MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology

Khuyagbaatar Batsuren[1], **Gábor Bella**[2], **and Fausto Giunchiglia**[2,3]
[1]National University of Mongolia, Mongolia
[2]DISI, University of Trento, Italy
[3]College of Computer Science and Technology, Jilin University, China
khuyagbaatar@num.edu.mn;{gabor.bella,fausto.giunchiglia}@unitn.it

Abstract

Large-scale morphological databases provide essential input to a wide range of NLP applications. Inflectional data is of particular importance for morphologically rich (agglutinative and highly inflecting) languages, and derivations can be used, e.g. to infer the semantics of out-of-vocabulary words. Extending the scope of state-of-the-art multilingual morphological databases, we announce the release of MorphyNet, a high-quality resource with 15 languages, 519k derivational and 10.1M inflectional entries, and a rich set of morphological features. MorphyNet was extracted from Wiktionary using both hand-crafted and automated methods, and was manually evaluated to be of a precision higher than 98%. Both the resource generation logic and the resulting database are made freely available[1][2] and are reusable as stand-alone tools or in combination with existing resources.

1 Introduction

Despite repeated paradigm shifts in computational linguistics and natural language processing, morphological analysis and its related tasks, such as lemmatization, stemming, or compound splitting, have always remained essential components within language processing systems. Recently, in the context of language models based on subword embeddings, a morphologically meaningful splitting of words has been shown to improve the efficiency of downstream tasks (Devlin et al., 2019; Sennrich et al., 2016; Bojanowski et al., 2017; Provilkov et al., 2020). In particular, the reintroduction of linguistically motivated approaches and high-quality linguistic resources into deep learning architectures has been crucial for dealing with morphologically rich—highly inflecting, agglutinative—languages more efficiently (Pinnis et al., 2017; Vylomova et al., 2017; Ataman and Federico, 2018; Gerz et al., 2018).

In response to such needs, and as simple and convenient substitutes for monolingual morphological analyzers, multilingual *morphological databases* have been developed, indicating for each word form entry one or more corresponding root or dictionary entries, as well as analysis (features) (Kirov et al., 2018; Metheniti and Neumann, 2020; Vidra et al., 2019). The precision and recall of these resources vary wildly, and there is still a lot of ground to cover with respect to the support of new languages, the modelling of the inflectional and derivational complexity of each language, as well as the richness of the information (features, affixes, parts of speech, etc.) provided.

As a further step towards extending online morphological data, we introduce *MorphyNet*, a new database that addresses both derivational and inflectional morphology. Its current version covers 15 languages and has 519k derivational and 10.1M inflectional entries, as well as a rich set of features (lemma, parts of speech, morphological tags, affixes, etc.). Similarly to certain existing databases, MorphyNet was built from *Wiktionary* data; however, our extraction logic allows for a more exhaustive coverage of both derivational and inflectional cases.

The contributions of this paper are the freely available MorphyNet resource, the description of the data extraction logic and tool, also made freely accessible, as well as its evaluation and comparison to state-of-the-art multilingual morphological databases. Due to the limited overlap between the contents of these resources and MorphyNet, we consider it as complementary and therefore usable in combination with them.

Section 2 of the paper presents the state of the art. Section 3 gives details on our method for generat-

[1] http://ukc.disi.unitn.it/index.php/MorphyNet
[2] http://github.com/kbatsuren/MorphyNet

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology,pages 39–48

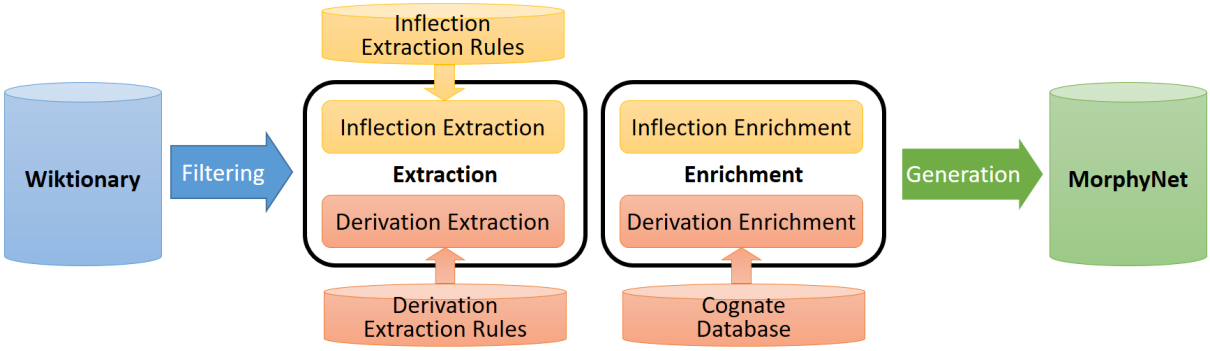

Figure 1: The MorphyNet generation process and the input datasets used.

ing MorphyNet data. Section 4 presents the resulting resource, and Section 5 evaluates it. Section 6 concludes the paper.

2 State of the Art

Ever since the early days of computational linguistics, morphological analysis and its related tasks—such as stemming and lemmatization—have been part of NLP systems. Earlier grammar-based systems used finite-state transducers or affix stripping techniques, and certain of them were already multilingual and were capable of tackling morphologically complex languages (Beesley and Karttunen, 2003; Trón et al., 2005; Inxight, 2005). However, due to the costliness of producing the grammar rules that drove them, many of these systems were only commercially available.

More recently, several projects have followed the approach of formalizing and/or integrating existing morphological data for multiple languages. *UDer (Universal Derivations)* (Kyjánek et al., 2020) integrates 27 derivational morphology resources in 20 languages. *UniMorph* (Kirov et al., 2016, 2018) and the *Wikinflection Corpus* (Metheniti and Neumann, 2020) rely mostly on *Wiktionary* from which they extract inflectional information. Beyond the data source, however, the two last projects have little in common: UniMorph is by far more precise and complete, and being used as gold standard for NLP community (Cotterell et al., 2017, 2018) (recently covering 133 languages (McCarthy et al., 2020)), while Wikinflection follows a naïve, linguistically uninformed approach of merely concatenating affixes, generating an abundance of ungrammatical word forms (e.g. for Hungarian or Finnish).

MorphyNet is also based on extracting morphological information from Wiktionary, extending the scope of UniMorph by new extraction rules and logic. The first version of MorphyNet covers 15 languages, and it is distinct from other resources in three aspects: (1) it includes both inflectional and derivational data; (2) it extracts a significantly higher number of inflections from Wiktionary; and (3) it provides a wider range of morphological information. While for the languages it covers MorphyNet can be considered a superset of UniMorph, the latter supports more languages. With UDer, as we show in section 4, the overlap is minor on all languages. For these reasons, we consider MorphyNet as complementary to these databases, considerably enriching their coverage on the 15 supported languages but not replacing them.

3 MorphyNet Generation

MorphyNet is generated mainly from Wiktionary, through the following steps.

1. *Filtering* returns XML-based Wiktionary content from specific sections of relevant lexical entries: headword lines, etymology sections, and inflectional tables are returned for nouns, verbs, and adjectives.

2. *Extraction* obtains raw morphological data by parsing the sections above.

3. *Enrichment* algorithmically extends the coverage of derivations and inflections obtained from Wiktionary, through entirely distinct methods for inflection and derivation.

4. *Resource generation*, finally, outputs MorphyNet data.

Below we explain the non-trivial Wiktionary extraction and enrichment steps, while Section 4 provides details on the generated resource itself.

3.1 Wiktionary Extraction

We extract inflectional and derivational data through hand-crafted extraction rules that target recurrent patterns in Wiktionary content both in source markdown and in HTML-rendered form. With respect to UniMorph that takes a similar approach and scrapes tables that provide inflectional paradigms, the scope of extraction is considerably extended, also including headword lines and etymology sections. This allows us to obtain new derivations, inflections, and features not covered by UniMorph, such as gender information or noun and adjective declensions for Catalan, French, Italian, Spanish, Russian, English, or Serbo-Croatian. Our rules target nouns, adjectives, and verbs in all languages covered.

Inflection extraction rules target two types of Wiktionary content: *inflectional tables* and *headword lines*. Inflectional tables provide conjugation and declension paradigms for a subset of verbs, nouns, and adjectives in Wiktionary. On tables, our extraction method was similar to that of UniMorph as described in (Kirov et al., 2016, 2018), with one major difference. UniMorph also extracted a large number of separate entries with modifier and auxiliary words, such as Spanish negative imperatives (*no comas*, *no coma*, *no comamos* etc.) or Finnish negative indicatives (*en puhu*, *et puhu*, *eivät puhu* etc.). MorphyNet, on the other hand, has a single entry for each distinct word form, regardless of the modifier word used. This policy had a particular impact on the size of the Finnish vocabulary.

As inflectional tables are only provided by Wiktionary for 62.5%[3] of nouns, verbs, and adjectives, we extended the scope of extraction to headword lines, such as

banca *f* (*plural* **banche**)

From this headword line, we extract two entries: one for *banca* is feminine singular and second for *banche* is feminine plural. We created specific parsing rules for nouns, verbs, and adjectives because each part of speech is described through a different set of morphological features. For example, valency (*transitive* or *reflexive*) and aspect (*perfective* or *imperfective*) are essential for verbs, while gender (*masculine* or *feminine*) and number (*singular* or *plural*) pertain to nouns and adjectives.

Derivation extraction rules were applied to the

[3]Computed over the 15 languages covered by MorphyNet.

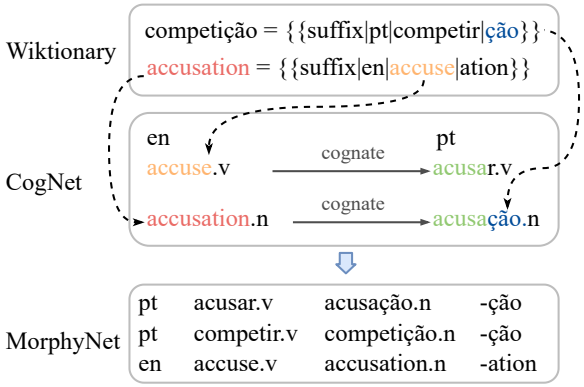

Figure 2: Derivation enrichment example: inference of the derivation of the Portuguese word *acusação*.

etymology sections of Wiktionary entries to collect the Morphology template usages, such as for the English *accusation*:

Equivalent to **accuse** + **-ation**.

where we have a morphology entry {{suffix|en|accuse|-ation}} from the Wiktionary XML dump. After collecting all morphology entries, we applied the enrichment method to increase its coverage.

3.2 Derivation Enrichment

Derivation enrichment is based on a linguistically informed cross-lingual generalization of derivational patterns observed in Wiktionary data, in order to extend the coverage of derivational data.

In the example shown in Figure 2, Wiktionary contains the Portuguese derivation *competir* (*to compete*) → *competição* (*competition*) but not *acusar* (*to accuse*) → *acusação* (*accusation*). An indiscriminate application of the suffix *-ção* to all verbs would, of course, generate lots of false positives, such as *chegar* (*to arrive*) ↛ **chegação*. Even when the target word does exist, the inferred derivation is often false, as in the case of *corar* (*to blush*) ↛ *coração* (*heart*). A counter-example from English could be *jewel* + *-ery* → *jewellery* but *gal* + *-ery* ↛ *gallery*.

For this reason, we use stronger cross-lingual derivational evidence to induce the applicability of the affix. In the example above, the existence of the English derivation *accuse* → *accusation*, where the meanings of the English and the corresponding Portuguese words are the same, serves as a strong hint for the applicability of the Portuguese pattern.

This intuition is formalized in MorphyNet as fol-

Table 1: Structure of MorphyNet inflectional data and its comparison to UniMorph. Data provided only by MorphyNet is highlighted in bold. The rest is provided by both resources in a nearly identical format.

Language	base_word	trg_word	features	**src_word**	**morpheme**
Hungarian	ház	házak	N;NOM;PL	**ház**	**-ak**
Hungarian	ház	házat	N;ACC;SG	**ház**	**-at**
Hungarian	ház	házakat	N;ACC;PL	**házak**	**-at**
Russian	играть	играть	V;NFIN**;IPFV;ACT**	**играть**	
Russian	играть	играют	V**;IND**;PRS;3;PL**;IPFV;ACT;FIN**	**играть**	**-ают**
Russian	играть	играющий	V;V.PTCP;ACT;PRS	**играют**	**-щий**

Table 2: Structure of MorphyNet derivational data and its comparison to UDer. Data only provided by MorphyNet is highlighted in bold. The rest is provided by both resources in a nearly identical format.

Language	src_word	trg_word	src_pos	trg_pos	**morpheme**
English	time	timeless	noun	adjective	**-less**
English	soda	sodium	noun	**substance.**noun	**-ium**
English	zoo	zoophobia	noun	**state.**noun	**-phobia**
Finnish	kirjoittaa	kirjoittaminen	verb	noun	**-minen**
Finnish	lyödä	lyöjä	verb	**person.**noun	**-jä**

lows: if in language A a derivation from source word w_s^A to target word w_t^A through the affix a^A is not explicitly asserted (e.g. by Wiktionary) but it is asserted for the corresponding *cognates* in at least one language B, then we infer its existence:

$$\mathrm{cog}(w_s^A, w_s^B) \wedge \mathrm{cog}(w_t^A, w_t^B) \wedge \mathrm{cog}(a^A, a^B)$$

$$\wedge\, \mathrm{der}(w_s^B, a^B) = w_t^B \Rightarrow \mathrm{der}(w_s^A, a^A) = w_t^A$$

where $\mathrm{cog}(x, y)$ means that the words x and y are cognates and $\mathrm{der}(b, a) = d$ that word d is derived from base word b and affix a. In our example, $A = \textit{Portuguese}$, $B = \textit{English}$, $w_s^A = \textit{acusar}$, $w_s^B = \textit{accuse}$, $w_t^A = \textit{acusação}$, $w_t^B = \textit{accusation}$, $a^A = \textit{-ção}$, and $a^B = \textit{-tion}$.

As shown in Figure 1, we exploited a cognate database, *CogNet*[4] (Batsuren et al., 2019, 2021), that has 8.1M cognate pairs, for evidence on cognacy: $\mathrm{cog}(w^A, w^B) = \textit{True}$ is asserted by the presence of the word pair in CogNet.

The result of enrichment was a total increase of 25.6% of the number of derivations in MorphyNet. Efficiency varies among languages, essentially depending on the completeness of the Wiktionary coverage: it was the lowest for English with 3% and the highest for Spanish with 57%.

3.3 Inflection Enrichment

The enrichment of inflectional data is based on the simple observation that Wiktionary does not provide the root word for all inflected forms. For example, for the Hungarian *múltjával* (*with his/her/its past*), Wiktionary provides the inflection *múltja* → *múltjával* (*his/her/its past* + instrumental). For *múltja*, in turn, it provides *múlt* → *múltja* (*past* + possessive). It does not, however, directly provide the combination of the two inflections: *múlt* → *múltjával* (*past* + possessive + instrumental). Inflection enrichment consists of inferring such missing rules from the existing data.

The case above is formalized as follows: if, after the Wiktionary extraction phase, the MorphyNet data contains the inflections $w_r \rightarrow w_1$ (with feature set F_1) as well as $w_1 \rightarrow w_2$ (with feature set F_2), then we create the new inflection $w_r \rightarrow w_2$ with feature set $F_1 \cup F_2$.

The application of this logic increased the inflectional coverage of MorphyNet by 10.8% and its recall (with respect to ground truth data presented in section 5) by 8.2% on average.

4 The MorphyNet Resource

Morphynet is freely available for download, both as text files containing the data and as the source code of the Wiktionary extractor.[5] Two text files are provided per language: one for inflections and one for derivations. The structure of the two types of files is illustrated in Tables 1 and 2, respectively. As shown, MorphyNet covers all data fields provided by UniMorph for inflections and by UDer for derivations. In addition, it extends UniMorph by indicating the affix and the immediate source word that produced the inflection. Such information is useful, for example, to NLP applications that rely on subword information for understand-

[4] http://github.com/kbatsuren/CogNet

[5] http://github.com/kbatsuren/WiktConv

Table 3: MorphyNet dataset statistics

		Inflectional morphology			Derivational morphology			
#	Languages	words	entries	morphemes	words	entries	morphemes	Total
1	Finnish	65,402	1,617,751	1,139	18,142	37,199	446	1,654,950
2	Serbo-Croatian	68,757	1,760,095	263	8,553	20,008	429	1,780,103
3	Italian	75,089	748,321	104	22,650	42,149	749	790,470
4	Hungarian	38,067	1,034,317	428	14,566	37,940	832	1,072,257
5	Russian	67,695	1,343,760	252	21,922	36,922	575	1,380,682
6	Spanish	67,796	677,423	145	16,268	27,633	490	705,056
7	French	44,729	453,229	98	15,473	37,203	636	490,432
8	Portuguese	30,969	329,861	161	10,504	15,974	387	345,835
9	Polish	36,940	663,545	251	9,518	18,404	405	681,949
10	German	35,086	214,401	243	13,070	23,867	465	238,268
11	Czech	9,781	298,888	112	4,875	9,660	318	307,935
12	English	149,265	652,487	8	67,412	200,365	2,445	852,852
13	Catalan	16,404	168,462	91	3,244	4,083	220	172,545
14	Swedish	14,485	131,693	32	3,190	5,810	217	137,503
15	Mongolian	2,085	14,592	35	1,410	1,940	229	16,532
	Total	722,550	10,108,825	3,362	230,797	519,157	8,843	10,627,369

Table 4: UniMorph and MorphyNet data sizes compared to Universal Dependencies content.

Language	UniMorph	MorphyNet	Univ. Dep.
Catalan	81,576	168,462	25,443
Czech	134,528	298,888	151,838
English	115,523	652,487	17,296
French	367,733	453,229	28,921
Finnish	2,490,377	1,617,751	47,813
Hungarian	552,950	1,034,317	3,685
Italian	509,575	748,321	24,002
Serbo-Croatian	840,799	1,760,095	35,936
Spanish	382,955	677,423	32,571
Swedish	78,411	131,693	15,030
Russian	473,482	1,343,760	18,774
Total	5,893,381	8,886,426	401,309

ing out-of-vocabulary words. MorphyNet also extends the UDer structure by indicating the affix and the semantic category for the target word when it can be inferred from the morpheme. Such information is again useful for subword regularization of derivationally rich languages, such as English.

Table 4 provides per-language statistics on MorphyNet data. The present version of the resource contains 10.6 million entries, of which 95% are inflections. Highly inflecting and agglutinative languages are dominating the resource as 55% of all entries belong to Finnish, Hungarian, Russian, and Serbo-Croatian. Language coverage above all depends on the completeness of Wiktionary, the main source of our data.

5 Evaluation

We evaluated MorphyNet through two different methods: (1) through *comparison to ground truth* and (2) through *manual validation* by experts.

Comparison to ground truth. The quality evaluation of morphology database is a challenging task due to many weird morphology aspects of languages evaluated (Gorman et al., 2019). As ground truth on inflections we used the *Universal Dependencies*[6] dataset (Nivre et al., 2016, 2017), which (among others) provides morphological analysis of inflected words over a multilingual corpus of hand-annotated sentences. McCarthy et al. (2018) built a Python tool[7] to convert these treebanks into UniMorph schema (Sylak-Glassman, 2016). We evaluated both UniMorph 2.0 and MorphyNet against this data (performing the necessary mapping of feature tags beforehand) over the 11 languages in the intersection of the two resources: Hungarian (Vincze et al., 2010), Catalan, Spanish (Taulé et al., 2008), Czech (Bejček et al., 2013), Finnish (Pyysalo et al., 2015), Russian (Lyashevskaya et al., 2016), Serbo-Croatian (De Melo, 2014), French (Guillaume et al., 2019), Italian (Bosco et al., 2013), Swedish (Nivre and Megyesi, 2007), and English (Silveira et al., 2014). Table 5 contains evaluation results over nouns, verbs, and adjectives separately, as well as totals per language. Missing data points (e.g. for Catalan nouns) indicate that UniMorph did not have any corresponding inflections. For languages and parts of speech where both resources provide data, MorphyNet always provides higher recall. The exception is Finnish because of our policy of not extracting conjugations with auxiliary and modifier words as separate entries (see Section 3.1). Overall, as

[6] https://universaldependencies.org/

[7] https://github.com/unimorph/ud-compatibility

Table 5: Inflectional morphology evaluation of MorphyNet against UniMorph on Universal Dependencies

Language	Resource	Noun			Verb			Adjective			Total		
		R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Catalan	UniMorph	-	-	-	71.9	99.3	83.4	-	-	-	21.3	99.3	35.1
	MorphyNet	66.0	98.4	79.0	73.8	99.1	84.6	48.2	99.6	65.0	64.3	98.8	77.9
Czech	UniMorph	28.2	99.1	43.9	9.5	18.1	12.5	17.6	44.8	25.3	21.0	72.7	32.6
	MorphyNet	33.2	98.9	49.7	28.2	93.8	43.4	36.1	98.1	52.8	34.2	98.0	50.7
English	UniMorph	-	-	-	96.1	90.9	93.4	-	-	-	28.3	90.9	43.2
	MorphyNet	81.5	99.1	89.4	97.1	96.8	96.9	85.3	99.7	91.9	83.2	98.8	90.3
French	UniMorph	-	-	-	70.6	98.5	82.2	-	-	-	20.6	98.5	34.1
	MorphyNet	80.2	98.6	88.5	94.4	98.5	96.4	60.1	94.6	73.5	79.7	97.9	87.9
Finnish	UniMorph	45.5	99.5	62.4	50.5	88.4	64.3	61.4	81.7	70.1	49.1	93.5	64.4
	MorphyNet	49.8	99.4	66.4	53.8	89.5	67.2	67.2	98.1	79.8	54.5	96.7	69.7
Hungarian	UniMorph	45.3	99.0	62.2	31.9	97.8	48.1	-	-	-	30.8	98.8	47.0
	MorphyNet	55.2	99.1	70.9	77.2	96.9	85.9	43.1	95.9	59.5	56.3	97.9	71.5
Italian	UniMorph	-	-	-	66.1	91.6	76.8	-	-	-	22.8	91.6	36.5
	MorphyNet	86.7	99.0	92.4	88.8	96.9	92.7	84.9	98.9	91.4	87.0	98.2	92.3
Serbo-Croatian	UniMorph	0.0	0.0	0.0	0.0	0.0	0.0	49.4	47.4	48.4	18.5	47.4	26.6
	MorphyNet	69.5	88.4	77.8	69.1	98.1	81.1	54.9	98.6	70.5	63.9	93.3	75.9
Spanish	UniMorph	-	-	-	93.0	99.8	96.3	-	-	-	32.1	99.8	48.6
	MorphyNet	88.3	99.2	93.4	97.0	99.5	98.2	81.9	99.2	89.7	89.7	99.3	94.3
Swedish	UniMorph	15.1	98.4	26.2	59.7	84.8	70.1	34.1	94.8	50.2	27.1	92.0	41.9
	MorphyNet	36.8	99.4	53.7	78.0	98.1	86.9	38.1	99.6	55.1	44.6	99.1	61.5
Russian	UniMorph	0.0	0.0	0.0	0.0	0.0	0.0	52.8	97.4	68.5	10.8	97.4	19.4
	MorphyNet	56.5	95.1	70.9	67.7	92.9	78.3	64.5	99.0	78.1	61.5	95.2	74.7

seen from Table 4, MorphyNet contains about 47% more entries over the 11 languages where it overlaps with UniMorph. In terms of precision, the two resources are comparable, except for Finnish (adjectives) and Swedish (adjectives and verbs) where MorphyNet appears to be significantly more precise.

UDer (Kyjánek et al., 2020) is a collection of individual monolingual resources of derivational morphology. Most of them have been carefully evaluated against their own datasets and offer high quality. We evaluated MorphyNet derivational data against UDer over the nine languages covered by both resources: French (Hathout and Namer, 2014), Portuguese (de Paiva et al., 2014), Czech (Vidra et al., 2019), German (Zeller et al., 2013), Russian (Vodolazsky, 2020), Italian (Talamo et al., 2016), Finnish (Lindén and Carlson, 2010; Lindén et al., 2012), Latin (Litta et al., 2016), and English (Habash and Dorr, 2003). Statistics and results are shown in Table 6. First of all, the overlap between MorphyNet and UDer is small, which is visible from our recall values relative to UDer that vary between 0.6% (Czech) and 59.5% (Italian). Among the languages evaluated, six were better covered by MorphyNet and the remaining three (Czech, German, and Russian) by UDer. The agreement between the two resources, computed as Cohen's Kappa, was 0.85 overall, varying between 0.74 (Finnish) and 0.97 (Portuguese). If we consider UDer as gold standard, we obtain precision figures between 87% and 99%.

Manual evaluation was carried out by language experts over sample data from five languages: English, Italian, French, Hungarian, and Mongolian. The sample consisted of 1,000 randomly selected entries per language, half of them inflectional and the other half derivational. The experts were asked to validate the correctness of source–target word pairs, of morphemes, as well as inflectional features and parts of speech (the latter for derivations). Table 7 shows detailed results. The overall precision is 98.9%, per-language values varying between 98.2% (Hungarian) and 99.5% (English). The good results are proof both of the high quality of Wiktionary data and of the general correctness of the data extraction and enrichment logic of MorphyNet. A manual checking of the incorrect entries revealed that most of them were due to the failure of extraction rules due to occasional deviations in Wiktionary from its own conventions.

6 Conclusions and Future Work

We consider the resource released and described here as an initial work-in-progress version that we plan to extend and improve. We are currently

Table 6: Derivational morphology evaluation of MorphyNet against Universal Derivations (UDer)

#	Language	MorphyNet	Univeral Derivations (UDer)		UDer ∩ MorphyNet	Recall	Precision	Kappa
1	French	37,203	Démonette	13,272	2,558	18.5	95.5	0.91
2	Portuguese	15,974	NomLex-PT	3,420	1,235	35.8	98.9	0.97
3	Czech	9,660	Derinet	804,011	5,347	0.6	94.1	0.88
4	German	23,867	DerivBase	35,528	5,878	15.6	93.5	0.87
5	Russian	36,922	DerivBase.RU	118,374	6,370	12.3	88.1	0.76
6	Italian	42,149	DerIvaTario	1,548	958	59.5	90.7	0.81
7	Finnish	37,199	FinnWordnet	8,337	2,664	30.6	87.0	0.74
8	Latin	9,191	WFL	2,792	4,037	14.0	93.7	0.87
9	English	200,365	CatVar	16,185	7,397	45.7	91.9	0.83
	Total	412,530		1,003,467	36,444	25.8	92.6	0.85

Table 7: Manual validation of language experts on MorphyNet

		Inflectional morphology			Derivational morphology			
#	Language	word pair	features	morphemes	trg words	POS	morphemes	Total
1	English	99.2	100.0	99.0	100.0	99.0	100.0	99.5
2	French	99.8	98.0	100.0	100.0	96.8	100.0	99.1
3	Hungarian	97.0	95.0	100.0	98.6	99.1	99.2	98.2
4	Italian	100.0	100.0	99.4	98.0	97.4	99.0	99.0
5	Mongolian	98.2	100.0	99.2	98.4	98.1	98.6	98.8
	Average.	98.8	98.6	99.5	99.0	98.1	99.4	98.9

working on increasing the coverage to 20 languages. We also plan to extend MorphyNet data with additional features and the semantic categories of words (e.g. animate or inanimate object, action) inferred from derivations. We are planning to conduct a more in-depth study of our evaluation results, especially with respect to UDer where it is not yet clear whether the occasional lower precision figures (87% for Finnish, 88% for Russian) are due to mistakes in MorphyNet, in the UDer resources, or are caused by other factors.

A major piece of ongoing work concerns the representation of MorphyNet derivational data as a lexico-semantic graph, as it is done in wordnets (Miller, 1998; Giunchiglia et al., 2017) where derivationally related word senses are interconnected by associative relationships. This effort, justifying the *-Net* in the name of our resource, will allow us to address completeness issues in existing wordnets by extending them by morphological relations and derived words.

We are happy to offer the MorphyNet extraction logic to be reused on a community basis. As extending the tool with new Wiktionary extraction rules is straightforward, we hope that the availability of the tool will allow language coverage to grow even further. We also hope that the MorphyNet data and the extraction logic can serve existing high-quality projects such as UniMorph and UDer.

References

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. A large and evolving cognate database. *Language Resources and Evaluation*.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. Cognet: a large-scale cognate database. In *Proceedings of ACL 2019, Florence, Italy*.

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovskỳ, Anna Nedoluzhko, Jarmila Panevová, et al. 2013. Prague dependency treebank 3.0.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of*

the Association for Computational Linguistics, 5:135–146.

Cristina Bosco, Montemagni Simonetta, and Simi Maria. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, et al. 2017. Conll-sigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30.

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154. Citeseer.

Valeria de Paiva, Livy Real, Alexandre Rademaker, and Gerard de Melo. 2014. Nomlex-pt: A lexicon of portuguese nominalizations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2851–2858.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017.

Kyle Gorman, Arya D McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but ok: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies. *Traitement Automatique des Langues*, 60(2):71–95.

Nizar Habash and Bonnie Dorr. 2003. Catvar: A database of categorial variations for english. In *Proceedings of the MT Summit*, pages 471–474. Citeseer.

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

Inxight. 2005. Linguistx natural language processing platform.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126.

Lukáš Kyjánek, Zdeněk Žabokrtskỳ, Magda Ševčíková, and Jonáš Vidra. 2020. Universal derivations 1.0, a growing collection of harmonised word-formation resources. *The Prague Bulletin of Mathematical Linguistics*, (115):5–30.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Krister Lindén, Jyrki Niemi, and Mirka Hyvärinen. 2012. Extending and updating the finnish wordnet. In *Shall We Play the Festschrift Game?*, pages 67–98. Springer.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLIC–IT 2016)*, pages 185–189.

Olga Lyashevskaya, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, Elena Shakurova, et al. 2016. Universal dependencies for russian: A new syntactic dependencies tagset. *Lyashevkaya, K. Droganova, D. Zeman, M. Alexeeva, T. Gavrilova, N. Mustafina, E. Shakurova//Higher School of Economics Research Paper No. WP BRP*, 44.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. Unimorph 3.0: Universal morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3922–3931.

Arya D McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101.

Eleni Metheniti and Günter Neumann. 2020. Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3905–3912.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre and Beata Megyesi. 2007. Bootstrapping a swedish treebank using cross-corpus harmonization and annotation projection. In *Proceedings of the 6th international workshop on treebanks and linguistic theories*, pages 97–102. Citeseer.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *International Conference on Text, Speech, and Dialogue*, pages 237–245. Springer.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–172.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014.

A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

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

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*.

Viktor Trón, Gyögy Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: open source word analysis. In *Proceedings of Workshop on Software*, pages 77–85.

Jonáš Vidra, Zdeněk Žabokrtskỳ, Magda Ševčíková, and Lukáš Kyjánek. 2019. Derinet 2.0: towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank.

Daniil Vodolazsky. 2020. Derivbase. ru: A derivational morphology resource for russian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3937–3943.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2017. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. Derivbase: Inducing and evaluating a derivational morphology resource for german. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1201–1211.