LREC 2022, Marseille

Towards Universal Segmentations: UniSegments 1.0

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek,
Emil Svoboda, Magda Ševčíková, Jonáš Vidra

📅 June 20-25, 2022

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

Outline

Introduction

Diversity in the Existing Resources

Our Harmonized Scheme and the Resulting Collection

Conclusions

Introduction

Basic Notions

- **Morphemes** are the smallest units of language that have a meaning.
  - E.g., *play+er+s*
  - They work as basic building blocks in various inflectional and word-formation processes.

- Types of morphemes:
  - root morphemes (convey lexical meaning)
  - prefixes : *re+play*
  - suffixes (incl. endings) : *teach+er*
  - interfixes (in compounds) : *speed+o+meter*

- **Allomorphy** – a morpheme can be possibly expressed with multiple different **morphs**.
  - E.g., *sheep vs. shep* (in *shepherd*)
  - Homonymy is possible:
    - E.g., *bear+s* (noun + plural *vs.* verb + 1^{st} person in singular)

Motivation for Harmonization Efforts

- Morpheme is a central linguistic notion, but – surprisingly – not properly substantiated in modern NLP, cf. Byte Pair Encoding.
- There are various data resources that are directly or indirectly related to morphological/morphemic segmentation.
- Different annotation schemes are applied in different resources.
- It is very difficult to perform e.g. multilingual/cross-lingual experiments.

- **Our goal**:
  to provide morpho-segmentation datasets for various languages in the same format.
  - Inspiration: the success story of UNIVERSAL DEPENDENCIES.

Diversity in the Existing Resources

Overview of Resources Included in Our Study

Abbreviated name	Original name, version	Languages	License
CroDeriV	CroDeriV 1.0	Croatian	CC BY-SA-3.0
Démonette	Démonette-1.2	French	CC BY-NC-SA 3.0
DeriNet	DeriNet 2.1	Czech	CC BY-NC-SA 3.0
DerIvaTario	DerIvaTario	Italian	CC BY-SA 4.0
DerivBaseDE	DErivBase 2.0	German	CC BY-SA 3.0
DerivBaseRU	DerivBase.Ru 1.0	Russian	Apache-2.0
Échantinom	Échantinom	French	CC BY 4.0
KCIS	KCIS Resources	Marathi, Hindi, Malayalam, Kannada, Bangla	CC BY-NC 4.0
MorphoLex	MorphoLex, MorphoLex-FR	English and French	CC BY-NC-SA 4.0
MorphyNet	MorphyNet v1	15 languages	CC BY-SA 3.0
PerSegLex	Persian Morph. Segmented Lexicon 0.5	Persian	CC BY-NC-SA 4.0
Uniparser	Uniparser morphological analyzer	7 languages	MIT License
WordFormationLatin	Word Formation Latin 1.1	Latin	CC BY-NC-SA 4.0
CELEX	CELEX Lexical Database 2.0	Dutch, English, German	non-free
KuznetsEfremDict	Dictionary of Morphemes of Russian	Russian	non-free
MorphoChallenge	MorphoChallenge 2005, 2007-2010	English, Finnish, German, Turkish, (Arabic)	non-free
TikhonovDict	Morphemic-spelling dict. of Russian	Russian	non-free

Crucial Differences among the Resources

Selection of the original lexical material

- Are **word forms or lemmas** segmented?
- Do they originate from **pre-existent lexicons or corpus based frequency lists**?
- What is the **distribution across POS categories**?
- **How many** [units, segments, ...] is processed?

Nature of segments

- Morphs : mostly delimited as contiguous sequences of characters.
- Morphemes : 3 different solutions:
  1. using a selected representative allomorph
  2. referring to the citation form of the base word
  3. a fully abstract unit, without mentioning any form (e.g., *dogs* → *dog* + *PL*)
- Both : possibility of hierarchical segmentation like in the Context-free Grammars.

Overview of the Original Resources

Resource	Number of segmented units:	POS categories:	Segmentation origin:	Segment info:		Completeness of segmentat.:	Classification of segments:	Zero morph.:	Hierarch. segm.:
	k = ×1,000, L = lemmas, W = word forms	N = noun, A = adjective, V = verb, D = adverb, O = other	M = manual, A = automatic	morphs or morpheme (or both)		C = complete, P = partial, S = single affix	T = stem, R = root, P = prefix, I = interfix, S = suffix, E = ending		
CroDeriV	16 kL	V	M	✓	–	C	R, P, S, E	✓	–
Démonette	42 kL	N, V, A	M + A	✓	–	S	T, S	–	✓
DeriNet	1,039 kL	N, A, D, V, O	M + A	✓	✓	C	R, P, S	–	✓
DerIvaTario	11 kL	N, A, V, O	M	–	✓	C	R	✓	✓
DerivBaseDE	61 kL	N, A, V	A	✓	–	S	P, S	–	✓
DerivBaseRU	156kL	N, V, A, D, O	A	✓	–	S	P, S, E	–	✓
Échantinom	5 kL	N	M	✓	–	S	R, P, S	–	–
KCIS	avg. 26 kW	N, V, O, A, D	M + A	–	✓	P	R, S	–	–
MorphoLex	avg. 43 kW	N, V, A, D, O	M	–	✓	C	R, P, S	–	–
MorphyNet	362 kW+kL	N, A, V, D, O	M + A	✓	–	S	R, P, S	–	–
PerSegLex	8 kW	–	M	✓	–	C	–	–	✓
Uniparser	avg. 277 kW	N, A, V, D, O	A	✓	–	P	T, P, S	✓	–
WordFormationLatin	36 kL	N, A, V, D, O	M + A	–	✓	P	R, P, S	–	✓
CELEX	avg. 77 kL	N, A, V, O, D	M	–	✓	C	R, P, I, S	✓	✓
KuznetsEfremDict	73 kL	N, V, A, D, O	M	✓	–	C	R	–	–
MorphoChallenge 2005	avg. 1 kL	–	M + A	✓	–	C	–	–	–
— 2007-2010	avg. 2.5 kL	–	M + A	✓	✓	C	–	–	–
TikhonovDict	103 kL	–	M	✓	–	C	–	–	–

Our Harmonized Scheme and the Resulting Collection

Scheme & Conversion

Basic design choices

- Segmentation to morphs is considered as primary.
- A simplifying assumption: words are fully decomposable into morphs (without overlaps).
- We unify POS category values.
- A simple line-oriented, five-column file format is used; e.g., Croatian *to scratch*.
  1. word form e.g., *podrapati*
  2. lemma e.g., *podrapati*
  3. part-of-speech category e.g., *VERB*
  4. simplified morphological segmentation e.g., $po + drap + a + ti$
  5. detailed annotations of indices and types of individual morphological segments (`JSON`)

Resource-specific conversion issues

- Aligning morphs and morphemes
- Making partial segmentation (more) complete

Conversion Examples

Ex.	Resource	Data samples in their original formats		UniSegments 1.0
1	CELEX	22845 \Leuchtbombe\1\C\1\Y\Y\Y\Leuchte+Bombe\NN\N\N\N\ (((licht)[A],(e)[N|A.])[N],(Bombe)[N])[N]\Y\N\N\N\S3/P3\N	→	Leucht + bombe (*photoflash bomb*)
2	CELEX	5290\brinksmanship\0\C\\1\N\N\N\N\Y\brink+s+man+ship\NxNx\SASA\N\N\Y\ ###\N\N\SASA\((brink)[N],(s)[N|N.Nx],(man)[N],(ship)[N|NxN.])[N]\N\N\Y	→	brink + s + man + ship (*brinksmanship*)
3	Démonette	"abaissement","tlfnome","abaisser","tlfnome","Ncms","tlfnome","Vmn—-","tlfnome","simple", "derif","suf","ment","derif",,,,"RES","demonette","","demonette","résultat de abaisser","derif", "résultat de ","demonette","descendant","demonette","abaiss","derif",,,"derif"	→	abaiss + e + ment (*lowering*)
4	DerIvaTario	3951;ABBATTIMENTO;BATTERE:vrb_th;ACons:ad:mt2:ms2b;MENTO:mento:mt4:ms1;;;;	→	ab + batt + i + mento (*breakdown*)
5	DerIvaTario	15744;CADENZAMENTO;CADERE:vrb_th;NZA:nza:mt1:ms2b;CONVERSION:N_V; MENTO:mento:mt1:ms1;;;	→	cade + nza + mento (*cadence*)
6	DerivBaseDE	Großstadt_Nf Großstädterin_Nf 2 Großstadt_Nf dNN05:(sfx "er" & opt uml & try (rsfx "er" "r" .||. dsfx "e" .||. opt (dsfx "en" .|. rsfx "en" "n") .||. try (dsfx "ien" .|. rsfx "ien" "i")) & try (rsfx "ia" "i") & opt (rsfx "a" "i")) nouns mNouns> Großstädter_Nm dNN02:(sfx "in" & try (dsfx "e")) nouns nouns> Großstädterin_Nf	→	Großstädt + er + in (*female city dweller*)
7	DerivBaseRU	вымор noun повыморить verb rule887(по + noun + и1(ть) -> verb) PFX,SFX	→	по + вымори + ть (*become extinct*)
8	Échantinom	alpiniste,m,al.pi.nist,1.49 1.96,5819,suffix,suffix,0,0,0,iste,iste,alpin,A,TRUE,alpin,ist,alpin,0,_~_ist, 53,0.569892473,0.4425928,0.454843023	→	alpin + iste (*alpinist*)
9	MorphoChallenge	act:act_V ion:ion_s s:+PL	→	act + ion + s (*actions*)

Resulting Collection

Universal Segmentations 1.0 includes 47 datasets for 32 different languages.

Public edition

- 13 harmonized resources whose original licenses were free enough
- available in the LINDAT/CLARIAH-CZ repository

Internal edition

- +4 resources which we are not allowed to distribute further due to license limitations
- we published the conversion scripts

Statistical Properties (15 out of 47 datasets)

		Distribution of morphs per unit [%]				Mean morphs per unit	Mean unit length [char]	Mean morph length [char]
Resource name	**Size**	**1**	**2**	**3**	**4+**			
deu-DerivBaseDE	61 kL	36	59	4	0	1.7	11.2	6.6
deu-MorphoChallenge	3 kL	4	27	42	27	3.0	10.5	3.5
deu-MorphyNet	29 kL	0	100	0	0	2.0	10.6	5.1
eng-CELEX	44 kL	30	51	16	3	1.9	8.6	4.5
eng-MorphoChallenge	3 kL	16	49	27	9	2.3	8.4	3.7
eng-MorphoLex	69 kW	21	45	27	7	2.2	8.3	3.8
eng-MorphyNet	292 kL	0	100	0	0	2.0	10.7	5.1
fra-Démonette	63 kL	46	80	3	0	1.7	9.9	5.9
fra-Échantinom	5 kL	53	40	6	1	1.5	7.8	5.1
fra-MorphoLex	16 kW	43	44	12	1	1.7	8.2	4.7
fra-MorphyNet	363 kL	0	100	0	0	2.0	10.7	5.1
rus-DerivBaseRU	156 kL	31	35	23	10	2.1	10.3	4.8
rus-KuznetsEfremDict	73 kL	1	7	17	75	4.3	9.9	2.3
rus-MorphyNet	692 kL	0	100	0	0	2.0	10.5	5.1
rus-TikhonovDict	103 kL	6	11	22	61	3.8	10.2	2.7

Conclusions

Conclusions & Future Work

Our Contribution

- We surveyed 17 existing data resources relevant for morphological segmentation and identified their similarities and differences.
- We designed a common annotation scheme.
- We converted the resources into the scheme.
- We released a subset of the harmonized resources publicly.

Future Work

- To harmonize more resources, including resources which deal with segmentation only very indirectly, such as UniMorph.
- If multiple resources available for the same language, to merge them.
- To develop multilingual segmentation tools.

Thank you!

If interested in Universal Segmentations, please have a look at

http://ufal.cz/universal-segmentations

where you will find:

- a link to the UniSegments 1.0 data on LINDAT/CLARIAH-CZ
- a comprehensive technical report on the existing resources
- future publications and presentations related to Universal Segmentations

Acknowledgement

We would like to thank all the authors of the original resources and our colleagues from various annotation projects who were so kind to give us access to their datasets, comments and advise on the data and annotation structure.

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Rep. (project LM2018101), and by the SVV project No. 260 575. It was using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project.