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

Basic Notions

- a morpheme – the smallest unit of language that has a meaning
  - *play+er+s*
- basic building blocks in various inflection and word-formation processes
- types of morphemes
  - a root morpheme conveys lexical meaning
  - prefixes
  - suffixes (incl. endings)
  - interfixes (in compounds)
- allomorphy: a morpheme possibly expressed with multiple different morphs
  - *sheep* vs. *shep* (in *shepherd*)
- homonymy possible
  - *bear+s*

Motivation for Harmonization Efforts

- morpheme is a central linguistic notion, but – surprisingly – not properly substantiated in modern NLP
- there are various data resources that are directly or indirectly related to morphological segmentation
- but different annotation schemes applied in different resources
- very difficult to perform e.g. multilingual/crosslingual experiments
- our goal: provide morphosegmentation datasets for various languages in one format
- inspiration: the success story of Universal Dependencies

Diversity of 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[a]	CC BY-SA 3.0
PerSegLex	Persian Morphologically Segmented Lexicon 0.5	Persian	CC BY-NC-SA 4.0
Uniparser	Uniparser morphological analyzer	7 languages[b]	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[c]
KuznetsEfremDict	Dictionary of Morphemes of Russian	Russian	non-free[c]
MorphoChallenge	MorphoChallenge 2005, 2007-2010	English, Finnish, German, Turkish, (Arabic[d])	non-free[c]
TikhonovDict	Morphemic-spelling dictionary of the Russian language	Russian	non-free[c]

Selection of Lexical Material

- lemmas or word forms?
- how many?
- distribution across POS categories?
- lists from pre-existent lexicons, or corpus based frequency lists?

Nature of Segments: Morphs, Morphemes, or Both

- morphs: mostly delimited as contiguous sequences of characters
- morphemes – 3 different solutions:
  1. using a canonically selected representative allomorph
  2. referring to the citation form of the base word
  3. a fully abstract unit, without mentioning any form (e.g. PL)
- possibly hierarchical segmentation

Overview of Resources Included in Our Study

Resource	Number of segmented units: k = ×1,000, L = lemmas, W = word forms	POS categories:[1] N = noun, A = adjective, V = verb, D = adverb, O = other	Segmentation origin: M = manual, A = automatic	Segment info: morphs or morpheme (or both)		Completeness of segmentation: C = complete, P = partial, S = single affix only	Classification of segments: T = stem, R = root, P = prefix, I = interfix, S = suffix, E = ending	Zero morpheme allowed:	Hierarchical segm.:
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	–	✓[n]	P	R, S	–	–
MorphoLex	avg. 43 kW	N, V, A, D, O[b]	M	–	✓	C	R, P, S	–	–
MorphyNet	362 kW+kL	N, A, V, D, O[c]	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	–	–	–
MorphoChallenge 2007-2010	avg. 2.5 kL	–	M + A	✓	✓	C	–	–	–
TikhonovDict	103 kL	–	M	✓	–	C	–	–	–

Our Harmonized Scheme and the Resulting Collection

Basic Design Choices

- segmentation to morphs considered as primary
- a simplifying assumption: words fully decomposable into morphs (without overlaps)
- unified POS values
- a simple line-oriented file format

Resource-Specific Conversion Issues

Examples:

- aligning morphs and morphemes
- making partial segmentation (more) complete

Conversion Examples

Ex.	Resource	Data samples in their original formats		Morph segmentation in 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	KCIS (Marathi)	2.2 हवामानामुळे N_NN <fs af=’हवामान,n,n,sg,,o,मुळे, ◌ा_मुळे’ name=’हवामानामुळे’>	→	हवामान + ◌ा + मळे (*due to the weather*)

Resulting Collection

- divided into two parts two parts
  - public edition – 13 harmonized resources whose original licenses were free enough
    - available in the LINDAT/CLARIAH-CZ repository
  - internal edition – 4 more resources which we are not allowed to distribute further due to license limitations
    - however, we published the conversion scripts
  - altogether 47 datasets for 32 different languages

Statistical Properties

Resource name	Size	Distribution of morphs per unit [%] 1	2	3	4+	Mean morphs per unit	Mean unit length [char]	Mean morph len [char]
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

Conclusions

Our Contribution

- we surveyed 17 existing data resources relevant for morphological segmentation, and identified 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

- if multiple resources available for the same language, to merge them
- to include more resources
- to include also resources which deal with segmentation only very indirectly, such as UniMorph
- to start developing multilingual segmentation tools

Thank you!

If interested in Universal Segmentations, please have a look at

https://ufal.mff.cuni.cz/universal-segmentations

where you will find

- a link to the UniSegments 1.0 data on Lindat/CLARIAH-CZ
- a comprehensive technical report
- future publications and presentations related to UniSegments