Compounds in Universal Dependencies: A Survey in Five European Languages

Emil Svoboda, Magda Ševčíková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{svoboda,sevcikova}@ufal.mff.cuni.cz

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Table of contents

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Introduction

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Introduction

- Compounds are words formed by a combination of two or more {words, bases, roots, stems}.
  - *waterfall*, *černobřichý*, *Jahreabschlussprüfung*, *городской-государство*, *blauäugig*, etc.
  - Compoundhood is difficult to rigidly define, as it borders with
    - Syntax (compounds vs. syntactic phrase)
    - Word formation (compounds vs. neoclassical formations vs. derivatives)
  - It is ***not*** defined by spelling:
    - *flowerpot* (closed compound), *flower-pot* (hyphenated compound), *flower pot* (open compound)
- We survey how compounds are treated in Universal Dependencies.
  - Orthographic conventions dictate tokenization
  - Tokenization dictates dependency building
  - Inter- and intralinguistic comparison is therefore difficult
  - A compound annotation scheme based on existing relations in UD is proposed
    - It explicitly shows the analogy between compounds and phrases
  - {English, German, Czech, Russian, Latin}

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Theoretical background

- Debate around compounding centered around:
  - Compounding vs. derivation — *biology*? *biodegradable*? *understand*?
  - POS of the components — *mother tongue* → [N + N]; *Umfrageteilnehmer* → [V/N? + N]
  - Headedness — *mother tongue* → [N + **(N)**]; *lady-in-waiting* → [**(N)** + N]
  - Endo- vs exo- centricity — *toothbrush* (type of brush); *bluestocking* (unrelated to legwear)
  - Internal structure — *Straßenbahnlinie* → [[[Straße] Bahn] Linie]
  - Component relation — coordinative vs. subordinate/determinative etc.
- Bisetto and Scalise (2006) propose a two-level multilingual classification scheme:
  - Component relation is level 1
    - subordinate: *sunglasses*;
    - attributive: *blue cheese*;
    - coordinative: *actor-director*
  - Centricity is level 2
    - endocentric: *apple cake*; *backyard*; *God Emperor*
    - exocentric: *killjoy*; *white-collar*; *mind-brain*
  - The classification is used in the construction of a database of compounds

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Compounds in language data resources 1/2

- MorboComp (20 languages; 2006):
  - All languages in scope except Czech covered
  - Based on Bisetto and Scalise's classification

Compound	POS	Struc	Class	End	Head-C	Head-S	1st-C	2nd-C	Gloss
madrelingua	N	[N+N]	SUB	Tru	right	right	madre	lingua	mother+tongue
mano lesta	N	[N+A]	ATT	Fal	none	none	mano	lesta	quick+hand=thief
dormiveglia	N	[V+V]	CRD	Fal	none	none	dormi	veglia	sleep+be awake=dozing

- CELEX2 (English, German, Dutch; 2014):
  - General lexical database
  - English: 6,267 compounds; German: 19,304
  - Morphs replaced by a representative form

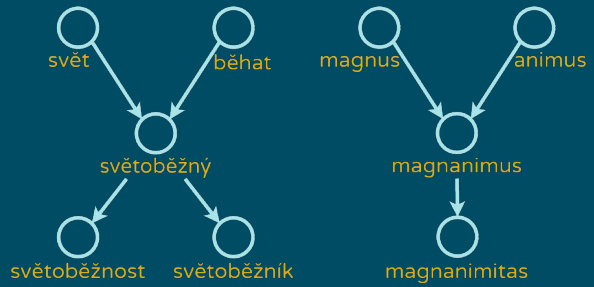
Umgangssprache	...	Umgang+s+Sprache	NxN	...	((((um)[V|.V],(geh)[V])[V])[N],\\(s)[N|N.N],((sprech)[V])[N])[N]	...
Grossmachtpolitik	...	Grossmacht+Politik	NN	...	(((gross)[A],(Macht)[N])[N],\\((polit)[R],(ik)[N|R.])[N])[N]	...
womenfolk	...	women+folk	NN	...	((women)[N],(folk)[N])[N]	...

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Compounds in language data resources 2/2

- GermaNet (German; 2014):
  - 120,000 compounds in its 2023 edition
  - Two closest existing ancestor words listed
- DeriNet 2.1 (Czech; 2006):
  - Lexical database of 432,000 attested lemmas
  - 45,473 compounds
  - Words linked to ancestors → tree-like structure
- Word Formation Latin (Latin; 2006):
  - Lexical database (36,258 entries; 3,198 compounds)
  - Similar structure to DeriNet 2.1
- Golden Compound Analyses (Russian; 2006):
  - 1,699 compounds linked to ancestors + POS
  - Compiled for the training of a compound splitter

Umgangssprache	Umgang	Sprache
Abbiegeassistent	abbiegen	Assistent
Umfrageteilnehmer	Umfrage|umfragen	Teilnehmer



полувсерьёз	adv	половина	noun	всерьёз	adv

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Current treatment of compounds in UD

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Annotation guidelines 1/2

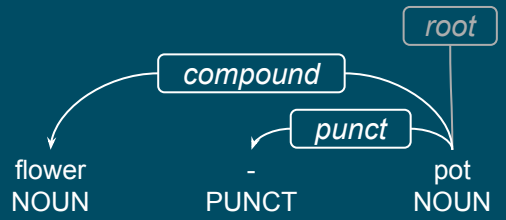
- Guidelines follow tokenization
  - Closed compounds treated as atomic (discrete, internally unstructured)
    - not distinguished from other words at all → amount in UD difficult to establish
    - usage of data sources allow us to present a **lower bound estimate**

Language	Closed compounds	Total words	Sentences with closed compounds	Total sentences
English	5,934 (0.82%)	726K	5,286 (11.57%)	46K
German	156,629 (4.11%)	3,810K	87,104 (50.14%)	208K
Czech	47,103 (2.11%)	2,222K	34,775 (27.27%)	128K
Latin	26,271 (2.62%)	983K	18,353 (31.27%)	59K
Russian	4,803 (0.27%)	1,830K	4,460 (4.00%)	111K

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Annotation guidelines 2/2

- Guidelines follow tokenization
  - Open and hyphenated compounds treated as subtrees
    - head component as root node
    - the rest as dependent node(s)

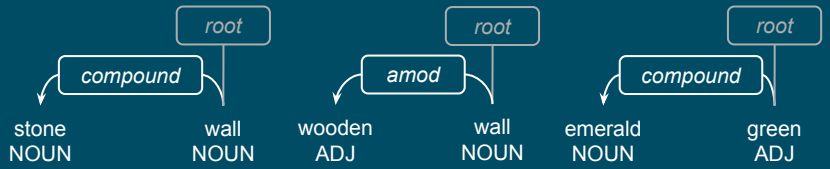


Language	*compound* relations	Sentences with *compound*	*compound:prt* relations	Sentences with *compound:prt*	Total words	Total sentences
English	22,017 (3.03%)	13,459 (29.27%)	2,485 (0.34%)	2,313 (5.0%)	726K	46K
German	1,787 (0.05%)	1,418 (0.68%)	22,349 (0.59%)	21,897 (10.5%)	3,810K	208K
Czech	2,690 (0.12%)	1,356 (1.06%)	0 (0.00%)	0 (0.0%)	2,222K	128K
Latin	85 (0.01%)	82 (0.1%)	0 (0.00%)	0 (0.0%)	983K	59K
Russian	1,973 (0.11%)	1,812 (1.6%)	0 (0.00%)	0 (0,0%)	1,830K	111K

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Compounds in UD treebanks - a closer look

English



German

Czech

Latin

Russian

Summary

- The *compound* relation is used inconsistently
- It is underspecified
- Interesting typological phenomena are lost

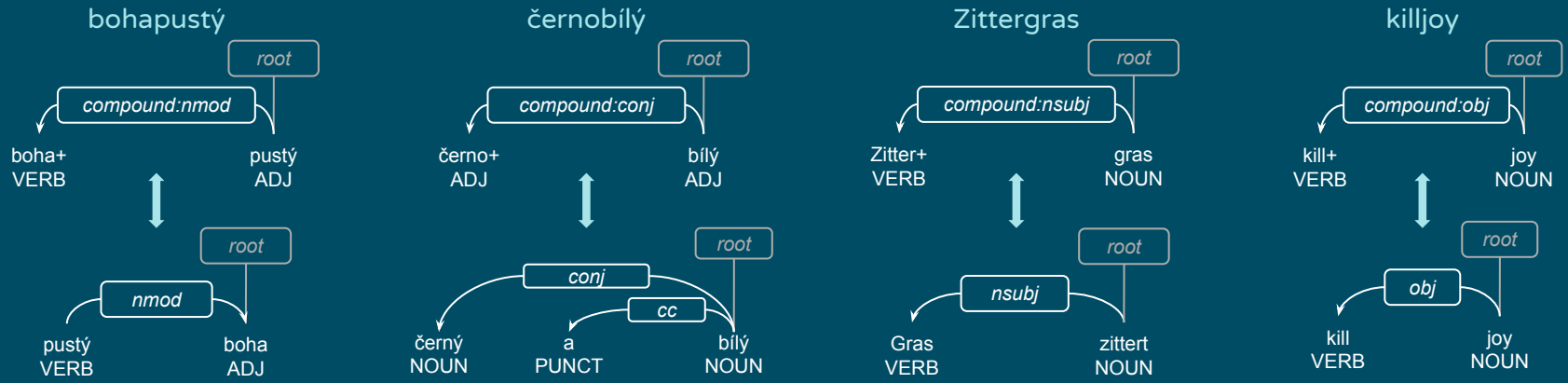
This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Syntax-based annotation proposal

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Proposal description (1/2)

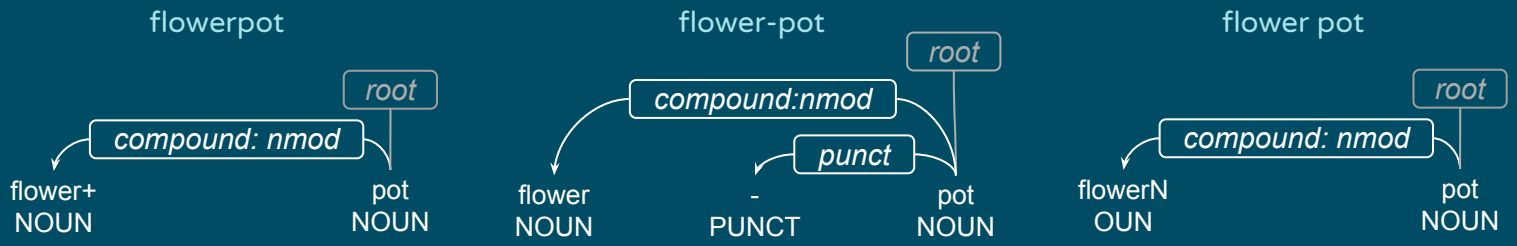
- We want to unify the treatment of compounds regardless of spelling:
  a. Closed compounds should be split
  b. Compound constituents should be assigned a lemma
  c. Compounds should be a subtree with a *compound: <relation>* tag



This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Proposal description (2/2)

- We want to unify the treatment of compounds regardless of spelling:
  a. Closed compounds should be split
  b. Compound constituents should be assigned a lemma
  c. Compounds should be a subtree with a *compound: <relation>* tag



This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Steps toward implementation

- Identification of closed compounds
  - No. of closed compounds varies – 50% of sentences in German vs. 4% in Russian
  - Higher-coverage resources or more sophisticated tools may reveal higher numbers
- Splitting of closed compounds and component lemmatization
  - Previously mentioned data sources provide a starting point
  - *PaReNT* provides splitting + component lemmatization for Czech
- Assigning syntactic relation labels
  - Limited data available (CELEX)
  - Pilot procedure based around finding phrases encoded by each example and feeding them into UDPipe
  - Possible pipeline:
    - *PaReNT*: *černobílý* → *černo+bílý* → *černý, bílý*
    - *ChatGPT/LLM*: černý, bílý → *černý a bílý*
    - *UDPipe: černý a bílý* → *cconj*

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Conclusion

- We explored the current treatment of compounds in UD.
  - We used English, German, Czech, German, Russian, and Latin
  - The handling of compound varies widely
- We propose that compound should be treated analogously to syntactic phrases.
  - Compounds should be treated consistently regardless of spelling
  - Existing relations should be used to create *compound: <relation>* tags
- We are currently implementing the proposed scheme.
  - A pipeline based on existing tools is being developed

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, Jonáš Vidra. Universal Derivations Kickoff: A Collection of Eleven Harmonized Derivational Resources for Eleven Languages. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*. Prague: Charles University. ISBN: 978-80-88132-08-0. 2022.

Emil Svoboda & Magda Ševčíková. Spliting and Identifying Czech Compounds: A Pilot Study. In *Proceedings of the Third Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*. France, 2021, pp. 125-134.

Emiliano Guevara, Sergio Scalise, Antonietta Bisetto, and Chiara Melloni. 2006. MORBO/COMP: A Multilingual Database of Compound Words. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation*, pages 2160–2163.

Emil Svoboda & Magda Ševčíková. *Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes*. The Prague Bulletin of Mathematical Linguistics. Prague, April 2022, 118 pp. 55–73.

Jonáš Vidra et al. DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the Second*

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, Lukáš Kyjánek. 2019. DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*. Prague: Charles University. ISBN: 978-80-88132-08-0.

Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 2014. *CELEX2 LDC96L14*, 1995. URL https://doi. org/10, 35111:3

Antonietta Bisetto and Sergio Scalise. 2005. *The classification of compounds.* Lingue e linguaggio,4(2):319–0.

Ivana Bozděchová. 1997. *Tvoření slov skládáním.* Institut sociálních vztahů, Praha.

Verena Henrich and Erhard Hinrichs. 2010. *GernEdiT - the GermaNet editing tool.* In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Sanjeet Khaitan, Arumay Das, Sandeep Gain, and Adithi Sampath. 2009. Data-Driven Compound Splitting Method for English Compounds in Domain Names. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, page 207–214, New York, NY, USA. Association for Computing Machinery.

Daniil Vodolazsky and Hermann Petrov. 2021. Compound Splitting and Analysis for Russian. *Resources and Tools for Derivational Morphology (DeriMo 2021)*, page 149

Daniel Zeman et al. 2021. *Universal Dependencies 2.12*. 2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

Thank you for your attention!

This research was supported by the Grant Agency of Charles University, project GA UK 128122. It uses data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz) supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).