2); in Figure (1) a sentence with two coordinated clauses with zero copula to show how the labels can combine.

Table 3 summarizes in which conditions the different approaches give syntactic structures different from regular content verbs analysis for copular sentences. The content head approach gives the most linguistically based distinction by drawing the line between copula and non-copula main verbs.

4 Experiments

We evaluated the three approaches in two parsing experiments. We used the same corpus with three different dependency annotations and the Bohnet parser (Bohnet, 2010) for both.

4.1 The corpus

We used a section of the Szeged Dependency Treebank that is available with all three analyses: the original annotation is function head based, there is an automatically converted complex label version, and the converted, manually corrected Universal Dependencies treebank for the content head version.

The section contains about 1300 sentences, 27000 tokens in total. The data contains 300 instances of virtual *V*, 230 overt copulas and 150 existential *van*s.

4.2 Experiment 1: Function head, content head or complex label

In the first experiment, the Bohnet parser was trained using the ten fold cross validation method on the same corpora of texts for the function head, the content head and the complex label representation separately, using gold POS tags and morphological features. In the evaluation of each model, we used UAS and LAS scores for the whole corpus as well as error analysis for the structures in question. Table 4 shows the UAS and LAS scores for each approach. We were interested in the parsing performances regarding different types of *van* sentences, so we created filtered subcorpora that contain only the sentences with existential *van*, only with overt copula and only with zero copula. We report results calculated for these datasets too.

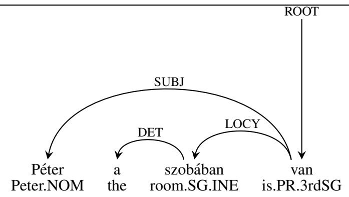
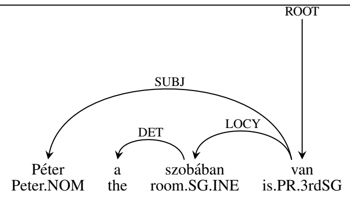
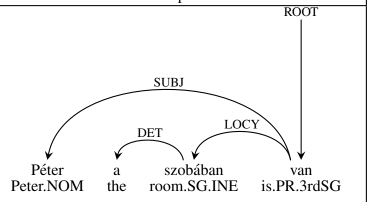
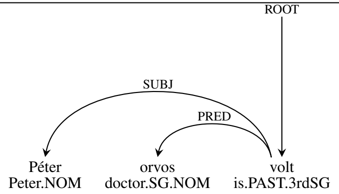
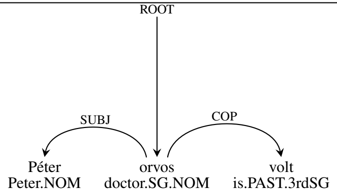
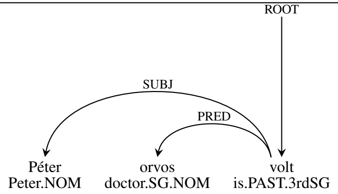
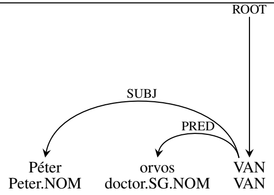
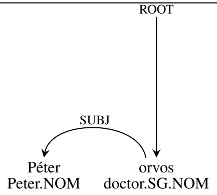
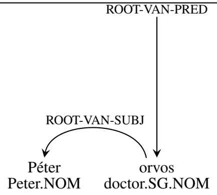
	Function	Content	Complex
Existential	 ROOT SUBJ DET LOCY Péter Peter.NOM a the szobában room.SG.INE van is.PR.3rdSG	 ROOT SUBJ DET LOCY Péter Peter.NOM a the szobában room.SG.INE van is.PR.3rdSG	 ROOT SUBJ DET LOCY Péter Peter.NOM a the szobában room.SG.INE van is.PR.3rdSG
Overt cop.	 ROOT SUBJ PRED Péter Peter.NOM orvos doctor.SG.NOM volt is.PAST.3rdSG	 ROOT SUBJ COP Péter Peter.NOM orvos doctor.SG.NOM volt is.PAST.3rdSG	 ROOT SUBJ PRED Péter Peter.NOM orvos doctor.SG.NOM volt is.PAST.3rdSG
Zero cop.	 ROOT SUBJ PRED Péter Peter.NOM orvos doctor.SG.NOM VAN VAN		 ROOT-VAN-PRED ROOT-VAN-SUBJ Péter Peter.NOM orvos doctor.SG.NOM

Table 2: Syntactic structures for existential, overt and zero copula sentences in function head, content head and complex label approaches. Note how all three trees for the existential sentence are the same, but the copular ones show differences in the analysis.

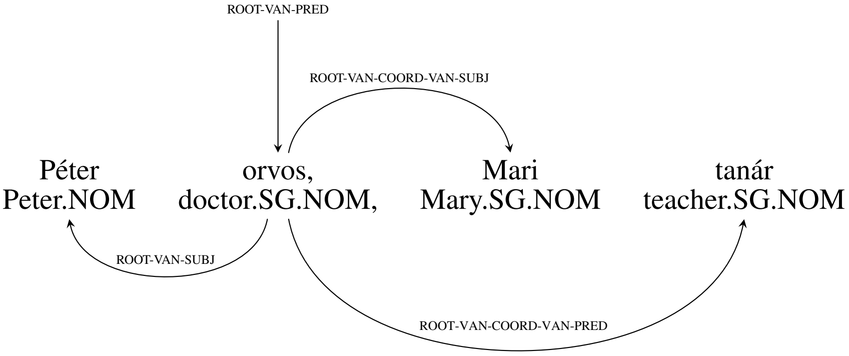


Figure 1: Complex label analysis of the coordinated copular clauses in *Péter orvos, Mari tanár* “Peter is a doctor, Mary is a teacher”.

	Function	Complex	Content
Verb	▨	▨	▨
Exist.	▨	▨	▨
Overt cop.	▨	▨	
Zero cop.	▨		

Table 3: Different analysis from conventional syntactic structure in the different approaches.

Based on these UAS and LAS scores, the function head analysis gives the best results with the complex label analysis as a close second, but we were interested in the specific relations of *van* and not the full sentences' parsing output. We did manual error analysis of the *van* verb's closest relations to investigate which dependency syntactic theory describes these relations best for computational parsing. We considered the following four errors in our analysis: incorrect head in the clause with *van*; incorrectly labeled or attached subject of *van*; incorrectly labeled or attached nominal predicate; subject and nominal predicate mixed up. Sentences showing none of the above errors were considered correct in the results shown below, regardless of other errors in the sentence. Table 5 shows the percentage of correct sentences for each analysis in the three above mentioned subcorpora and the overall results in the bottom row.

4.3 Experiment 2: POS-based approach to the copula

In the second experiment, we investigated a way to improve the content head analysis with a POS-based approach. Our hypothesis is that the existential *van* and the overt copula *van* are better disambiguated on the level of POS tagging: as the copular *van* has a syntactic structure (in the content head analysis), which is very different from the one of all other verbs, not treating it as a normal verb makes sense from a syntactic parsing point of view. For this reason, the level of POS tagging is a better fit to disambiguate existential and copular *van* than the actual parsing. We used the previously introduced Hungarian Universal Dependencies treebank with the content head annotation and created a new, POS-based copula version, where the copula *van* has a new POS tag, **COP** distinguishing it from all other verbs including the existential *van*, as shown in examples (5) and (6).

(5) Péter a szobában volt.
NOUN DET NOUN **VERB**
Peter.NOM the room.INE is.PAST.3rdSG
Peter is in the room.

(6) Péter orvos volt.
NOUN NOUN **COP**
Peter.NOM doctor.NOM is.PAST.3rdSG
Peter was a doctor.

In the experiment, we applied the Bohnet parser this time for POS tagger, morphology tagger, and dependency parser training and evaluation, using ten fold cross validation on the original content head treebank and the new version with the **COP** POS tag. Table 6 gives the UAS and LAS results for the two analyses on a subcorpus with only the sentences with existential, overt or zero *van* and on the full corpus.

The results in Table 6 show very little change on the full corpus and marginally better results on the *van* sentences for the POS-based approach. Again, we focus on manual error analysis of the affected structures.

In the new POS-based content head approach, the new **COP** POS tag for the copula *van* is assigned with 0.699 F-score over the whole corpus and the **COP** POS tag triggers the dependency parser to assign the content head copula structure as expected.

To evaluate the approach, we created a subcorpus of the sentences with existential and overt *van*s, as those are the ones we aim to better disambiguate. On these sentences, we evaluated the accuracy of dependency label prediction of *van*. In both versions in the gold analysis the overt copula *van* has the dependency label **cop**, while the existential *van* has the appropriate verbal dependency label. In our results, for the original content head analysis, the correct label is assigned with 58.14% accuracy, while our POS tag based content head approach assigns the correct label with 60.35% accuracy. Although this not a statistically significant improvement, we believe that the tendencies reported on this relatively small corpus are of importance for parsing sentences with copulas.

5 Discussion

The most common error for all three linguistically plausible analyses is incorrectly labeling or attaching the subject of *van* and mixing it up with the

	Function	Content	Complex
Existential - UAS	86.18	80.48	86.84
Existential - LAS	91.04	77.21	82.46
Overt copula - UAS	82.8	75.05	83.62
Overt copula - LAS	77.31	71.67	77.82
Zero copula - UAS	84.42	78.39	77.5
Zero copula - LAS	79.17	75.15	69.59
Full corpus - UAS	85.75	84.41	84.76
Full corpus - LAS	81.24	81.2	79.89

Table 4: UAS and LAS scores with the three analyses on different subcorpora.

	Function	Content	Complex
Existential	78	80	80
Overt copula	62	42	52
Zero copula	70	68	30
Overall	70	63	54

Table 5: Percentage of correct sentences in the manual error analysis.

nominal predicate. Correctly identifying the subject and the nominal predicate is very hard: both are nominative case nominal phrases and while with first or second person subjects, the agreement with the verb makes them easier to tell apart, when both subject and predicate are third person noun phrases, even native speakers of Hungarian find it difficult to assign the correct structure to the sentence (which can be further complicated by the free word order). With the free word order in Hungarian, both sentences in Figures (2) and (3) can express the same meaning (without having any additional contextual information or information about stress patterns in spoken language), but the subject and predicate relations are not straightforward to assign. In the gold annotation, the annotator must decide on one of the options, but in some cases, both options are plausible, causing issues for the parser.

The manual error analysis shows that the complex label approach gives the worst results for copula constructions: it gives fewer correct copula structures and wrongly assigns the complex labels to parts of the sentence without zero copulas. The training time is also an issue as the complex label model trains almost twice as long as the other two because of the huge number of different labels - the function head approach uses 26 different labels, the content head 50, while the complex label analysis in our case used over 200 distinct labels – theoretically, an infinite number of labels are possible for it. The huge number of distinct dependency labels used in this approach probably influences the lower scores achieved by the system as well, as statistically the system has a much lower chance of assigning the correct label out of a set of 200, than that of 26 or 50 labels.

The function and content head approaches achieved similar results in most cases. Both show the lowest scores for the overt copula cases that are very hard to disambiguate between existential and copular *van*. The two approaches score very similarly on the different error types as well. In interpreting the results, it is important to note that the function head analysis requires a preprocessing step to add the virtual *VAN*s to the corpus in order for them to be analyzed parallel to all other types of verbs; these virtual nodes were already present in both training and test data in the experiment.

Our two experiments were done on the relatively small (approximately 1800 sentences) section of the Szeged Corpus available with function head, content head and complex label gold syntactic analysis, therefore our results are preliminary, but we think the tendencies shown would hold using bigger corpora.

Based on the results of our two experiments, we propose using content head dependency syntactic structures for the analysis of Hungarian copula constructions with our addition of treating the distinction of existential and copular *van* on the level of POS tagging.

6 Conclusions

Our paper discussed Hungarian copula *van* and different possible analyses of copula constructions in dependency syntax, evaluating them in com-

	Original	POS-based
Only *van* sentences- UAS	71.67	72.08
Only *van* sentences- LAS	65.87	66.3
Full corpus - UAS	77.8	77.77
Full corpus - LAS	72.02	72.05

Table 6: UAS and LAS scores for the original and POS-based content head analyses.

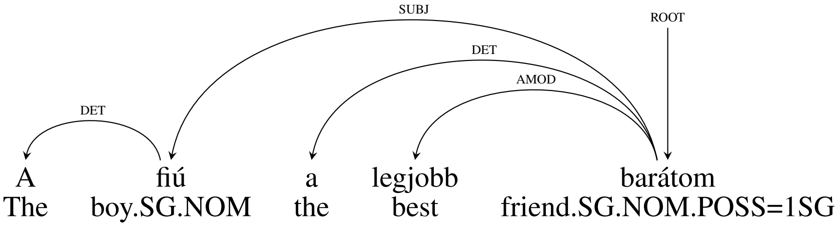


Figure 2: Content head analysis of the copular sentence, *A fiú a legjobb barátom* "The boy is my best friend".

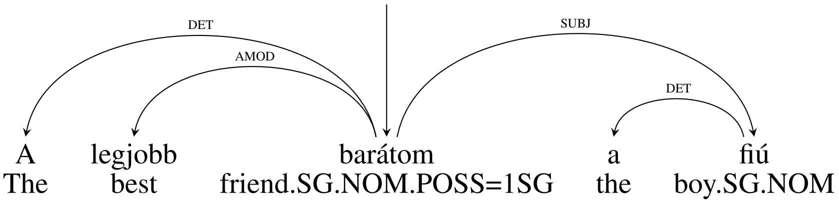


Figure 3: Content head analysis of the copular sentence, *A legjobb barátom a fiú* "My best friend is the boy.".

putational linguistics. We introduced the Hungarian verb *van* and its main linguistic properties, described the function head, content head and complex label approaches to represent copula constructions and showed the results of two parsing experiments focusing on the Hungarian copula. Based on the outcome of our experiments, we support the use of the content head approach with the POS tagging based additions proposed in this paper for the treatment of Hungarian copula constructions.

Our goals in this paper were to show how syntactic analysis can be influenced by not just the syntactic framework, but the specific approach within it and to highlight the importance of manual error analysis alongside the UAS and LAS values. Manual error analysis often shows nuances in the analysis of specific phenomena hidden in overall precision scores and offers more informative results from both computational and linguistics points of view.

In the future, we plan to repeat our experiments on bigger corpora and also for other languages, as well as to investigate other challenging syntactic constructions in a similar fashion.

Acknowledgements

This research was supported by the EU-funded Hungarian grant EFOP-3.6.1- 16-2016- 00008.

References

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.

Timothy Jowan Curnow. 2000. Towards a Cross-Linguistic Typology of Copula Constructions. In John Henderson, editor, *Proceedings of the 1999 Conference of the Australian Linguistic Society*, pages 1–9.

Mary Dalrymple, Helge Dyvik, and Tracy H. King. 2004. Copular Complements: Closed or Open? In *Proceedings of the LFG '04 Conference*, pages 188–198, New Zealand. University of Canterbury.

Marcel Den Dikken. 2006. *Relators and Linkers: The Syntax of Predication, Predicate Inversion, and Copulas*. MIT Press.

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

Tibor Laczkó. 2012. On the (Un)Bearable Lightness of Being an LFG Style Copula in Hungarian. In *The Proceedings of the LFG12 Conference*, pages 341–361, Stanford. CSLI Publications.

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.

Joakim Nivre, 2015. *Towards a Universal Grammar for Natural Language Processing*, pages 3–16. Springer International Publishing, Cham.

Barbara Partee. 1998. Copular Inversion Puzzles in English and Russian. In Katarzyna Dziwirek, Herbert Coats, and Cynthia Vakareliyska, editors, *Formal Approaches to Slavic Linguistics*, pages 361–395.

Alain Polguère and Igor Aleksandrovič Mel'čuk, editors. 2009. *Dependency in Linguistic Description*. Studies in language companion series. Amsterdam Philadelphia, Pa. J. Benjamins.

Wolfgang Seeker, Richárd Farkas, Bernd Bohnet, Helmut Schmid, and Jonas Kuhn. 2012. Data-driven dependency parsing with empty heads. In *Proceedings of COLING 2012: Posters*, pages 1081–1090, Mumbai, India, December. The COLING 2012 Organizing Committee.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of LREC 2010*, Valletta, Malta, May. ELRA.

Veronika Vincze, Richárd Farkas, Katalin Ilona Simkó, Zsolt Szántó, and Viktor Varga. 2015. Univerzális dependencia és morfológia magyar nyelvre. In *XII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 322–329, Szeged.

Veronika Vincze, Katalin Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal Dependencies and Morphology for Hungarian - and on the Price of Universality. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 356–365, Valencia, Spain, April. Association for Computational Linguistics.