Universal Feature-based Morphological Trees

Federica Gamba, Abishek Stephen, Zdeněk Žabokrtský

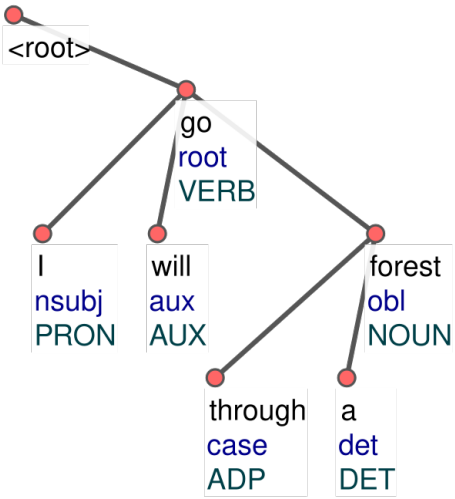
May 25, 2024

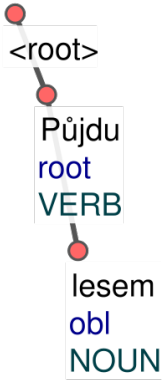
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

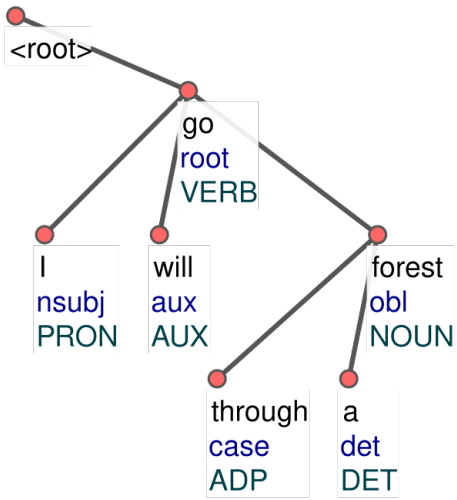
Outline - An Overview

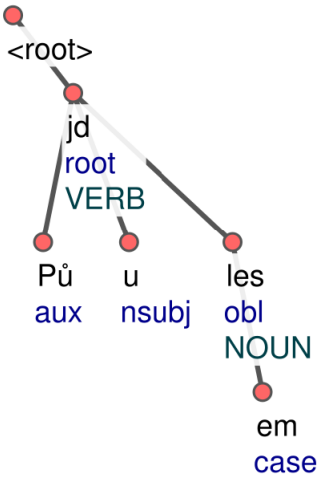
Comparability of Trees I





Comparability of Trees II





Features and Morphs

```
14  ten      ten      NUM    CD   NumForm=Word|NumType=Card      15  nummod  15:nummod  _
15  letters  letter   NOUN   NNS  Number=Plur                    13  obj     13:obj     _
16  from     from     ADP    IN   _                              18  case    18:case    _
17  the      the      DET    DT   Definite=Def|PronType=Art      18  det     18:det     _
18  kings    king     NOUN   NNS  Number=Plur                    15  nmod    15:nmod:from  _
19  of       of       ADP    IN   _                              20  case    20:case    _
20  Gezer    Gezer    PROPN  NNP  Number=Sing                    18  nmod    18:nmod:of  _
21  swearing         swear  VERB VBG  VerbForm=Ger              18  acl     18:acl     _
22  loyalty  loyalty  NOUN   NN   Number=Sing                    21  obj     21:obj     _
23  to       to       ADP    IN   _                              26  case    26:case    _
24  the      the      DET    DT   Definite=Def|PronType=Art      26  det     26:det     _
25  Egyptian         Egyptian      ADJ  JJ   Degree=Pos         26  amod    26:amod  Proper=True
26  pharaoh  pharaoh  NOUN   NN   Number=Sing                    21  obl     21:obl:to  SpaceAfter=No
27  .        .        PUNCT  .    _                              13  punct   13:punct   _
```

Outline - Exploited Resources

UniSegments

- **UniSegments** (Žabokrtský et al., 2022): collection of harmonized versions of 17 segmentation resources covering 32 languages.

Language	Resource
Czech	DeriNet
English	MorphoLex
French	Demonette
Italian	DerIvaTario
Latin	WordFormationLatin
Catalan	MorphyNet
Finnish	MorphyNet
German	MorphyNet
Hungarian	MorphyNet
Portuguese	MorphyNet

UniMorph and SIGMORPHON data

- **UniMorph** (McCarthy et al., 2020): collection of morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema.

UniMorph and SIGMORPHON data

- **UniMorph** (McCarthy et al., 2020): collection of morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema.

- **SIGMORPHON**: manually annotated Czech dataset made available for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022).

Universal Dependencies

- **UD** (de Marneffe et al., 2021): selected treebanks from version 2.12.

Language	Treebank
Czech	PUD, PDT
English	PUD, GUM
Finnish	PUD, TDT
French	PUD, GSD
German	PUD, GSD
Italian	PUD, ISDT
Portuguese	PUD, Bosque

Language	Treebank
Catalan	AnCora
Hungarian	Szeged
Latin	ITTB

Outline - Workflow

Manipulation of Nodes I

0. In **SIGMORPHON** data?
   - 0.1 N: continue to 1
   - 0.2 Y: segment and quit

Manipulation of Nodes I

0. In **SIGMORPHON** data?
   - 0.1 N: continue to 1
   - 0.2 Y: segment and quit
1. **Lemma** segmented in UniSegments?
   - 1.1 N: cs. *rok* → *rok*; continue to 3
   - 1.2 Y: cs. *prokonzul* ‘proconsul’ → *pro* + *konzul*; continue to 2

Manipulation of Nodes II

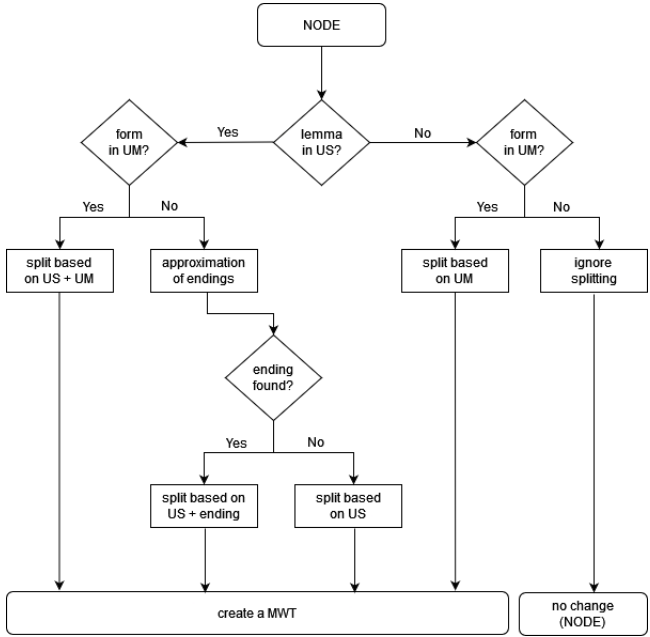
2. **Inflected** form of a segmented lemma?
   - 2.1 N: cs. *prokonzul*, *rok*
   - 2.2 Y: S2.1 Form in **UniMorph**?

Manipulation of Nodes II

2. **Inflected** form of a segmented lemma?
   - 2.1 N: cs. *prokonzul*, *rok*
   - 2.2 Y: S2.1 Form in **UniMorph**?
     - 2.2.1 N: approximation of inflectional ending by string comparison
       en. *shortened* → *short* + *en* (US) + *ed* (string comparison)
     - 2.2.2 Y: ca. *culturals*: *cultur* + *al* (US) + *s* (UM)

Manipulation of Nodes II

2. **Inflected** form of a segmented lemma?
   - 2.1 N: cs. *prokonzul, rok*
   - 2.2 Y: S2.1 Form in **UniMorph**?
     - 2.2.1 N: approximation of inflectional ending by string comparison
       en. *shortened* → *short* + *en* (US) + *ed* (string comparison)
     - 2.2.2 Y: ca. *culturals*: *cultur* + *al* (US) + *s* (UM)
3. Unsegmented lemma, form inflected in UM?
   - 3.1 N: la. *caelum* → *caelum*; no splitting
   - 3.2 Y: fr. *travaillait* → *travailler* + *ait*


NODE
lemma in US?
Yes
No
form in UM?
form in UM?
Yes
No
Yes
No
split based on US + UM
approximation of endings
split based on UM
ignore splitting
ending found?
Yes
No
split based on US + ending
split based on US
create a MWT
no change (NODE)

Feature Extraction I

- *gyerekek* 'children': `Number=Plur|Case=Nom`
- *gyerek* + *ek* → *gyerek*: `Number=Plur|Case=Nom`
  ek: `Number=Plur|Case=Nom`
- *gyerek*: `Number=Plur`
  gyerek: `Case=Nom`
  ek: `Number=Plur`
  ek: `Case=Nom`
- **△P scores** (Jenkins and Ward, 1965) as a measure of cue validity, i.e. measuring how strongly two events are linked.

$$\Delta P_{forward} = P(m|f) - P(m|\neg f) \quad (1)$$

$$\Delta P_{backward} = P(f|m) - P(f|\neg m) \quad (2)$$

Feature Extraction II

Morph	Feature	ΔP forward	ΔP backward
ek	Case=Nom	-0.006	-0.146
ek	Number=Sing	-0.033	-0.431
ek	Person=3	0.031	0.427
ek	Definite=Ind	0.026	0.328
ek	PronType=Ind	0.064	0.099
ek	Mood=Ind	0.030	0.340
ek	Tense=Pres	0.032	0.344
ek	VerbForm=Fin	0.028	0.333
ek	Voice=Act	0.028	0.333
ek	**Number=Plur**	**0.163**	**0.531**

Morph	`Number=Plur`	ΔP forward	ΔP backward
tunk	**1**	**0.033**	**0.972**
ok	**7**	**0.232**	**0.852**
ak	5	0.165	0.690
ek	5	0.163	0.531
ai	1	0.033	0.972

Conforming to UD

- **Lemma**:
  - info about morpheme in US (if available).
    la. *averto* 'to turn away' → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.

Conforming to UD

- **Lemma**:
  - info about morpheme in US (if available).
    la. *averto* 'to turn away' → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.
- **POS**:
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): `X`.

Conforming to UD

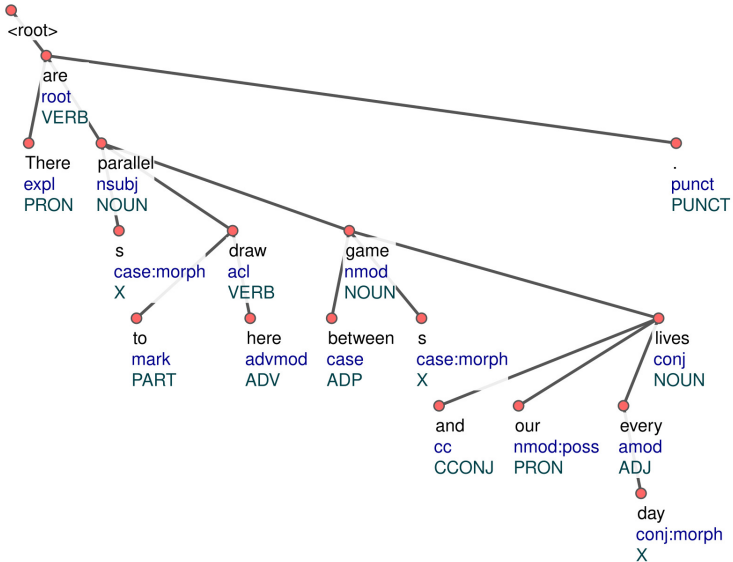
- **Lemma**:
  - info about morpheme in US (if available).
    la. *averto* 'to turn away' → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.
- **POS**:
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): `X`.
- **Features**: feature-based alignment.

Conforming to UD

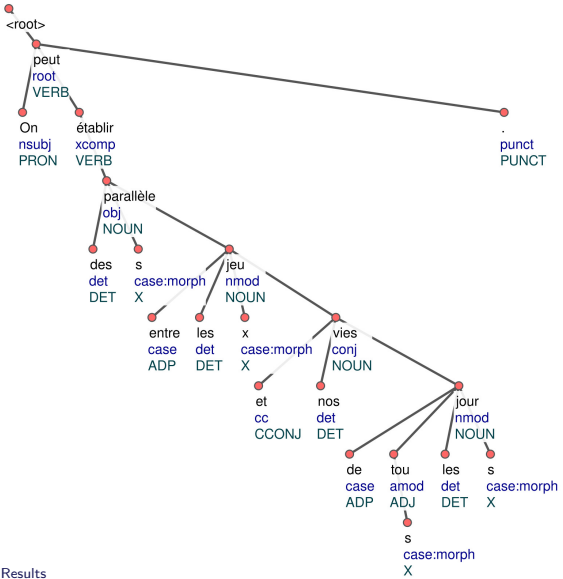
- **Lemma**:
  - info about morpheme in US (if available).
    Ia. *averto* 'to turn away' → *a* + *verto*; morph *a* associated to morpheme *a(b)*.
  - else, lemma = form.
- **POS**:
  - head of MWT (stem): POS of the manipulated node.
  - other tokens of MWT (morphs): `X`.
- **Features**: feature-based alignment.
- **Deprel**:
  - Prefixes: `nmod:morph` if NOUN/PROPN, else `advmod:morph`.
  - If single root: deprel of the manipulated node; `conj:morph` for the second (or +).
  - Suffixes:
    - `aux:morph` for VERBs and AUXs.
    - `case:morph` for NOUNs, PROPNs, ADJs, DETs, PRONs, ADVs, NUMs, very rare ADPs.
    - else `dep:morph`.

Outline - Results

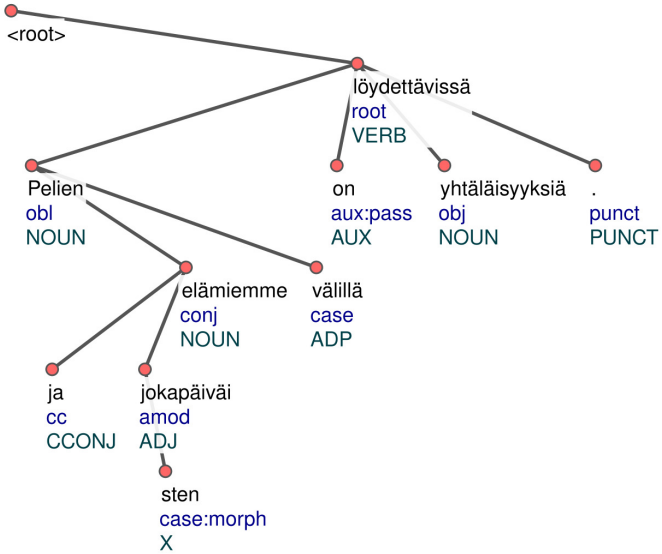
Morphological Trees I



Morphological Trees II



Morphological Trees III



CoNLL-U Representation

```
1       There   there   PRON    EX      _       2       expl
2       are     be      VERB    VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        0
3-4     parallels       _       _       _       _       _
3       parallel        parallel        NOUN    NNS     _       2       nsubj
4       s       s       X       _       Number=Plur     3       case:morph
5       to      to      PART    TO      _       6       mark    5:mark  _
6       draw    draw    VERB    VB      VerbForm=Inf    3       acl
7       here    here    ADV     RB      PronType=Dem    6       advmod
8       between between ADP     IN      _       9       case
9-10    games   _       _       _       _       _       _
9       game    game    NOUN    NNS     _       3       nmod
10      s       s       X       _       Number=Plur     9       case:morph
11      and     and     CCONJ   CC      _       15      cc
12      our     we      PRON    PRP$    Number=Plur|Person=1|Poss=Yes|PronType=Prs      15      nmod:poss
13-14   everyday        _       _       _       _       _       _
13      every   every   ADJ     JJ      _       15      amod
14      day     day     X       _       Degree=Pos      13      conj:morph
15      lives   life    NOUN    NNS     Number=Plur     9       conj
16      .       .       PUNCT   .       _       2       punct
```

To Sum Up

- Novel data structure:
  - Integration of the morphological internal structure of words into a UD-like sentence representation.
  - To enhance comparability of languages that express comparable meaning through different grammatical strategies.
  - Focus on **cross-lingual correspondence of morphs**.
- Case study of 10 languages, leading to a prototype of methodology to manipulate UD treebanks.
- Existing segmentation resources employed:
  - Approach that ties the quality of our data to that of the employed resources.
  - Some limitations observed.

Thank you!

gamba,stephen,zabokrtsky@ufal.mff.cuni.cz

References I

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online, August 2021. ACL. doi: 10.18653/v1/2021.sigmorphon-1.5.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington, July 2022. ACL. doi: 10.18653/v1/2022.sigmorphon-1.11.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August 2013. ACL.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/coli_a_00402.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]. *Traitement automatique des langues*, 60(2):71–95, 2019.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France, May 2020. ELRA.

Nabil Hathout and Fiammetta Namer. Démonette, a French derivational morpho-semantic network. In *Linguistic Issues in Language Technology, Volume 11, 2014-Theoretical and Computational Morphology: New Trends and Synergies*, 2014.

Herbert M. Jenkins and William C. Ward. Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79(1): 1–17, 1965. doi: 10.1037/h0093874. URL `https://doi.org/10.1037/h0093874`.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria, December 2021. ACL.

Eleonora Litta, Marco Passarotti, and Chris Culy. Formatio formosa est. Building a word formation lexicon for Latin. In *CLiC-it/EVALITA*, 2016.

References II

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France, May 2020. ELRA.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August 2013. ACL.

Marco Passarotti. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320, 2019. doi: 10.1515/9783110599572-017.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT, 2015.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017.

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior research methods*, 50:1568–1580, 2018.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102, 2016.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. ELRA.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. DeriNet 2.1, 2021.

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

Veronika Vincze, Katalin Simkó, Zsolt Szántó, and Richárd Farkas. 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 2017. ACL.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. Towards Universal Segmentations: UniSegments 1.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149, Marseille, France, June 2022. ELRA.

Amir Zeldes. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017. doi: http://dx.doi.org/10.1007/s10579-016-9343-x.