Building a dependency treebank from the Hungarian Gigaword Corpus

Ágnes Kalivoda [1], Robert Malouf [2], Farrell Ackerman [3], Bálint Sass [1]

[1] Hungarian Research Centre for Linguistics; [2] San Diego State University; [3] University of California San Diego

1. Aim

Creation of a Hungarian dependency treebank which is suited for large-scale linguistic analysis. That is:

- it has high-quality linguistic annotation
- core metadata are provided, e.g. text style and year of publication
- it is as large as it can get

2. Background

Universal Dependencies (UD): a framework for crosslinguistically consistent morphosyntactic annotation [1]. UD Treebanks for Hungarian:

Szeged Dependency Treebank [9]:

- the first and only manually annotated dependency corpus for Hungarian
- often used as training data for NLP models
- too small (82,000 sentences, 1.2 million words)

Demszky (2021) [2]:

- an attempt to parse the Hungarian Gigaword Corpus [5] using Stanza [7]
- several errors at start → they are amplified in the processing pipeline

This project sets out to build a dependency treebank from the Hungarian Gigaword Corpus, **differing from Demszky (2021) in two key respects:**

- taking several pre-processing steps to clean the text
- taking a different approach to the parsing process

3. Text source

The Hungarian Gigaword Corpus [5] (HGC) is a 1 billion-word general corpus. It represents a wide cross-section of Hungarian from the 20th and 21st centuries.
The size of its two main components (region and style), given in million words:

	Hungary	Slovakia	Transylvania	Subcarpathia	Vojvodina	SUM
press	350.5	11.6	0.6	0.7	1.5	**364.8**
personal	300.3	–	0.4	0.4	0.1	**301.1**
science	112.0	3.3	1.6	0.7	0.3	**117.9**
official	98.0	0.2	0.6	0.3	0.1	**99.0**
literature	77.0	2.3	0.8	0.4	0.2	**80.6**
spoken	76.2	–	–	–	–	**76.2**
SUM	**1013.9**	**17.3**	**3.9**	**2.5**	**2.0**	**1039.7**

The corpus is tokenized, lemmatized and POS-tagged, but no higher level annotations are available in it.

4. Getting started

The HGC is stored as a collection of XML files containing metadata as XML tags and tokenized text in vertical format. These files had to undergo several processing steps, some of which required TSV format.

Conversion from XML to TSV:

1. add unique IDs to the 3 main structural tags

```
<div type="article" id="div_3">
    <p id="p_1">
        <s id="s_1">
            Hamarosan
            fordulat
            jöhet
            a
            magyar
            lakáspiacon
            <g/>
            .
        </s>
```

2. keep the IDs in each token line: the metadata remains accessible via the IDs

```
div_3  p_1  s_1   Hamarosan     " "
div_3  p_1  s_1   fordulat      " "
div_3  p_1  s_1   jöhet         " "
div_3  p_1  s_1   a             " "
div_3  p_1  s_1   magyar        " "
div_3  p_1  s_1   lakáspiacon   ""
div_3  p_1  s_1   .             "\n"
```

5. Corpus cleaning

Four types of text were removed:

1. **sentences longer than 500 tokens** – an upper limit on sentence length had to be set, as the parsing of extremely long sentences might cause parsers to crash
2. **duplicate paragraphs**, using Onion [6] – in the *personal* subcorpus, several paragraphs occurred hundreds of times due to insufficient pre-processing
3. **non-Hungarian paragraphs**, using LangID [4] – if the paragraph contained at least 6 tokens, else it was labeled as Hungarian
4. **paragraphs where all diacritics are missing** – it happens often that Hungarians use English keyboard for typing in Hungarian; diacritic restoration tools could be applied to such texts, but they have a hard time dealing with short and frequent words, e.g. *el* ('away') – *él* 'live', *le* ('down') – *lé* 'juice'

6. Corpus parsing

Two scenarios were considered:

1. parsing the corpus **exclusively with Stanza** [7]
2. using `emtsv` [3] up to POS-tagging **and Stanza** only for dependency parsing

`emtsv` outperforms Stanza in the quality of morphological analysis and lemmatization → feeding the output of `emtsv` into the input of Stanza gives better results

7. The final treebank

The treebank consists of 928.69 million tokens. The original metadata of the HGC are preserved. In addition, the year of publication could be obtained for 93.8% of the texts (871.11 million tokens).

The source files are formatted according to the **CoNLL-U standard** [1]. An excerpt from the treebank (comment lines and the first 8 columns) is shown below:

```
# div_id = 3
# par_id = 1
# sent_id = 1
# text = A hallgatók kötelesek részt venni az előadásokon.
1   A           a           DET   [/Det|Art.Def]   Definite=Def|…   2   det
2   hallgatók   hallgató    NOUN  [/N][Pl][Nom]    Case=Nom|…       3   nsubj
3   kötelesek   köteles     ADJ   [/Adj][Pl][Nom]  Case=Nom|…       0   root
4   részt       rész        NOUN  [/N][Acc]        Case=Acc|…       5   obj
5   venni       vesz        VERB  [/V][Inf]        VerbForm=Inf|…   3   xcomp
6   az          az          DET   [/Det|Art.Def]   Definite=Def|…   7   det
7   előadásokon előadás     NOUN  [/N][Pl][Supe]   Case=Sup|…       5   obl
8   .           .           PUNCT [Punct]          _                3   punct
```

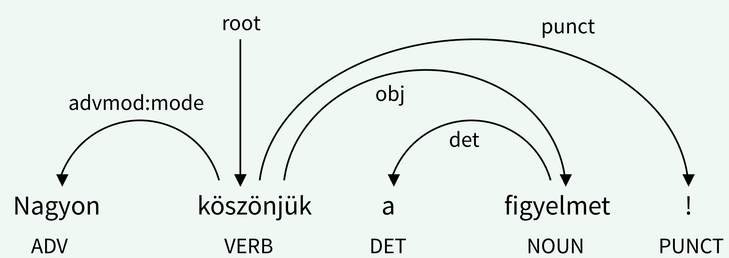
8. How to access the treebank

Demo site using NoSketchEngine [8]: https://corpus.rilex.nytud.hu

The annotation of each word is augmented with the features of its head, thus *one* dependency relationship can be queried easily. After running the queries below, a frequency list on the *lemma* attribute should be created to see the desired result.

- Adjective modifiers of *lakás* 'flat':
  `[head_lemma="lakás" & deprel="amod:att"]` → *new*, *own*, *large* etc.
- Objects of *köszön* 'thank':
  `[head_lemma="köszön" & deprel="obj"]` → 2nd hit: *figyelem* 'attention'
- Take part in what?
  `[head_lemma="vesz" & xpos=".*Sup.*"]`
  `+ filter -5..5 [word="részt"]` → 22nd hit: *előadás* 'lecture'

Thank you for your attention.



[1] de Marneffe, Marie-Catherine et al. "Universal Dependencies". In: *Computational Linguistics* 47.2 (2021), pp. 255–308. URL: https://doi.org/10.1162/coli%5C_a%5C_00402.

[2] Demszky, Dorottya. "The role of verb semantics in Hungarian verb-object order". In: *Proceedings of the Linguistic Society of America*. 2021 Annual Meeting of the LSA. Vol. 6. Online: LSA, 2021, pp. 54–68.

[3] Balázs Indig et al. "One format to rule them all - The emtsv pipeline for Hungarian". In: *Proceedings of the 13th Linguistic Annotation Workshop*. LAW 13. Florence, Italy: ACL, 2019, pp. 155–165.

[4] Marco Lui and Timothy Baldwin. "langid.py: An Off-the-shelf Language Identification Tool". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. ACL 2012. Jeju, Republic of Korea: ACL, 2012, pp. 25–30.

[5] Oravecz, Csaba, Váradi, Tamás, and Sass, Bálint. "The Hungarian Gigaword Corpus". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. LREC 2014. Reykjavik, Iceland: ELRA, 2014, pp. 1719–1723.

[6] Jan Pomikálek. "Removing boilerplate and duplicate content from web corpora". PhD thesis. Brno, Czech Republic: Masaryk University, Faculty of Informatics, 2011.

[7] Qi, Peng et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*. ACL 58. Online: ACL, 2020, pp. 101–108.

[8] Rychlý, Pavel. "Manatee/Bonito - A Modular Corpus Manager". In: *Proceedings of Recent Advances in Slavonic Natural Language Processing*. RASLAN 2007). Brno, Czech Republic: Masaryk University, 2007, pp. 65–70.

[9] Vincze, Veronika et al. "Hungarian Dependency Treebank". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. LREC'10. Valletta, Malta: ELRA, 2010, pp. 1855–1862.