

Morphology in multilingual data resources: A brief survey

DOI: <http://dx.doi.org/10.21165/el.v54i1.4031>

Magda Ševčíková¹

Abstract

This paper presents selected multilingual language resources that capture various aspects of morphology. Text corpora, dedicated lexical datasets, and typological databases provide insights into inflection, derivation, and, in some cases, the internal structure of words. With continued improvements in coverage, consistency, and compatibility, these resources have the potential to become a solid basis for a realistic understanding of morphological structures across natural languages.

Keywords: morphology; typology; segmentation; corpora; databases.

¹ Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic; sevcikova@ufal.mff.cuni.cz; <https://orcid.org/0000-0003-4780-7912>

Morfologia em recursos de dados multilíngues: uma breve revisão

Resumo

Este artigo apresenta recursos linguísticos multilíngues selecionados que capturam vários aspectos da morfologia. *Corpora* de texto, conjuntos de dados lexicais dedicados e bancos de dados tipológicos fornecem visão sobre inflexão, derivação e, em alguns casos, a estrutura interna das palavras. Com melhorias contínuas em cobertura, consistência e compatibilidade, esses recursos têm o potencial de se tornar uma base sólida para uma compreensão realista das estruturas morfológicas em todas as línguas naturais.


Palavras-chave: morfologia; tipologia; segmentação; *corpora*; bancos de dados.

1. Introduction

In modern linguistics, the comparison of languages aims not only at revealing similarities and differences between the languages being studied but also to uncover universal features of language as a means of human communication. The number of languages whose structures are compared for the purpose of typological research has grown from dozens to hundreds and thousands over the last six decades; cf. the 30-language sample underlying Greenberg's 1963 universals of grammar and the Universal Dependencies project (de Marneffe *et al.*, 2021) covering more than 160 languages in its current version, or the typological database Grambank (Skirgård *et al.*, 2023) providing selected features for nearly 2,500 languages.

Despite these impressive advances, creating a realistic, data-driven picture of more than seven thousand languages spoken around the world (Eberhard *et al.*, 2025) remains far from achievable. This is not only due to the majority of languages being undocumented but also because even for those that are documented, the information is scattered across various datasets. Data and annotations are often fragmented according to traditional linguistic sub-disciplines such as syntax, morphology, phonology, etc. In the present paper, I will specifically focus on phenomena traditionally classified under morphology in theoretical discussions and provide an overview of the linguistic data resources that capture them.

The paper is structured as follows. Section 2 briefly summons up the broad scope of morphology as a linguistic discipline and its reflection in linguistic data. The following sections are then restricted to datasets that cover a larger number of languages and thus have the potential to be used in linguistic typology. Section 3 introduces morphological annotation in multilingual text corpora, with examples of parallel and non-parallel data. Section 4 gives examples of lexical datasets that contain different types of morphological



information, and in Section 5 typological databases are introduced. A few concluding notes in Section 6 close the paper.

2. Morphology in the theoretical discussion and in language data resources

Morphology “deals with words, their internal structure, and how they are formed” (Aronoff; Fudeman, 2005, p. 2). It is traditionally subdivided into inflectional and derivational morphology. While inflectional morphology involves the modification of words to express grammatical features such as tense, number, or case, without changing the word’s lexical meaning, derivational morphology deals with the creation of new words by adding prefixes, suffixes, or other morphemes to existing words. The scope of morphology can also be broadened to other word-formation processes, esp. conversion, compounding, or blending.

Morphology is positioned between phonology, which deals with how sounds function in languages, and syntax, which examines how words are combined to form sentences. Morphology also has significant ties to semantics and pragmatics. In terms of semantics, morphology plays a crucial role in how meaning is encoded within words. In pragmatics, morphology can influence how meaning is modified in different contexts. The way words are formed and used can depend on social factors, registers, or discourse contexts.

Although linguistic discussions often point out that the disciplines thus defined are merely useful constructs that help grasp the compact linguistic reality, these debates have had only limited influence on the creation of linguistic data. Here, the traditional distinctions are projected into fundamental design decisions: The distinction between derivation and inflection manifests in grouping individual word forms under representative forms (lemmas), the separation of morphology and syntax is reflected in the delineation of layers in corpora with multiple types of linguistic annotation.

3. Morphology in multilingual text corpora

Text corpora are large collections of texts that document words in their natural contexts. Both parallel and non-parallel corpora have been used to compare morphology across languages. While parallel corpora such as the JHU Bible Corpus or the InterCorp project contain equivalent texts in two or more languages that are aligned at the sentence or segment level (Section 3.1), the Universal Dependencies project has developed a unified annotation scheme that is applied to non-parallel texts from individual languages (Section 3.2).

3.1 Parallel corpora

The John Hopkins University Bible Corpus contains more than four thousand translations of the Christian Bible in over 1,600 languages; the texts are verse-aligned. McCarthy *et al.* (2020) summarize Natural Language Processing experiments in which the Bible texts (typically for a specific subset of the languages covered by the corpus) were annotated with part-of-speech tags, dependency relations, or features monitored by the typological database Ethnologue (Eberhard *et al.*, 2025). They also demonstrated the use of unannotated data to calculate the type-token ratio, which was interpreted as an indicator of morphological richness of the languages.

Figure 1. Parallel sentences from the English and Portuguese sections of the InterCorp corpus. The forms of the English verb *smile* are annotated with modified PennTreebank part-of-speech tags (explained in the oval), while the corresponding Portuguese forms are assigned EAGLES tags (in the rectangle)

- A worm ca n't smile /VB .	- Uma minhoca não ri /VMI . Testou o sentido de humor delas ?
Keep on smiling /VBG	Quando você rir /VMN
Why do you smile /VBP ?	Por que está rindo /VMG ?
Why are you smiling /VBG ?	Por que está rindo /VMG ?

VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non-3rd person singular present
VBZ Verb, 3rd person singular present

VERBOS			
Pos.	Atributo	Valor	Código
1	Categoria	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P

Source: Own elaboration

Biblical texts are also part of the InterCorp parallel corpus (Čermák; Rosen, 2012).² In addition, the corpus also includes movie subtitles, legal texts, proceedings of the European Parliament, or fiction. In this corpus, Czech serves as the pivot language, meaning that for every text in a language other than Czech, a corresponding Czech version is available. However, texts can also correspond between other language pairs. The parallel corpora are aligned at the sentence level. The current version of InterCorp (InterCorp Release 16;³ Rosen *et al.*, 2023) includes parallel texts in Czech and 61 other languages. In almost half of the languages, texts are provided with morphological annotation, i.e. the individual word forms are assigned the respective dictionary forms (lemmas) and morphological tags. The morphological tags used in the individual language sections of InterCorp come from various sources, which means that they describe the part of speech and selected morphological categories of individual word forms (such as person, tense, mood) by using

2 <https://intercorp.korpus.cz/?lang=en>

3 https://wiki.korpus.cz/doku.php/en:cnk:intercorp:historie#release_16

labels from different tagsets. Figure 1 provides example sentences from the InterCorp Release 16, where the English section is labeled with modified PennTreebank part-of-speech tags (Marcus *et al.*, 1993), while Portuguese texts use EAGLES tags (EAGLES 1996).

This diversity has been eliminated in the InterCorp Release 16ud,⁴ where the part-of-speech category and morphological features are provided in the unified Universal Dependencies scheme; see Figure 2. On top of morphological annotation, syntactic annotation is provided for each sentence, following the Universal Dependencies format. Some more details on the morphological annotation in Universal Dependencies are given in the following subsection.

Figure 2. Parallel sentences from the English and Portuguese sections of the InterCorp corpus with the unified morphological annotation based on Universal Dependencies

When she smiles /VERB//Sing/3 , you feel like -- I don't know .	Quando ela se ri /VERB//Sing/3 , sentimos ... não sei .
He smiles /VERB//Sing/3 .	Ele ri /VERB//Sing/3 .
When he smiles /VERB//Sing/3 , I can see through his head .	Buracos ! Quando se ri /VERB//Sing/3 , até se lhe vê o cérebro .
- She smiles /VERB//Sing/3 .	- Ela sorri , ela ri /VERB//Sing/3 .

Source: Own elaboration

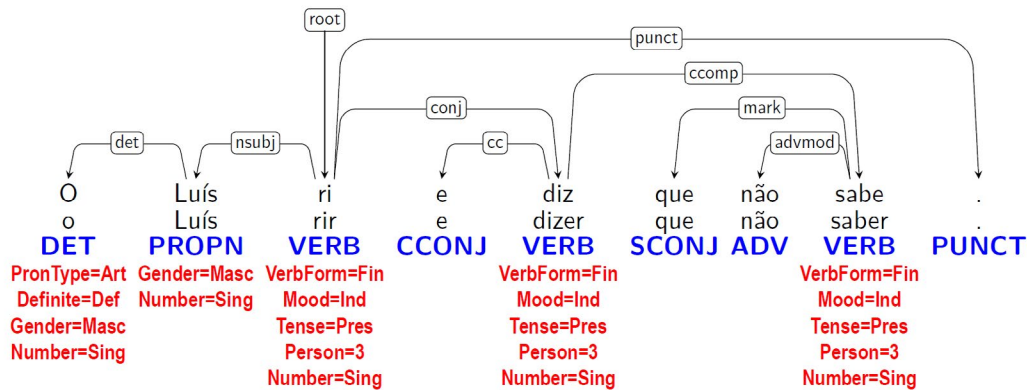
3.2 Universal Dependencies

In its current version 2.15, the Universal Dependencies⁵ collection contains 269 treebanks in 168 languages. The selection of texts and their size vary across the individual treebanks, yet this collection has excellent potential for cross-linguistic comparison due to the fact that the treebanks are annotated with cross-linguistically consistent morphological and syntactic information, which in Figure 3 is displayed below and above the sentence, respectively. The morphological annotation of a word consists of a lemma (displayed in black directly below each word form in the figure), a universal part-of-speech tag (in blue), and a set of universal features (in red). The universal part-of-speech tag is chosen from a fixed list of 17 tags. Universal features further specify the lexical and grammatical properties of the word. If any of the 28 features, each of which has an individual repertoire of values, is not relevant to a given word (e.g. tense with nouns in English), its value is not specified.

4 https://wiki.korpus.cz/doku.php/en:cnk:intercorp:historie#release_16ud

5 <https://universaldependencies.org/>

Figure 3. A Portuguese sentence provided by morphological annotation (below the sentence) and syntactic analysis (above the sentence) according to the Universal Dependencies scheme



Source: Own elaboration

Figure 4. Morphological analyses provided by some of the resource cited

resource	analysis provided				
UniMorph – English (inflections)	collaborationist	collaborationist	N;SG	-	
	collaborationist	collaborationists	N;PL	collaborationist s	
	overwrite	overwriting	V V.PTCP;PRS	overwrite ing	
	overwrite	overwritten	V V.PTCP;PST	overwrite en	
– German	beschreiben	beschreibe	V SBJV;PRS;3;SG	beschreib -e	
	riren	riremos	V IND;PST;PRF;1;PL	riren riremos	
	riso	riso	N;MASC;SG	-	
UniMorph – English (derivations)	collaboration	collaborationist	N:N	-ist	
	central	decentral	ADJ:ADJ	de-	
	centrally	decentrally	ADV:ADV	de-	
	central	subcentral	ADJ:ADJ	sub-	
	subcentral	subcentrally	ADJ:ADV	-ly	
	dezentral	dezentralisieren	ADJ:V	-isieren	
	dezentralisieren	dezentralisieren	V:V	de-	
	riren	riável	V:ADJ	-ável	
– Portug.	riso	risada	N:N	-ada	
CELEX – English	collaborationism	collaboration+ism	(((col)[V].Nx),((labour)[V])[N],(ate)[V xN.])[V],(ion)[N V.])[N],(ism)[N N.])		
	womenfolk	women+folk	((women)[N],(folk)[N])[N]		
	Umgangssprache	Umgang+s+Sprache	(((um)[V].V],(geh)[V])[V])[N],(s)[N N.N],((sprech)[V])[N])[N]		
– German	Grossmachtpolitik	Grossmacht+Politik	(((gross)[A],(Macht)[N])[N],((polit)[R],(ik)[N R.])[N])[N]		
MorphoLex – English	collaborationists	{<co<(labor)>ate>}>ion>>ist>			
UniSegments – English	rewriting	VERB	re+ writ+ing (morph: re, span: [0, 1], type: prefix; morph: writ, span: [2, 3, 4, 5], type: root;		
		morph: ing, span: [6, 7, 8], type: suffix)			
	risada	NOUN	ris+ada (span: [0, 1, 2], type: root; span: [3, 4, 5], type: suffix)		
	riável	ADJ	ri+ável (span: [0, 1], type: root; span: [2, 3, 4, 5], type: suffix)		
Corpus Kadiwéu	idinaGataGatinigi				
	i	di	n	aGata	Ga
	1Absolutive	Inverse.voice	Antipassive	hide	Plural
					tinigi
					Applicative

Source: Own elaboration

4. Morphology in lexical datasets

I use “lexical datasets” as an umbrella term for data sources that, in contrast to text corpora just discussed, are repositories of words without contexts. Examples of lexical datasets are given where words are provided with inflectional features (Section 4.1), with information about their derivational history (Section 4.2), and with information about their internal structure (Section 4.3).

4.1 Inflection

The UniMorph⁶ database (Batsuren *et al.*, 2022) currently provides word lists for 169 languages and analyses them in terms of inflectional and/or derivational morphology. The inflectional data consist of quadruples:

- a lemma (cf. the first column in the example entry from the UniMorph inflectional files in Figure 4; e.g. the Portuguese *rir*),
- one of the inflected forms of the lemma (cf. *ríramos* in the second column in the example),
- a part-of-speech category along with morphological features characterizing the particular inflected form (cf. third column). The features are specified by the Universal Morphological Feature Schema (Sylak-Glassman 2016), which is heavily based on the Leipzig Glossing Rules (Comrie *et al.*, 2008). The features are separated from the part-of-speech value if an inflectional affix is identified in the word form, as with *ríramos*, but not with *riso*.
- morphological segmentation of the particular word form (fourth column); in Portuguese and English, a so-called canonical segmentation is provided, where the affix (in this case, inflectional) is separated and the remaining string (root or stem) is replaced by the citation form of the word form analyzed (e.g. *rir|ríramos*; cf. Kann *et al.*, 2016). In German, in contrast, the verb form is cut into two substrings without replacing the actual string (*beschreib*) by an infinitive (*beschreiben* ‘to describe’). This column is not used with words without overt inflectional markers.

4.2 Derivation

Derivational analysis is provided for 30 languages in UniMorph 4.0. It differs from the inflectional data produced by this project in several respects. While the inflectional files contain both lemmas and inflected forms, the derivational section is limited to lemmas. As illustrated in Figure 4, the derivational analysis consists of four pieces of information:

⁶ <https://unimorph.github.io/>

- the lemma of the word is listed in the first column that is assumed to enter into derivation to form the word in the second column (cf. the verb *rir* as the motivating word for the derived adjective *riável* in the Portuguese example in Figure 4),
- the part-of-speech category of the input and output words is provided in the third column (cf. V:ADJ related to the above example pair),
- the derivational affix that is assumed to attach to the predecessor to form the derivative is listed in the fourth column.

Some of the analyses provided may be conceived of as inconsistencies; cf. the English examples in Figure 4, where the adverb *decentrally* is formed by a prefix from *centrally*, while *subcentrally* by a suffix from *subcentral*. In the derivational data from German, two immediate predecessors are listed for the German verb *dezentralisieren* 'to decentralize'.

The Universal Derivations⁷ project is another attempt to describe derivational morphology across languages in a unified manner. This resource, which currently covers 21 languages, was created by harmonizing available resources into a unified scheme. According to this scheme, each derivative is linked to its motivating word. By connecting these pairs, a rooted tree structure is formed, with the root node being an unmotivated word. Other words sharing the same root are then organized around it according to increasing morphological complexity; see Figure 5. Nevertheless, the size of the trees and the data for individual languages vary significantly due to the different design decisions made by the original resources, such as limitations to particular part-of-speech categories or to particular types of derivatives (cf. Kyjánek *et al.*, 2020, also for the references to the original datasets).

It should be noted that while the inflectional analysis – often more generally referred to as morphological analysis – in corpora and lexical datasets (cf. Section 3 and 4.1, respectively) commonly involves specification of inflectional meanings, such as number and tense, semantic aspects are not addressed in the derivational data of UniMorph or in Universal Derivations. This is related to the fact that the cross-linguistic discussion of meanings in derivation has only recently been initiated by Bagasheva's (2017) pilot proposal, which was utilized by Körtvélyessy *et al.*, (2020) in their pioneering project on derivational networks.

4.3 Morphological segmentation

Morphological segmentation is understood here as the task of breaking down words into sequences of minimal meaning-bearing units (morphemes), such as roots, prefixes, and suffixes. The delimitation of final inflections and derivational affixes, which has been

⁷ <https://ufal.mff.cuni.cz/universal-derivations>

described as part of the analyses provided by the UniMorph database, can be viewed as a partial segmentation while other resources aim at identifying all morphemes within a word's structure (complete segmentation). Figure 4, which contains examples from the resources that are briefly compared in this section, illustrate some of the differences between the existing approaches.

As described above, UniMorph analyses inflection and derivation of each language separately and according to different principles. In the inflection analysis, an inflectional marker is identified; if not available, no segmentation is carried out. The segmentation complies with the principles of canonical segmentation, where the non-affixal part (root or stem) is replaced by the complete citation form of the word. Derivational analysis attempts to identify solely the morpheme by which a given word may have been formed from a word with a simpler morphological structure. As exemplified above, though, the segmentations are not always consistent within specific languages and across them.

A similar approach to the segmentation of derivatives (and other complex words) was applied by the 1990s project CELEX, which is a forerunner in the field of morphological data (Baayen *et al.*, 1995) and besides English and German (with examples in Figure 4) it covers also Dutch. The resource identifies, in derivatives, the morpheme that is assumed to have been added last (i.e. *-ism* in *collaborationism*) while compounds are decomposed into individual components (*Umgang+s+Sprache*). On top of that, complete segmentation is also provided, i.e. all morphemes in the structure of the word are identified. This second analysis complies with the principles of canonical segmentation (cf. *ate* instead of *at* in *collaborationism*; *geh* instead of *gang* and *sprech* instead of *sprach* in the German compound *Umgangssprache* 'colloquial language'), though some cases deviate from this pattern (cf. *womenfolk*, in which *women* is not replaced by *woman*).

Complete segmentation adhering to the canonical principles is also provided in MorphoLex,⁸ which is available for English and French (Sánchez-Gutiérrez *et al.*, 2018, Mailhot *et al.*, 2020). The MorphoLex example in Figure 4 documents that the final inflection is omitted in the analysis.

In the Universal Segmentations⁹ project (Žabokrtský *et al.*, 2022), a unified annotation scheme for morphological segmentation has been proposed. Words are completely segmented into substrings (morphs), that are not replaced by citation forms. Each morph is classified as prefix, root, or suffix (cf. Figure 4). Segmented data for 32 languages were published in the first version of this resource.

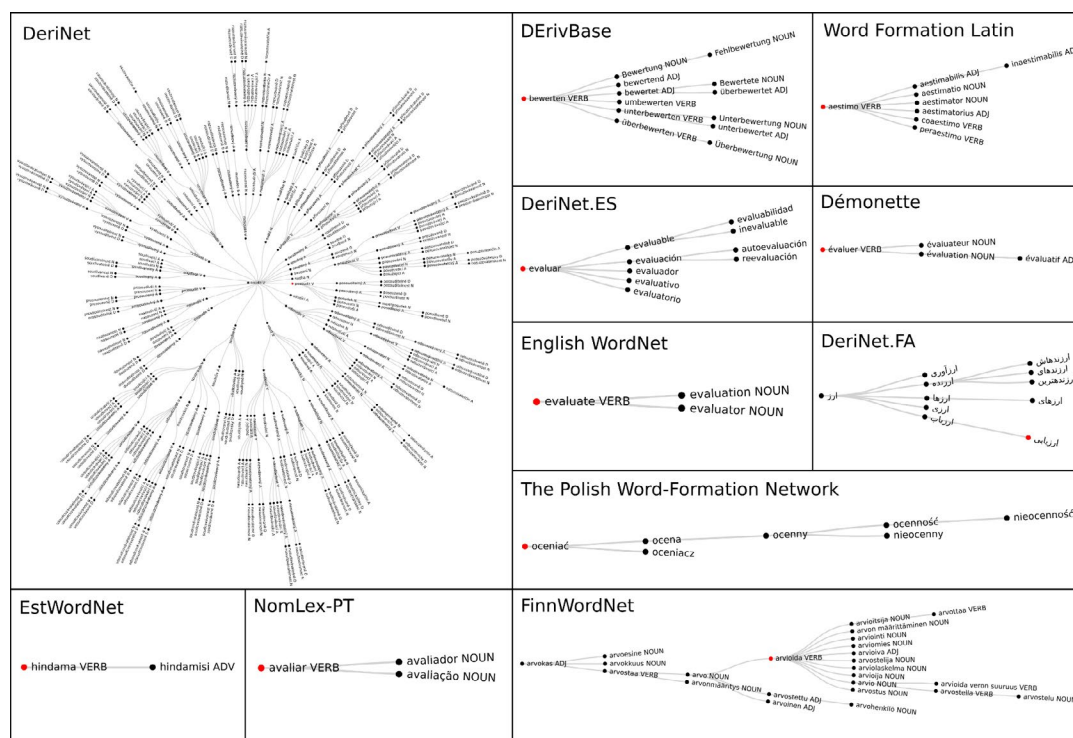
8 <https://github.com/hugomailhot/MorphoLex-en>, <https://github.com/hugomailhot/morpholex-fr>

9 <https://ufal.mff.cuni.cz/universal-segmentations>

While none of the previously mentioned sources providing complete segmentation have identified the meanings of morphemes, there are projects that have undertaken this challenging task for individual languages or, so far, rather modest multilingual collections. It is worth noting that these are projects where annotation at the level of morphemes is part of a multi-level analysis going up to the level of syntactic relations: Corpus Kadiwéu¹⁰ (Galves *et al.*, 2017, Sandalo; Galves 2023), the Universal Dependencies treebank for Beja (Kahane *et al.*, 2021), and Guillaume *et al.*, (2024) proposing annotation guidelines for morpheme-level analysis in Universal Dependencies treebanks and implementing this annotation in three languages. The morphological segmentation of the word form *idinaGataGatinigi* (meaning 'we hid in') from Corpus Kadiwéu is provided in Figure 4 for the sake of comparison.

Moreover, these pioneering projects also pave the way for connecting information about the internal structure of words with analysis at the level of words and sentences, thus escaping the separation that we currently see in linguistic data sources.

Figure 5. Graphs from the Universal Derivations database modelling derivations in Czech (DeriNet), German (DErивBase), Latin (Word Formation Latin), Spanish (DeriNet.ES), French (Démonette), English (English WordNet), Farsí (DeriNet.FA), Polish (Polish Word-Formation Network), Estonia, (EstWordNet), Portuguese (NomLex-PT), and Finnish (FinnWordNet)



Source: Own elaboration

¹⁰ <https://www.tycho.iel.unicamp.br/viewer/C12>

Figure 6. Examples of features and language-specific values from the typological databases
WALS and Grambank

WALS		
Feature 33A: Coding of nominal plurality		
English: plural suffix	Portuguese: plural suffix	Kadiwéu: plural suffix
Feature 112A: Negative morphemes		
English: negative particle	Portuguese: negative particle	Kadiwéu: negative affix
Grambank		
Feature GB030: Is there a gender distinction in independent 3rd person pronouns?		
English: present	Portuguese: present	Hungarian: absent
Feature GB039: Is there nonphonological allomorphy of noun number markers?		
English: present	Portuguese: absent	Hungarian: absent

Source: Own elaboration

5. Typological databases

Typological databases differ from the resources described so far in that they do not contain authentic language material (words and sentences), but they gather structural characteristics extracted from available reference books and other sources for up to thousands of languages. The information in these databases takes the form of metadata on the phonological, morphological, syntactic, and lexical features of individual languages.

Of 192 structural features listed in the World Atlas of Language Structures (WALS; Dryer; Haspelmath, 2013),¹¹ twelve are classified as morphological, another 29 as describing nominal categories, and 17 reporting on verbal categories. All these features are related to inflectional morphology, while derivation is not considered.

The more recent Grambank¹² database (Skirgård *et al.*, 2023) draws on WALS, but differs from it in the inclusion of both inflection and derivation, and also differs in other aspects, including the formulation of features, as illustrated in Figure 6.

As the typological databases do not list all features for every language, the prediction of missing entries has become one of the tasks in the emerging field of computational typology. It has been addressed by using various natural language processing techniques

¹¹ <https://wals.info/>

¹² <https://grambank.clld.org/>

(for summary, cf. Ponti *et al.*, 2019) and became the topic of the 2020 shared task organized by the Association for Computational Linguistics Special Interest Group on Typology (SIGTYP 2020 Shared Task; Bjerva *et al.*, 2020).

Apart from the broad-scope typological resources, dedicated databases have been compiled, such as the Atlas of Pidgin and Creole Language Structures Online (APiCS, Michaelis *et al.*, 2013).¹³

Platforms for sharing typological data have been developed, such as the AUTOTYP (Bickel; Nichols 2002)¹⁴ or URIEL (Litell *et al.*, 2017)¹⁵ meta-databases and, more recently, the CLDF initiative, which additionally calls for data integration through language codes and concepts (Forkel *et al.*, 2018).¹⁶

6. Concluding remarks

The present paper has provided a selective survey of what information about the morphology of natural languages is captured in existing linguistic resources. Since viewed from the perspective of the potential use of these data for language comparison and linguistic typology, I concentrated – apart from a few noteworthy exceptions – on sources covering multiple languages. Existing multilingual resources, into which significant effort has been invested in the recent decades, essentially provide data on all aspects of word structure that are included when defining the scope of morphology as a linguistic sub-discipline. Text corpora document the use of individual inflectional forms in sentence contexts. The available lexical datasets contain information on inflectional morphology, derivational morphology, and the internal structure of words.

A closer look at the selected resources revealed that, regardless of the number of languages covered, maintaining consistent annotation is a significant challenge. When comparing the resources, it became clear that although they aim to address the same phenomenon, the analyses are often based on different principles.

Efforts aimed at improving the internal consistency of the different types of morphological resources and expanding their coverage, as well as ensuring their mutual compatibility and combinability, and, last but not least, linking these resources with those containing other types of linguistic annotations (such as lexical and semantic), could considerably enhance their exploitability for comparative and typological studies.

¹³ <https://apics-online.info/>

¹⁴ <https://www.autotyp.uzh.ch/>

¹⁵ http://www.cs.cmu.edu/~dmortens/projects/7_project/

¹⁶ <https://cldf.clld.org/>

Acknowledgements

I would like to express my sincere gratitude to the organizers of the 70th GEL Seminar, in particular to Filomena Sandalo and Livia Oushiro, for inviting me to participate in the roundtable on Linguistically Annotated Computational Resources for NLP & Language Typology and for providing the opportunity to present my contribution in its published form. The research reported on in the present paper was supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

- ARONOFF, M.; FUDEMAN, K. *What is Morphology?* Oxford: Blackwell, 2005.
- BAAZEN, R. H.; PIEPENBROCK, R.; GULIKERS, L. *CELEX2*. Philadelphia: Linguistic Data Consortium, 1995.
- BAGASHEVA, A. Comparative Semantic Concepts in Affixation. In: SANTANA-LARIO, J.; VALERA, S. (ed.). *Competing patterns in English affixation*. Bern: Peter Lang, 2017. p. 33-65.
- BATSUREN, K.; GOLDMAN, O.; KHALIFA, S.; HABASH, N.; KIERAŚ, W.; BELLA, G.; LEONARD, B.; NICOLAI, G.; GORMAN, K.; GHANGGO ATE, Y.; RYSKINA, M.; MIELKE, S.; BUDIANSKAYA, E.; EL-KHAISSI, C.; PIMENTEL, T.; GASSER, M.; LANE, W. A.; RAJ, M.; COLER, M.; MONTOYA SAMAME, J. R.; SITICONATZI CAMAITERI, D.; ZUMAETA ROJAS, E.; LÓPEZ FRANCIS, D.; ONCEVAY, A.; LÓPEZ BAUTISTA, J.; SILVA VILLEGAS, G. C.; TORROBA HENNIGEN, L.; EK, A.; GURIEL, D.; DIRIX, P.; BERNARDY, J.-P.; SCHERBAKOV, A.; BAYYR-OOL, A.; ANASTASOPOULOS, A.; ZARIQUIEY, R.; SHEIFER, K.; GANIEVA, S.; CRUZ, H.; KARAHÓĞA, R.; MARKANTONATOU, S.; PAVLIDIS, G.; PLUGARYOV, M.; KLYACHKO, E.; SALEHI, A.; ANGULO, C.; BAXI, J.; KRIZHANOVSKY, A.; KRIZHANOVSKAYA, N.; SALESKY, E.; VANIA, C.; IVANOVA, S.; WHITE, J.; HALL MAUDSLAY, R.; VALVODA, J.; ZMIGROD, R.; CZARNOWSKA, P.; NIKKARINEN, I.; SALCHAK, A.; BHATT, B.; STRAUGHN, C.; LIU, Z.; NORTH WASHINGTON, J.; PINTER, Y.; ATAMAN, D.; WOLINSKI, M.; SUHARDIJANTO, T.; YABLONSKAYA, A.; STOEHR, N.; DOLATIAN, H.; NURIAH, Z.; RATAN, S.; TYERS, F. M.; PONTI, E. M.; AITON, G.; ARORA, A.; HATCHER, R. J.; KUMAR, R.; YOUNG, J.; RODIONOVA, D.; YEMELINA, A.; ANDRUSHKO, T.; MARCHENKO, I.; MASHKOVTSOVA, P.; SEROVA, A.; PRUD'HOMMEAUX, E.; NEPOMNIASHCHAYA, M.; GIUNCHIGLIA, F.; CHODROFF, E.; HULDEN, M.; SILFVERBERG, M.; MCCARTHY, A. D.; YAROWSKY, D.; COTTERELL, R.; TSARFATY, R.; VYLOMOVA, E. UniMorph 4.0: Universal Morphology. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: ELRA, 2022. p. 840-855.

BICKEL, B.; NICHOLS, J. Autotypologizing databases and their use in fieldwork. In: *Proceedings of the LREC 2002 Workshop on Resources and Tools in Field Linguistics*. Las Palmas: ELRA, 2002.

BJERVA, J.; SALESKY, E.; MIELKE, S. J.; CHAUDHARY, A.; CELANO, G. G. A.; PONTI, E. M.; VYLOMOVA, E.; COTTERELL, R.; AUGENSTEIN, I. SIGTYP 2020 shared task: Prediction of typological features. In: *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. 2020. p. 1-11.

ČERMÁK, F.; ROSEN, A. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, v. 13, n. 3, 2012. p. 411-427.

COMRIE, B.; HASPELMATH, M.; BICKEL, B. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Disponível em: <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>. Acesso em: 07 jul. 2025.

DRYER, M. S.; HASPELMATH, M. (ed.). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013.

EAGLES. *Recommendations for the Morphosyntactic Annotation of Corpora*. Disponível em: <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html>. Acesso em: 07 jul. 2025.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (ed.). *Ethnologue: Languages of the World*. 28. ed. Dallas, Texas: SIL International, 2025.

FORKEL, R.; LIST, J.-M.; GREENHILL, S. J.; RZYMSKI, C.; BANK, S.; CYSOUW, M.; HAMMARSTRÖM, H.; HASPELMATH, M.; KAIPING, G. A.; GRAY, R. D. Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, v. 5, 2018. p. 180205.

GALVES, C.; SANDALO, F.; VERONESI, L. Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, v. 59, n. 2, 2017. p. 631-648.

GREENBERG, J. H. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In: GREENBERG, J. H. (ed.). *Universals of Language*. Cambridge, Mass: MIT Press, 1963. p. 73-113.

GUILLAUME, B.; GERDES, K.; GUILLER, K.; KAHANE, S.; LI, Y. Joint Annotation of Morphology and Syntax in Dependency Treebanks. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino: ELRA; ICCL, 2024. p. 9568-9577.

KAHANE, S.; VANHOVE, M.; ZIANE, R.; GUILLAUME, B. 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)*. Sofia: ACL, 2021. p. 48-60.

KANN, K.; COTTERELL, R.; SCHÜTZE, H. Neural Morphological Analysis: Encoding-Decoding Canonical Segments. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: ACL, 2016. p. 961-967.

KÖRTVÉLYESSY, L.; BAGASHEVA, A.; ŠTEKAUER, P. *Derivational Networks across Languages*. Berlin: De Gruyter, 2020.

KYJÁNEK, L.; ŽABOKRTSKÝ, Z.; ŠEVČÍKOVÁ, M.; VIDRA, J. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, v. 115, n. 2, 2020. p. 5-30.

LITTELL, P.; MORTENSEN, D.; LIN, K.; KAIRIS, K.; TURNER, C.; LEVIN, L. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In: *Proceedings of the 15th Conference of EACL*. Valencia: ACL, 2017. p. 8-14.

MAILHOT, H.; WILSON, M. A.; MACOIR, J.; DEACON, S. H.; SÁNCHEZ-GUTIÉRREZ, C. MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, v. 52, n. 4, 2020. p. 1008-1025.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, v. 19, n. 2, p. 313-330, 1993.

DE MARNEFFE, M.-C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, 2021. p. 255-308.

MCCARTHY, A. D.; WICKS, R.; LEWIS, D.; MUELLER, A.; WU, W.; ADAMS, O.; NICOLAI, G.; POST, M.; YAROWSKY, D. The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*. Marseille: ELRA, 2020. p. 2884-2892.

MICHAELIS, S. M.; MAURER, P.; HASPELMATH, M.; HUBER, M. (ed.). *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013.

PONTI, E. M.; O'HORAN, H.; BERZAK, Y.; VULIĆ, I.; REICHART, R.; POIBEAU, T.; SHUTOVA, E.; KORHONEN, A. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, v. 45, n. 3, p. 559-601, 2019.

ROSEN, A.; VAVŘÍN, M.; ZASINA, A. J. InterCorp corpus version 16. Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University, 2023. Disponível em: <http://www.korpus.cz>. Acesso em: 07 jul. 2025.

SÁNCHEZ-GUTIÉRREZ, C. H.; MAILHOT, H.; DEACON, S. H.; WILSON, M. A. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, v. 50, n. 4, p. 1568-1580, 2018.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: o problema de Anchieta. *Cadernos de Estudos Linguísticos*, v. 65, n. 1, p. 1-26, 2023.

SKIRGÅRD, H.; HAYNIE, H. J.; BLASI, D. E.; HAMMARSTRÖM, H.; COLLINS, J.; *et al.* Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, v. 9, 2023. DOI: 10.1126/sciadv.adg6.

SYLAK-GLASSMAN, J. The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema). Baltimore: John Hopkins University, 2016.

ŽABOKRTSKÝ, Z.; BAFNA, N.; BODNÁR, J.; KYJÁNEK, L.; SVOBODA, E.; ŠEVČÍKOVÁ, M.; VIDRA, J. Towards Universal Segmentations: UniSegments 1.0. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: ELRA, 2022. p. 1137-1149.