Deciphering and Characterizing Out-of-Vocabulary Words for Morphologically Rich Languages

Georgie Botev, Arya D. McCarthy, Winston Wu, and **David Yarowsky**
Johns Hopkins University

Abstract

This paper presents a detailed foundational empirical case study of the nature of out-of-vocabulary words encountered in modern text in a moderate-resource language such as Bulgarian, and a multi-faceted distributional analysis of the underlying word-formation processes that can aid in their compositional translation, tagging, parsing, language modeling, and other NLP tasks. Given that out-of-vocabulary (OOV) words generally present a key open challenge to NLP and machine translation systems, especially toward the lower limit of resource availability, there are useful practical insights, as well as corpus-linguistic insights, from both a detailed manual and automatic taxonomic analysis of the types, multidimensional properties, and processing potential for multiple representative OOV data samples.

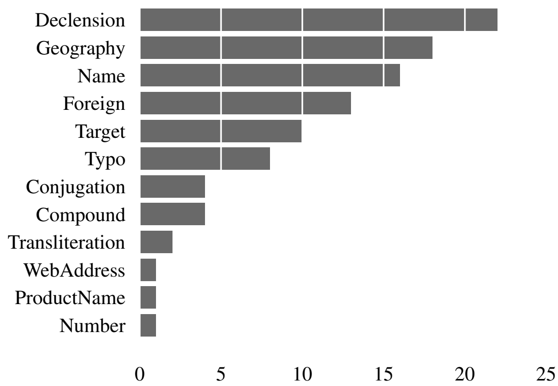

Figure 1: Taxonomized distribution of out-of-vocabulary types in Bulgarian Wikipedia, random sample of 100 types

1 Introduction

Even in a familiar language, unfamiliar words cause trouble for machine processing or comprehension of text. Any dictionary is innately incomplete in its coverage, unable to provide novel coinages and exhaustive forms. Without finding the word in a dictionary, the surface form and context afford only weak evidence for its meaning. The situation is even worse for languages other than English, especially morphologically rich languages, for two reasons: first, there is usually less annotated data available; and second, the coverage of such data is much lower due to the high number of different forms. Moreover, many words not found in even a small training corpus are in fact related to quite common words by processes such as inflection, derivation, compounding, or misspelling.

In the work described herein, we therefore concentrate on the problem of characterizing unknown words in terms of the processes by which they arise, and especially the relative frequencies at which such processes occur. This informs us of the *distribution* of out-of-vocabulary (OOV) words with respect to different dictionary sources.

To do so, we conduct a study on a sample of two Bulgarian language corpora annotated by a native speaker. Rather than treat OOV tokens as a monolithic and undifferentiated problem, we progressively apply multi-faceted linguistic analyses to these corpora, characterizing both the words that these analyses explain and words yet to be explained, which we shall call the **residual vocabulary**. Our methods are a mixture of the vintage and the vogue: specialized edit distances, composition of finite-state transducers, a noisy channel model for language identification fitted with empirical Bayes, and neural network–based part of speech taggers. Collectively, our processes accurately explain more than two in three (69%) unknown Bulgarian words in a held-out set according to whether they are proper names, inflections, derivations, compounds, foreign words, or misspellings (as illustrated in both Figures 1 and 3, discussed in more depth in §5). We release our native speaker–annotated lexicon, intermediate analyses, and software at `www.github.com/gbotev1/bg`.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5309–5326
October 12–17, 2022.

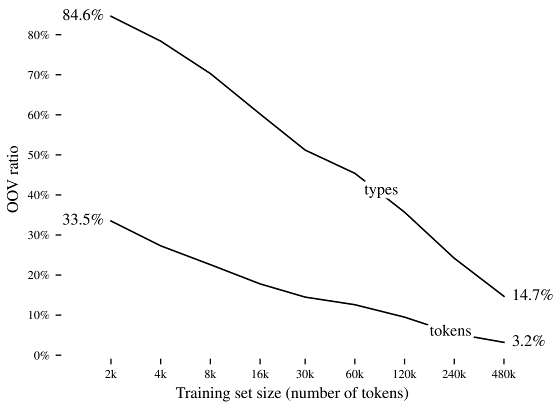

Figure 2: OOV rate as a function of data size (Bulgarian Wikipedia). Note the logarithmic horizontal axis.

2 Motivation and Related Work

Previously unseen words often represent a significant portion of the vocabulary, due in part to the Zipfian nature of language. Figure 2 illustrates this for various vocabulary sizes. Note that for the Bulgarian training data, the OOV rate remains high for both tokens (corpus instances of words) and types (vocabulary words) as found in a held-out set of 20,000 tokens. The rates are computed ignoring capitalization, punctuation, and numbers, so that these do not skew the count of unknown words.

The frontier of natural language processing as an engineering discipline has adopted information-theoretic subword tokenization (Sennrich et al., 2016; Kudo, 2018) to constrain the vocabulary size and provide a representation of all words, preventing any words from being out-of-vocabulary. Because such models dominate so much of the field of NLP, one may ask what value there is in analyzing the residual vocabulary today. Foremost, there is the corpus-linguistic and lexicographic value of characterizing this aspect of text: it is instructive about the patterns of lacunae in dictionaries or word formation processes in particular domains such as color (McCarthy et al., 2019). There are engineering applications as well. In languages with insufficient data for training large neural machine translation systems (Mueller et al., 2020) (or even for fine-tuning to new languages; see Lee et al., 2022), statistical methods dominate (Koehn and Knowles, 2017). The methods described in this paper are of value for populating the phrase tables of statistical MT models beyond what can be done with existing bilingual dictionaries, as in Vilar et al. (2007) who address spelling variants by online retokenization, or de Gispert (2006) who aims to reduce morphological variety. Moreover, entity linking and the use of gazetteers in named entity recognition both benefit from exact word representations. We underscore the fact that resource-poor languages are the norm, not the exception. Out of the world's roughly 7,000 languages, only 216 have more than 1,000 gloss definitions in Wiktionary, a popular multilingual dictionary.[1] For the remaining ≈6,800 data-poor languages, unknown words are not only neologisms and proper names; items of the core vocabulary are regularly absent from bilingual dictionaries or small but extant corpora.

Lexicon stratification, the splitting of the lexicon based on words' origin and degree of assimilation into the language (Ito and Mester, 1995), is a powerful technique to hone the processing of OOV words (Tsvetkov and Dyer, 2015). The four identified levels are the core vocabulary, the partially assimilated words, the fully assimilated words, and peripheral lexemes. This paper proffers empirical relative frequencies of these degrees and showcases a series of models that roughly correspond to these degrees.

3 The Bulgarian Language

Bulgarian is a member of the South Slavic branch of the Indo-European family, written in the Cyrillic script. As a member of the Balkan sprachbund, its lexis[2] and grammar have been influenced by areal effects. It thus displays several traits uncharacteristic of other Slavic languages (except Macedonian) which affect the apparent size of the lexicon: a postposed definite article marked for gender, the use of clitic pronouns, a lack of verbal infinitive, and limited case declension (Corbett and Comrie, 2003).

As a case study, Bulgarian is useful because it uses several widespread strategies for word formation. Its rich verbal morphology yields over 50 forms per verb lexeme. Derivational affixation and compounding are prevalent processes. In fact, derivation for nouns is both productive and regular (Krushkov, 2001). Finally, a significant fraction of the Bulgarian lexis is borrowed from Russian, Greek, or other languages, especially in technical contexts.

These properties have made Bulgarian a focus for linguistic examination and an area of interest in natural language processing. For example, Slavcheva (2003) devise a rich morphological tag set for Bulgarian verbs. Koeva et al. (2020) build a richly anno-

[1] https://en.wiktionary.org/wiki/Wiktionary:Statistics

[2] We distinguish between the *lexis*, i.e., the set of all words in a language, and the *lexicon*, i.e., the set of all lexemes.

tated corpus of web-crawled Bulgarian. Popov et al. (2020) construct a battery of models for multi-stage analysis of Bulgarian text, including lemmatization, parsing, and named entity recognition. Notably, the latter relies on a dictionary-based lemmatizer with a statistical model for fallback.

In contrast to these works, which offer an engineering approach to modeling Bulgarian, our work relies on computational tools insofar as they help characterize *properties* of Bulgarian text. Namely, we explore the relative frequency of various processes by which words—especially unknown words—arise in naturally occurring Bulgarian text.

4 Data

For our study, we need a large and representative corpus of Bulgarian text. We use the entirety of Bulgarian Wikipedia, which contains 1.3 million word types and 73.6 million word tokens (type–token ratio 0.018) after tokenization; a random sample of these is summarized in Figure 1.

We also must define the set of known words. We merge three broad-coverage bilingual dictionaries:

LanguageNet. 364,327 entries covering 155,703 unique English words.[3]

PanLex. 180,023 entries covering 70,986 unique English words (Baldwin et al., 2010).

Wiktionary. 51,537 entries covering 22,856 unique English words. We extract these with `Yawipa` (Wu and Yarowsky, 2020a,b).

In aggregate, these cover 165,644 unique Engish words, with a median number of translations 1 and mean approximately 2.360.[4]

To identify the residual vocabulary, we remove from Bulgarian Wikipedia all entries in our dictionaries as well as non-alphabetic entries, leaving 371,475 novel words—about one in every 200 tokens.[5] A random sample of 100 is summarized in Figure 3. The complete word lists and analyses are given in Appendix A. All annotations were validated or adjudicated by a non-author professional Bulgarian translator who is a native speaker.

What becomes immediately apparent is that the residual vocabulary after dictionary entries are removed comes from five major groups: morphological variants of other words, foreign words, misspellings, compound words, and proper names like place names or people. We devise computational approaches to tackle these five major categories.

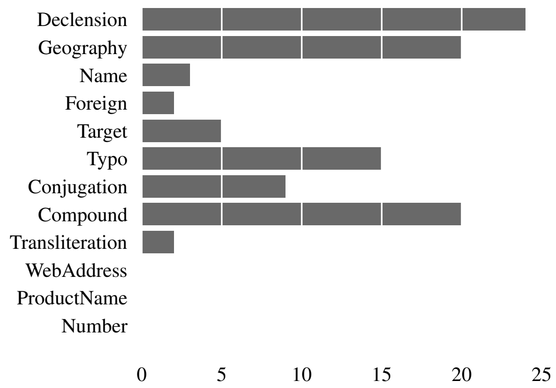

Figure 3: Taxonomized distribution of out-of-vocabulary types in Bulgarian Wikipedia that are unseen in Wiktionary, PanLex, and LanguageNet; random sample of 100 types. Compare with Figure 1.

Because we discovered an abundance of Russian words interspersed in the Bulgarian text, we also extract Russian–English bilingual entries from the same three dictionaries. We find 232,094 entries in Wiktionary covering 75,284 unique English words; 2,379,638 entries in PanLex covering 859,279 unique English words, and 1,633,709 unique entries in LanguageNet covering 879,438 unique English words. Their union covers 932,738 unique English words, with potentially multiple Russian candidate translations. The median number of translations was one, and the mean was 1.888.

Preprocessing To identify Bulgarian tokens in context, we first preprocess the text using the rule-based spaCy sentence segmenter and tokenizer (Honnibal and Montani, 2017). We found this to be faster than the Stanza neural tokenizer (Qi et al., 2020). We use Stanza for POS tagging, though its poor performance motivates the 'vintage' models we introduce below. In preliminary experiments, we also explored `TreeTagger` (Schmid, 1994, 1999).[6]

[3] `uakari.ling.washington.edu/languagenet/`

[4] The English pronoun *we* had the most translations: 306, due largely to inappropriate Bulgarian translations in LanguageNet which were first-person plural verb forms.

[5] This is approximately the same rate as Min and Wilson (1998) observe; they report that at this rate an out-of-vocabulary word occurs in 12% of sentences.

[6] Several avenues exist to improve part-of-speech tagging with minimal available resources. The most notable is projecting part-of-speech annotations across unsupervised word alignments into the language of interest, then using these silver annotations to train a new tagger (Yarowsky and Ngai, 2001; Täckström et al., 2013; Wang and Manning, 2014; Buys and Botha, 2016; Nicolai and Yarowsky, 2019; Eskander et al., 2020). Such methods could either complement a tagger such as Stanza trained in the language of interest via classifier combination or

We normalize all text to Unicode NFKD form to increase coverage.[7] This also allowed us to remove accents, which were predominantly used to mark stress. We subsequently remove tokens with any letter not in the Bulgarian alphabet. While this removes a few interesting cases like mp3-файлове 'MP3 files' and 2-то 'the second [thing]', on the whole the eliminations were useful: filtering URLs, email addresses, and also less structured non-words.

We found the need to preprocess the dictionaries by *hyphen flattening*. If a dictionary entry begins or ends with a hyphen, indicating that it is a prefix or suffix, we associate it with its non-hyphenated translational counterpart. For instance, the nonsensical English entry 'pra' is linked to the Bulgarian transliteration 'пра', and the Bulgarian prefix 'пра-' is correctly (and uniquely) associated with the English prefix 'great-'. After flattening, the Bulgarian entry 'пра' would have both 'pra' and 'great' listed as candidate translations. This preprocessing both reduces the dictionary's size and is crucial to increasing the impact of the compound analysis (§5.5).

Moreover, we define a heuristic to eliminate Old Bulgarian words, based on a 1945 orthographic reform that forbids word-final 'ъ'. Inspecting a sample of 50 words captured by this heuristic reveals that while none of the words filtered here were modern Bulgarian, 44% were in fact Old Bulgarian. The remainder were transliterations (the "unassimilated foreign words" of Tsvetkov and Dyer, 2015) from disparate languages: Italian (18%), Turkish[8] (16%), Kazakh (6%), Chinese (6%), Albanian (4%), and single exemplars of Irish, Portuguese, and Moldovan.[9]

5 Modeling and Analysis

This work by its nature differs from a great deal of the empirical work in natural language processing. The object of its inquiry is language itself, not computational models, and so we do not evaluate in the standard positivist paradigm of comparing scores on standard benchmarks. Instead, we build computational models to help sift through the millions of words in our corpus, study their distribution, and discover what can be modeled about them. After all, if we seek to tame the lexis, we must first understand it. In this regard, we follow the guidance of Hajič and Hajičová (2007) who recognize the value of objective assessment of models or theories on annotated corpora, grounded in linguistic intuition about the phenomenon to be modeled. Our characterization of the residual vocabulary helps to extend the linguistic intuition in an empirical manner.

The modularity of our approach lets us leverage prior tools and research in the language, and components can be upgraded as better models are devised (e.g., Nicolai et al., 2020 and Wiemerslage et al., 2022 for morphological analysis, Lewis et al., 2020 for inferring cognates). Moreover, disparate models for a single word formation process can be combined *in situ* via classifier combination or meta learning.

While many of the tools we use are tailored to the Bulgarian language, such as hand-crafted derivational rules from a grammar, in principle our approach makes minimal assumptions about the nature of the language. It could easily be adapted to other Slavic languages or, given sufficient prior typological information, other written languages writ large.

The overall sequence of method application is given in Figure 4. In the following sections, we elaborate on the most telling among these: language identification, then modeling morphology, misspellings, and compounds. Table 1 gives complete analyses for the held-out set of Wikipedia residual vocabulary, coupled with computer-predicted analyses.

5.1 Russian language filtering

A substantial fraction of the residual vocabulary is direct borrowings (loanwords) from other languages; cross-lingually this can be between 10% and 70% of the lexicon (Haspelmath and Tadmor, 2009). While our preprocessing eliminates several directly imported words that were not transliterated, a significant number of borrowings comes from Russian, which largely shares an alphabet with Bulgarian.

Some words can be clearly identified as non-Bulgarian by means of straightforward linguistic heuristics. The filtered words were mostly Russian, with a few exceptions that were Ukrainian or Serbian. We employ the following heuristics:

1. A Bulgarian word cannot begin or end with the soft sign 'ь'.
2. If the soft sign 'ь' occurs in the middle of a

annotate the language in the absence of in-language annotations.

[7] Kyle Gorman notes an increase of 0.3 in labeled attachment score for dependency parsing of Hindi, purely from normalization: http://www.wellformedness.com/blog/text-encoding-issues-in-universal-dependencies/.

[8] Note that due to both areal effects in the Balkan sprachbund and Bulgaria's past as an Ottoman territory, many Turkish lexemes have entered the Bulgarian lexicon as *fully assimilated* lexical items (Ito and Mester, 1995).

[9] We will not engage with the question of whether Romanian and Moldovan are dialects or separate languages; here, we use this as a shorthand for the Daco-Romance language written with the Cyrillic script.

Index	Word	Human Trans.	Alg. Trans.	Human Type	Human Sub-Type	Alg. Type	Alg. Sub-Type	Features	POS
1	звероферма	beast farm	beast|@	Compound	–	Compound	Partial	FEM	NOUN
2	неоспорван	uncontested	new|contested	Compound	–	Compound	–	–	ADJ
3	солокариера	solo career	solo|career	Compound	–	Compound	–	FEM	NOUN
4	битовофекални	household faeces	household faeces	Compound	–	Compound	–	PL	NOUN
5	светлооранжев	light orange	light|orange	Compound	–	Compound	–	MASC	ADJ
6	контраразузнаване	counter intelligence	counter|intelligence	Compound	–	Compound	–	NEUT	NOUN
7	удавил	drowned	–	Conjugation	–	Conjugation	–	–	PART
8	завзели	conquered	–	Conjugation	–	Conjugation	–	PL	PART
9	далекомером	distance meter	–	Foreign	Russian	Foreign	Russian	–	NOUN
10	мацелумът	Macellum	kitten|noise	Geography	Italian	Compound	–	–	PROPN
11	койбалската	Koybalska	koybali|times	Geography	Russian	Compound	–	FEM+DEF	ADJ (Proper)
12	горнобродчани	Inhabitants of Gorno Brod	gorno|brod	Geography	Bulgarian	Compound	–	PL	NOUN
13	костойчиновият	Kostoychinov	kostov|new	Geography	Bulgarian	Compound	–	DEF	ADJ (Proper)
14	ашоташен	Ashotashen	ashot	Geography	Armenian	Declension	Fuzzy	–	PROPN
15	Уайя	Huaya	–	Geography	Mexican	Proper	Likely	–	PROPN
16	Бишина	Bishina	–	Geography	Serbian	Proper	Likely	–	PROPN
17	Кастей	Castei	–	Geography	Italian	Proper	Likely	–	PROPN
18	Бозовая	Bozovaya	–	Geography	Bulgarian	Proper	Likely	–	PROPN
19	Исаково	Isakovo	–	Geography	Russian	Proper	Likely	–	PROPN
20	Кеседжи	Kesdji	–	Geography	Greek	Proper	Likely	–	PROPN
21	Сигнора	Signora	–	Geography	Italian	Proper	Likely	–	PROPN
22	Соулънт	Solent	–	Geography	English	Proper	Likely	–	PROPN
23	Харагуа	Jaragua	–	Geography	Dominican Republic	Proper	Likely	–	PROPN
24	Ябълчице	Yabaltchitse	–	Geography	Bulgarian	Proper	Likely	–	PROPN
25	Байенбург	Bayenburg	–	Geography	German	Proper	Likely	–	PROPN
26	Петъчници	Petachnitsi	–	Geography	Bulgarian	Proper	Likely	–	PROPN
27	Валтотопион	Valtotopion	–	Geography	Greek	Proper	Likely	–	PROPN
28	Казакевичево	Kazakevichevo	–	Geography	Bulgarian	Proper	Likely	–	PROPN
29	енорияшкото	parish	parish|@	Declension	–	Compound	–	MAS+DEF	ADJ
30	апоплектичната	the apoplectic	Apoel|Ethic	Declension	–	Compound	–	DEF	ADJ
31	будени	awake	awake	Declension	–	Declension	Simple	PL	ADJ
32	ашерова	ashura	ashur	Declension	–	Declension	Fuzzy	FEM	ADJ
33	подобия	similarity	similarity	Declension	–	Declension	–	PL	NOUN
34	потника	tank top	tank top	Declension	–	Declension	Simple	MASC+DEF	NOUN
35	пролози	prologues	mercury	Declension	–	Declension	Fuzzy	PL	NOUN
36	грацията	The grace	grace	Declension	–	Declension	Simple	FEM+DEF	NOUN
37	ослепяло	became blind	blindness	Declension	–	Declension	–	NEUT	PART
38	смутното	turmoiled	turmoil	Declension	–	Declension	Simple	NEUT+DEF	ADJ
39	сталинци	stalinists	stalin	Declension	–	Declension	Fuzzy	PL	NOUN
40	суглинки	loams	suli	Declension	–	Declension	Simple	FEM+PL	NOUN
41	тръбеста	tubular	tubular	Declension	–	Declension	Fuzzy	FEM	ADJ
42	записната	pertaining to recording	recording	Declension	–	Declension	Simple	FEM+DEF	ADJ
43	потурчено	stamp down	stamp down	Declension	–	Declension	Simple	NEUT	ADJ
44	неголямото	not so big	rare	Declension	–	Declension	Simple	DEF	ADJ
45	еклектиката	eclecticism	eclectic	Declension	–	Declension	Fuzzy	FEM+DEF	NOUN
46	съблечената	The undressed	undressed	Declension	–	Declension	Simple	FEM+DEF	ADJ
47	персистиращи	persistent	persistence	Declension	–	Declension	Fuzzy	PL	NOUN
48	превърналата	the one that became	became	Declension	–	Declension	Simple	FEM+DEF	ADJ
49	кибернетизация	cybernetization	cybernetics	Declension	–	Declension	Fuzzy	FEM	NOUN
50	мултиетническия	the multiethnic	multiethnic	Declension	–	Declension	Simple	DEF	ADJ
51	Шотландска	Scotish	–	Declension	–	Proper	Standard	FEM	ADJ (Proper)
52	кодокан	Kodokan	kodo|kan	Name	School	Compound	–	–	PROPN
53	айдънидите	Aydin	@|nit	Name	Dynasty	Compound	–	–	PROPN
54	аморейско	Amorite	–	Name	Ethnicity	Foreign	Russian	NEUT	ADJ (Proper)
55	себрите	Ancestors of Serbians	seri	Name	Tribe	Declension	Fuzzy	PL+DEF	PROPN
56	ЦТА	Central Tibet Administration	–	Name	Organization	Proper	Likely	–	PROPN
57	Азел	Azel	–	Name	Person	Proper	Likely	–	PROPN
58	Юджи	Yuji	–	Name	Person	Proper	Likely	–	PROPN
59	ЗЕЛПО	ZELPO	–	Name	Building	Proper	Likely	–	PROPN
60	Какаи	Kakai	–	Name	Person	Proper	Likely	–	PROPN
61	Лопов	Lopov	–	Name	Person	Proper	Likely	–	PROPN
62	Мусан	Musan	–	Name	Person	Proper	Likely	–	PROPN
63	Пийбо	Peebo	–	Name	Person	Proper	Likely	–	PROPN
64	Дарбес	Darbez	–	Name	Person	Proper	Likely	–	PROPN
65	Прицак	Pritsak	–	Name	Person	Proper	Likely	–	PROPN
66	Халиду	Halidu	–	Name	Person	Proper	Likely	–	PROPN
67	Бейтлър	Beightler	–	Name	Person	Proper	Likely	–	PROPN
68	Вигберт	Witbert	–	Name	Person	Proper	Likely	–	PROPN
69	Евтахий	Evtahiy	–	Name	Person	Proper	Likely	–	PROPN
70	Оливиър	Olivier	–	Name	Person	Proper	Likely	–	PROPN
71	Ризберг	Rieseberg	–	Name	Person	Proper	Likely	–	PROPN
72	Памтивек	Pamtivek (colloquial for ancient)	–	Name	Book	Proper	Likely	–	NOUN
73	Харелсън	Harrelson	–	Name	Person	Proper	Likely	–	PROPN
74	Цибисова	Cybisowa	–	Name	Person	Proper	Likely	–	PROPN
75	Гроновиус	Gronovius	–	Name	Person	Proper	Likely	–	PROPN
76	Орочимаро	Orochimaro	–	Name	Person	Proper	Likely	–	PROPN
77	Настоплиси	Nastoplisi	–	Name	Person	Proper	Likely	–	PROPN
78	Присовский	Prisovskii	–	Name	Person	Proper	Likely	–	PROPN
79	Гутомсдатер	Gutomsdater	–	Name	Person	Proper	Likely	–	PROPN
80	Хаджипопова	Hadjipopova	–	Name	Person	Proper	Likely	–	PROPN
81	Христодоров	Christodorov	–	Name	Person	Proper	Likely	–	PROPN
82	буганин	gohst (archaic)	buga|nin	Target	–	Compound	–	MASC	NOUN
83	реверсира	reverse	reversed	Target	–	Declension	Fuzzy	–	VERB
84	Фесенджан	Fesenjan	–	Target	Iranian	Proper	Standard	–	PROPN
85	баунс	bounce	–	Transliteration	English	Misspelling	Substitution	MASC	NOUN
86	потът	the pot	floor, sweat, sex	Transliteration	English	Misspelling	Substitution	MASC+DEF	NOUN
87	футуризмо	futurism	The futurism	Transliteration	Italian	Misspelling	Substitution	–	NOUN
88	Форматър	formatter	–	Transliteration	English	Proper	Likely	–	NOUN
89	фрагмет	fragment	fra|@, @|met	Typo	Omission	Compound	–	MASC	NOUN
90	денудаци	akin to denudational	day|odd person	Typo	Omission	Compound	–	–	NOUN
91	клавесинистикатегория	category|harpsichordists	category|harpsichordists	Typo	Concatenation	Compound	–	PL	NOUN
92	реакциии	reactions	reactions	Typo	Misspelling	Declension	–	PL	NOUN
93	същотото	same	same	Typo	Addition	Declension	–	NEUT+DEF	ADJ
94	домантите	The Odomanti	tomatoes	Typo	Omission	Declension	–	PL+DEF	NOUN
95	ренгеново	x-ray	reinette	Typo	Omission	Declension	Fuzzy	–	ADJ
96	дестващ	acting	–	Typo	Omission	Misspelling	–	MASC	PART
97	отркрили	discovered	discovered	Typo	Addition	Misspelling	–	PL	PART
98	юзозападна	Southwestern	Southwestern	Typo	substitution	Misspelling	–	FEM	ADJ
99	Памятник	Monument	–	Typo	Substitution	Proper	Standard	MASC	NOUN
100	ПашиКатегория	Pasha Category	–	Typo	Concatenation	Proper	Likely	MASC	NOUN

Table 1: Manual classification of 100 randomly sampled words after classifying all of Bulgarian Wikipedia.

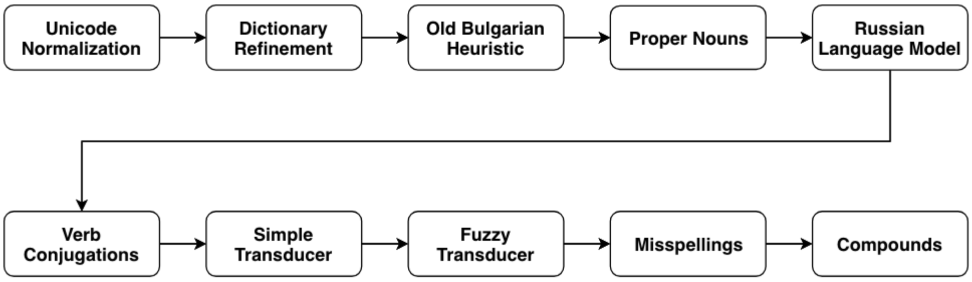

Figure 4: Sequence of methods applied to computationally analyze residual vocabulary

word, it must be followed by an ‘о’. This is the only character that may follow the soft sign in modern Bulgarian. In Russian, however, many characters are attested following ‘ь’ (e.g., улыбаться ‘to smile’ and семья ‘a family’).

For words not covered by these heuristics, we require a different approach to distinguish them. Cognate identification and transliteration empirically identify borrowings poorly (Ciobanu and Dinu, 2015; Tsvetkov et al., 2015). We instead employ language identification to disambiguate the remainder as Bulgarian or Russian words. We use a noisy channel model of the language ℓ of word form $\boldsymbol{\xi}$:

$$p_\theta(\ell \mid \boldsymbol{\xi}) \propto p_\theta(\boldsymbol{\xi} \mid \ell)\, \pi(\ell).$$

In factoring this generative model, we use character 5-gram models as the language models $p_\theta(\boldsymbol{\xi} \mid \ell)$. The Bulgarian model is trained on Bulgarian ParlaMint 1.0, which comprises 10.5 million tokens covering 123,000 word types. The Russian model is trained on the Russian SynTagRus Universal Dependencies data, which comprises 496,000 tokens and 94,000 word types. The prior probability $\pi(\ell)$ is optimized on the data; that is, we use empirical Bayes to infer a point estimate.

After this process, every one of 50 randomly sampled non-Bulgarian words was filtered as foreign, though some were Ukrainian or Slovenian instead of Russian. We note that 15 of these words were ambiguous; their character sequences could have represented valid Bulgarian or Russian words.

5.2 Verbal morphology

While Bulgarian nominal declension is much simpler than its Slavic sibling languages (presenting only nominative and vocative cases) (Gribble, 1987; Townsend and Janda, 1996), its verbal conjugation system is rich, embodying “the morphologically richest and most problematic part-of-speech category” (Slavcheva, 2003). Bulgarian verbs reflect voice, tense, mood, person, number, and evidentiality.

To analyze Bulgarian verbs, we construct a finite-state transducer that builds on the UniMorph project (Sylak-Glassman et al., 2015a,b; Kirov et al., 2016, 2018; McCarthy et al., 2020) and Apertium (Forcada et al., 2011; Forcada and Tyers, 2016).[10] This enables fast, interpretable analysis by composition and union of machines. Composition corresponds to application of a morphological rule (Roark and Sproat, 2007), and union collects alternative rules (or candidate manifestations of a single rule) into one machine. Our finite-state transducer is designed to map inflected word forms to their citation forms (their *lemmas*), if the word forms were tagged as verbs by Stanza. We construct one finite-state transducer for each form–lemma pair in UniMorph and Apertium, then take the union of these machines.

Transforming a word $\boldsymbol{\xi}$ to its citation form is equivalent to composing a finite-state acceptor representing $\boldsymbol{\xi}$ with the transducer. If the two cannot compose (because $\boldsymbol{\xi}$ is not in the domain of definition (i.e., input language) of the transducer), then we do not suppose that $\boldsymbol{\xi}$ is an inflected verb form.

When applied to identified verbs in the residual vocabulary, a spot check of 50 supposed Bulgarian verbs shows that 46 are correctly predicted. Of the remaining four, two are Russian words that passed through the filter from §5.1. The others are охрени ‘ocher’ (a plural adjective) and *собено, a misspelling of the Bulgarian adverb особено ‘specifically’.

5.3 Derivational morphology

Bulgarian has a productive set of derivational processes. Following the efficacy of the transducer for inflectional morphology, we introduce one for derivational morphology. We draw on the 22 derivational rules in Manova (2010) which explored the parsabil-

[10]UniMorph is a collection of morphological lexica in 167 languages, annotated in a cross-lingually consistent schema. Apertium is a rule-based machine translation system which includes a finite-state morphological analyzer and generator.

$$d_{\mathbf{x},\mathbf{y}}(i,j) = \min \begin{cases} 0 & \text{if } i = j = 0, \\ d_{\mathbf{x},\mathbf{y}}(i-1,j)+1 & \text{if } i > 0, \\ d_{\mathbf{x},\mathbf{y}}(i,j-1)+1 & \text{if } j > 0, \\ d_{\mathbf{x},\mathbf{y}}(i-1,j-1)+\mathbf{1}_{(\mathbf{x}_i \neq \mathbf{y}_j)} & \text{if } i,j > 0, \\ d_{\mathbf{x},\mathbf{y}}(i-2,j-2)+1 & \text{if } i,j > 1 \text{ and } \mathbf{x}_i = \mathbf{y}_{j-1} \text{ and } \mathbf{x}_{i-1} = \mathbf{y}_j, \end{cases}$$

Figure 5: Recurrence relation the Damerau–Levenshtein distance between two strings $\mathbf{x}$ and $\mathbf{y}$. The dynamic program to tractably compute this is a modification of the Wagner–Fisher algorithm (1975) for Levenshtein distance.

ity hypothesis (Hay, 2001; Aronoff and Fuhrhop, 2002) for Bulgarian. Patseva (2017) was also a basis for derivational rules.

Composing the finite-state transducer for derivational analysis with itself, or with a finite-state transducer for modeling inflections, expands the coverage by capturing forms with multiple derivations, as is the relationship between хиндуистките 'the Hinduistics' and хинду 'Hindu':

хинду → хиндуист	(nominal derivation)
хиндуист → хиндуистка	(diminutive feminine)
хиндуистка → хиндуистки	(plural)
хиндуистки → хиндуистките	(definite article)

Such considerations are crucial because derived forms may themselves be inflected. Moreover, certain forms are more amenable to derivation. For instance, adverbs are often formed from the neuter singular form of adjectives, except for adjectives that end in -ки. These motivate a single transducer to consider the two jointly (Fischer et al., 2016).

This model of morphology is 68% accurate on a random sample. While some errors are due to misspellings, it also ignores stem alterations which may arise but are not encoded in the derivational transformations. While fine-tuning the transduction rules to handle cases like мед 'copper' → медникар 'coppersmith' or злато 'gold' → златар 'goldsmith' is possible based on prior knowledge, the approach gives a reasonable grounding in using the available linguistic resources for a language.

5.4 Misspelling

The analysis and recovery of misspellings has a long history in the computational processing of language (McIlroy, 1982; Kernighan et al., 1990; Kukich, 1992). Rather than simply *identifying* misspellings, which can be easily done by checking against an existing wordlist, we also seek to identify the correct spelling of the misspelled word. To do so, we employ the Damerau-Levenshtein distance (Damerau, 1964), a modification of Levenshtein's edit distance that also allows character transpositions as an edit operation. It is well known that transposition errors (e.g. **lang**au**ge* instead of *language*) are common typing errors (Salthouse, 1984, 1986), and the Damerau–Levenshtein distance gives a more parsimonious backtrace for them.

In the residual space, we identify misspellings as words with a Damerau–Levenshtein distance of 1 from an item in the vocabulary. Exactly computing the Damerau–Levenshtein distance requires a nontrivial extension of the standard edit distance (see Figure 5); however, the asymptotic complexity remains proportional to the product of the string pair's lengths—as in the standard edit distance.

We find that one in six words from the residual vocabulary of the Wikipedia corpus is a misspelling of a word into a non-word (Figure 3). To decipher the meanings of these words, we link them to existing words in the Bulgarian vocabulary by finding the in-vocabulary word with the smallest Damerau–Levenshtein distance. On a random sample of 50 Bulgarian words classified as misspellings (Table A.3), 35 of these were indeed misspellings (for an accuracy of 70%). The remainder were largely transliterations, inflected forms of verbs that were not identified via the methods described in §5.2, and some proper nouns.

Our approach targets correcting the spellings of non-words into valid words. A context-driven model could also identify misspellings of words into other words which are valid but infelicitous.

5.5 Compounds

Finally, we consider the word formation process of compounding. Unlike morphological derivation (which affixes bound morphemes to a lexeme to create a new lexeme), compounding combines *free* morphemes to create a lexeme, as with the English word *candlestick*. We find it useful to process compounds after inflections because compounds as novel lexemes invite the same inflectional processes as non-compound lexemes of their core part of speech.

Following Wu and Yarowsky (2018), we consider compounds as words with two morphemes concatenated together, potentially with surface alterations. (McCarthy et al. (2019) used this to find compound color words in thousands of languages.) We split a word into all possible morpheme pairs, such that each morpheme has a length of at least 3 and at least one component has an edit distance at most 2 from some dictionary entry.[11] Thus, this method also identifies the decomposition of the compound word. When only one component fits the edit distance criterion, the decomposition omits the component with high edit distance. To make detection of compounds tractable, our implementation relies on fast prefix and suffix tries. A related alternative is the finite-state representation by Oflazer (1996).

We apply our compound analysis method to identify compounds in the residual words, and we manually evaluate a random sample of 50 predicted compounds Table A.4. Of these, 30 were correctly identified as compounds, and 22 were correctly decomposed. We observed a high number of false positives, which can be easily filtered out by examining the total edit distance of the components to known words. Every correctly identified compound has components whose combined edit distance is ≤ 2 (note that earlier we consider a compound to be valid if at least *one* component's edit distance to a known word is ≤ 2). Removing false positives with a total edit distance greater than 2 removes 18 incorrectly classified compounds, improving precision.

Many correctly identified compounds had a combined edit distance of zero or one (e.g., джазформация as джаз 'jazz' + формация 'formation'). Some errors were particularly instructive. For example, the word калейдоскопът 'the kaleidoscope', is incorrectly identified as a compound word whose second component is път 'road'. In fact, this word is a definite inflection of калейдоскоп 'kaleidoscope' using the suffix -ът. This reveals a transduction missing from our list in §5.2. In fact, we found the compound analysis to be quite helpful in identifying new inflectional suffixes, with which we augmented our FST for inflectional morphology.

[11] These values likely need to be adapted to new languages.

6 Discussion and Conclusion

We have investigated the space of unknown lexical items in naturally occurring text. In a case study on Bulgarian, a host of analytical models applied sequentially characterize the residual space of out-of-vocabulary words. Our models identify myriad processes responsible for these unknown words and map from such words to known words via heuristic and probabilistic processes. In this way, it complements Cucerzan and Yarowsky (2000) who model unknown words based on affixal or contextual similarity, and it affords means to improve machine translation.

The complete results of the residual space analyses are given in Table 1. Of the held-out set of 100 randomly sampled OOV words, our sequence of analyses properly taxonomized 69 of these. To confirm the robustness of these findings, a parallel study using the same series of techniques was conducted on the BulTreeBank corpus (Simov et al., 2002). In this case, 78% of a random sample of unknown words was correctly classified (see Table A.5), affirming the validity of the approach.

Initially one might suspect the need for less aggressive inflection and compounding models, given that so many errors were typos. On balance, significant fractions of the analyses were reasonable: even if an inflected form is misspelled, it is useful to reduce it to a lemma that can then reduce the space of possible correct spellings to which it can be mapped. While our annotation convention allows for only a single category per word, several examples show the benefit of using annotations as heuristics with shades of nuance worthy of human validation. For instance, several misspelled proper names are identified as names rather than typos, and a case of two words inadvertently joined by a deleted space (i.e., a typo) is correctly decomposed into those words by the compounding model.

In light of continued challenges in designing computational tools that effectively serve the world's thousands of languages, and that ignoring the linguistic traits of a language does not absolve the designer but rather induces greater harm (Bender, 2009), a detailed and taxonomized understanding of the behaviors of the language is vital. Our analysis of the word formation processes in such a way that can be grounded in the known lexicon affords both broad-scale familiarity with the language and practical value: it can tailor the design of core NLP tools to the residual vocabulary of a new language.

Acknowledgments

Arya D. McCarthy is supported by an Amazon AI2AI Fellowship and a Frederick Jelinek Fellowship.

References

Mark Aronoff and Nanna Fuhrhop. 2002. Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint. *Natural Language & Linguistic Theory*, 20:451–490.

Timothy Baldwin, Jonathan Pool, and Susan Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Coling 2010: Demonstrations*, pages 37–40, Beijing, China. Coling 2010 Organizing Committee.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 431–437, Beijing, China. Association for Computational Linguistics.

P.G. Corbett and P.B. Comrie. 2003. *The Slavonic Languages*. Routledge Language Family Series. Taylor & Francis.

Silviu Cucerzan and David Yarowsky. 2000. Language independent, minimally supervised induction of lexical probabilities. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 270–277, Hong Kong. Association for Computational Linguistics.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Adrià de Gispert. 2006. *Introducing linguistic knowledge into statistical machine translation*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Andrea Fischer, Klára Jágrová, Irina Stenger, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2016. Orthographic and morphological correspondences between related Slavic languages as a base for modeling of mutual intelligibility. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4202–4209, Portorož, Slovenia. European Language Resources Association (ELRA).

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Mikel L. Forcada and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

C.E. Gribble. 1987. *Reading Bulgarian Through Russian*. Slavica Publishers.

Jan Hajič and Eva Hajičová. 2007. Some of our best friends are statisticians. In *Text, Speech and Dialogue*, pages 2–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Jennifer Hay. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics*, 39:1041–1070.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Junko Ito and Armin Mester. 1995. The core-periphery structure of the lexicon and constraints on reranking. In Jill Beckman, Suzanne Urbanczyk, and Laura Walsh, editors, *Papers in Optimality Theory*, volume 18 of *University of Massachusetts Occasional Papers in Linguistics [UMOP]*, pages 181–209. University of Massachusetts, Amherst: GLSA.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Portorož, Slovenia. European Language Resources Association (ELRA).

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.

Hristo Krushkov. 2001. Automatic morphological processing of Bulgarian proper nouns.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24(4):377–439.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Dylan Lewis, Winston Wu, Arya D. McCarthy, and David Yarowsky. 2020. Neural transduction for multilingual lexical translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4373–4384, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stela Manova. 2010. Suffix combinations in Bulgarian: parsability and hierarchy-based ordering. *Morphology*, 20(1):267–296.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Winston Wu, Aaron Mueller, William Watson, and David Yarowsky. 2019. Modeling color terminology across thousands of languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2241–2250, Hong Kong, China. Association for Computational Linguistics.

M. McIlroy. 1982. Development of a spelling list. *IEEE Transactions on Communications*, 30(1):91–99.

Kyongho Min and William H. Wilson. 1998. Integrated control of chart items for error repair. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 862–868, Montreal, Quebec, Canada. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.

Kemal Oflazer. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89.

Mirena Patseva. 2017. Bulgarian word stress analysis in the frame of prosody morphology interface. Technical report, Rutgers University.

Alexander Popov, Petya Osenova, and Kiril Simov. 2020. Implementing an end-to-end treebank-informed pipeline for Bulgarian. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 162–167, Düsseldorf, Germany. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*, volume 4. OUP Oxford.

Timothy A Salthouse. 1984. Effects of age and skill in typing. *Journal of Experimental Psychology: General*, 113(3):345.

Timothy A Salthouse. 1986. Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological bulletin*, 99(3):303.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 13–25. Springer Netherlands, Dordrecht.

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, Berlin, Germany. Association for Computational Linguistics.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Milena Slavcheva. 2003. Some aspects of the morphological processing of Bulgarian. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 71–77, Budapest, Hungary. Association for Computational Linguistics.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015a. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *Systems and Frameworks for Computational Morphology*, pages 72–93, Cham. Springer International Publishing.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015b. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

C.E. Townsend and L.A. Janda. 1996. *Common and Comparative Slavic: Phonology and Inflection : with Special Attention to Russian, Polish, Czech, Serbo-Croatian, Bulgarian*. Slavica Publishers.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 598–608, Denver, Colorado. Association for Computational Linguistics.

Yulia Tsvetkov and Chris Dyer. 2015. Lexicon stratification for translating out-of-vocabulary words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 125–131, Beijing, China. Association for Computational Linguistics.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33–39, Prague, Czech Republic. Association for Computational Linguistics.

Robert A. Wagner and Roy Lowrance. 1975. An extension of the string-to-string correction problem. *J. ACM*, 22(2):177–183.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Winston Wu and David Yarowsky. 2018. Massively translingual compound analysis and translation discovery. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Winston Wu and David Yarowsky. 2020a. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Winston Wu and David Yarowsky. 2020b. Wiktionary normalization of translations and morphological information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4683–4692, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

A Supplemental Material

In the following pages, we provide specific analyses, both hand-crafted and computationally performed, of the residual vocabulary. All tables are referred to in the main text.

Index	Word	Translation	Type	Sub-Type	Features	POS
1	петминутни	five minute	Compound	Declension	N/A	ADJ
2	кръглоскулови	round-cheeked	Compound	Declension	N/A	ADJ
3	Неврофиброматоза	Neurofibromatosis	Compound	N/A	N/A	NOUN
4	киноадаптация	adapted for movie	Compound	N/A	FEM	NOUN
5	постулиращ	postulating	Conjugation	N/A	N/A	PART
6	вселяват	inspire	Conjugation	N/A	PL	VERB
7	осквернява	desecrate	Conjugation	N/A	N/A	VERB
8	интерпретираща	interpretating	Conjugation	N/A	N/A	PART
9	ϵκάστου	each	Foreign	Ancient Greek	N/A	PRON
10	Fiyafi	savannah	Foreign	Bashkir	N/A	NOUN
11	18bit	18-bit	Foreign	English	N/A	ADJ
12	World";<br	World	Foreign	English	N/A	NOUN
13	goto	goto	Foreign	English	N/A	VERB
14	nü	nü	Foreign	English	N/A	NOUN
15	Darvin	Darvin	Foreign	English	N/A	PROPN
16	ordinatorium	ordinatorium	Foreign	Latin	N/A	NOUN
17	Branchinella	Branchinella	Foreign	Latin	N/A	NOUN
18	vagrans	vagrans	Foreign	Latin	N/A	NOUN
19	Genealogia	genealogy	Foreign	Portuguese	N/A	NOUN
20	Obrero	Workers'*	Foreign	Spanish	N/A	PROPADJ
21	Bilmeden	without knowing	Foreign	Turkish	N/A	PART
22	Предарица	Predaritsa	Geography	Bulgarian	N/A	PROPN
23	Устренският	Ustrenskiat	Geography	Bulgarian	DEF	PROPADJ
24	Чиевци	Chievtsi	Geography	Bulgarian	N/A	PROPN
25	еркечкия	erkechki	Geography	Bulgarian	DEF	PROPADJ
26	харманлии	Harmanlii	Geography	Bulgarian (Capital)	N/A	PROPN
27	Хетфилд	Hatfield	Geography	English	N/A	PROPN
28	Чансълърсвилската	Chancellorsville	Geography	English	DEF	PROPADJ
29	Норвич	Norwich	Geography	English	N/A	PROPN
30	келия	Kelia	Geography	Greek	N/A	PROPN
31	Карино	Karino	Geography	Italian	N/A	PROPN
32	Креспано	Crespano	Geography	Italian	N/A	PROPN
33	Триесткият	Triest	Geography	Italian	DEF	PROPADJ
34	Колàца	Colazza	Geography	Italian	N/A	PROPN
35	Майкубенския	Maikuben	Geography	Kazakh	DEF	PROPADJ
36	Хайдаркан	Khaidarkan	Geography	Kyrgyz	N/A	PROPN
37	Донданген	Dondagen	Geography	Latvian	N/A	PROPN
38	джемат	Djemat	Geography	Serbian (Capital)	N/A	PROPN
39	Пунтаренас	Puntarenas	Geography	Spanish	N/A	PROPN
40	презаснемането	retake	Declension	N/A	DEF	NOUN
41	сънародника	compatriot	Declension	N/A	DEF	NOUN
42	хидропланът	the hydroplane	Declension	N/A	DEF	NOUN
43	закалени	hardened	Declension	N/A	PL	ADJ
44	амидразоните	the amidrazones	Declension	N/A	PL+DEF	NOUN
45	умственото	the mental	Declension	N/A	DEF	ADJ
46	прибоят	the surf	Declension	N/A	DEF	NOUN
47	панкреатичната	the pancreatic	Declension	N/A	DEF	ADJ
48	представителят	the representative	Declension	N/A	DEF	ADJ
49	интригуваща	intriguing	Declension	N/A	FEM	ADJ
50	алтъни	gold coins	Declension	N/A	PL	NOUN
51	абонаментни	subscription	Declension	N/A	PL	ADJ
52	ракетка	small rocket	Declension	N/A	DIM	NOUN
53	Тежките	the heavy	Declension	N/A	PL+DEF	ADJ
54	квесторското	the quaestoring	Declension	N/A	DEF	NOUN
55	кармелитка	Carmelite	Declension	N/A	FEM	PROPN
56	плутония	plutonium	Declension	N/A	DEF	PROPN
57	Баритонът	the baritone	Declension	N/A	DEF	NOUN
58	трактове	tracts	Declension	N/A	PL+DEF	NOUN
59	тюрколожка	turkologist	Declension	N/A	FEM	NOUN
60	лигандното	the ligand	Declension	N/A	NEUT+DEF	ADJ
61	Мухльовци	ninnies	Declension	N/A	PL	ADJ
62	Амфибио	Amphibio	Name	Car	N/A	PROPN
63	Керуал	Keroualle	Name	French	N/A	PROPN
64	Ерхфрид	Erchanfried	Name	German	N/A	PROPN
65	Арагами	Aragami	Name	Japanese	N/A	PROPN
66	Иродиади	Herodias	Name	Latin	N/A	PROPN
67	Евто	Evto	Name	Person	N/A	PROPN
68	Посълуайт	Postlethwaite	Name	Person	N/A	PROPN
69	Васcалли	Vassalli	Name	Person	N/A	PROPN
70	чихо	Chiho	Name	Person	N/A	PROPN
71	Рисова	Risova	Name	Person	N/A	PROPN
72	Фридлендър	Friedlander	Name	Person	N/A	PROPN
73	Тремитус	Tremitus	Name	Person	N/A	PROPN
74	Върчаковски	Varchakovski	Name	Person	N/A	PROPN
75	Камбанийски	Kambaniiski	Name	Person	N/A	PROPN
76	Адольф	Adolf	Name	Russian	N/A	PROPN
77	Киспии	Kaspii	Name	Tribe	N/A	PROPN
78	611.8	611.8	Number	N/A	N/A	NUM
79	Ми34	Mi-34	Product Name	N/A	N/A	PROPN
80	Турбина	turbine	Target	N/A	FEM	NOUN
81	Настигайки	catching up	Target	N/A	N/A	PART
82	Покосен	Stricken	Target	N/A	N/A	PART
83	Студентство	College experience	Target	N/A	N/A	NOUN
84	месинг	brass	Target	N/A	N/A	NOUN
85	Питомен	domesticated	Target	N/A	N/A	ADJ
86	Салкъм	Offshoot	Target	N/A	N/A	NOUN
87	мерило	measure	Target	N/A	N/A	NOUN
88	черничев	mulberry	Target	N/A	MASC	ADJ
89	асимптотически	asymptotic	Target	N/A	MASC	ADJ
90	нюзпейпър	newspaper	Transliteration	English	N/A	NOUN
91	шейдъри	shaders	Transliteration	English	PL	NOUN
92	стигнатдо	reached up (to)	Typo	Concatenation	N/A	PART
93	1948.През	1948.Through	Typo	Concatenation	N/A	PRON
94	про\xadвеж\xadда\xadне	N/A	Typo	Formatting	N/A	PRON
95	Алфсосо	Alfonso	Typo	Mixed	N/A	PROPN
96	популярзират	polularize	Typo	N/A	PL	VERB
97	прожа	will continue	Typo	Omission	N/A	VERB
98	низинà	valley	Typo	Punctuation	N/A	NOUN
99	блодове	fruits	Typo	Substitution	N/A	NOUN
100	Ricochet.com	Ricochet	Web Address	English	N/A	PROPN

Table A.1: Manual classification of 100 randomly sampled words from the tokenized Bulgarian Wikipedia corpus **before** any further processing from our pipeline is performed. We use the UD part-of-speech tags from: `https://universaldependencies.org/u/pos/`. Table is summarized in Figure 1.

Index	Word	Translation	Type	Sub-Type	Features	POS
1	овесопроизводител	oat producer	Compound	N/A	MASC	NOUN
2	непорнографски	non-pornographic	Compound	N/A	MASC	ADJ
3	бързоразрастващи	fast growing	Compound	N/A	PL	ADJ
4	окомерни	eye sketching	Compound	N/A	PL	ADJ
5	преформирването	the reformatting	Compound	N/A	NEU+DEF	NOUN
6	метапознавателните	the meta cognitive	Compound	N/A	PL+DEF	ADJ
7	трискатна	of three gables	Compound	N/A	FEM	ADJ
8	бундестреньор	coach of German national team	Compound	N/A	MASC	NOUN
9	многоцветниците	nymphalidae	Compound	N/A	PL+DEF	NOUN
10	авиокосмическо	(pertaining to) aerospace	Compound	N/A	NEUT	ADJ
11	геолокацията	geolocation	Compound	N/A	FEM+DEF	NOUN
12	следамерикански	post American	Compound	N/A	MASC	ADJ
13	китоловния	whaling	Compound	N/A	MASC+DEF	ADJ
14	видеоизкуство	videoart	Compound	N/A	NEUT	NOUN
15	новоизграденият	the newly built	Compound	N/A	MASC+DEF	ADJ
16	първооснови	primary basis	Compound	N/A	PL	NOUN
17	лейбгвардейците	the life guards	Compound	N/A	PL+DEF	NOUN
18	сухолубиви	drought-tolerant	Compound	N/A	PL	ADJ
19	наскоропоявилия	appeared recently	Compound	N/A	MASC+DEF	ADJ
20	леководолазът	the scuba diver	Compound	N/A	MASC+DEF	NOUN
21	косиха	mowed	Conjugation	Bulgarian	N/A	VERB
22	разбъркаха	mixed	Conjugation	N/A	N/A	VERB
23	обследвайки	investigating, inquiring	Conjugation	N/A	N/A	PART
24	заобиколил	go around, circumvent	Conjugation	N/A	N/A	PART
25	преместилите	moved around	Conjugation	N/A	PL+DEF	PART
26	нарекъл	called, named	Conjugation	N/A	MASC	PART
27	недостигнатия	unattainable	Conjugation	N/A	MASC+DEF	PART
28	тероризиращ	terrorizing	Conjugation	N/A	MASC	PART
29	досмила	digesting, grinding	Conjugation	N/A	N/A	PART
30	руководителей	leaders	Foreign	Russian	PL	NOUN
31	паутина	spider web	Foreign	Russian	FEM	NOUN
32	питянтятяра	Pitjantjatjara	Geography	Australian	N/A	PROPN
33	широкольшки	(of) Shiroka Laka	Geography	Bulgarian	PL	ADJ
34	сетница	Setnica	Geography	Bulgarian	N/A	PROPN
35	замфировска	zamphyrovska	Geography	Bulgarian	FEM	ADJ (Proper)
36	гулянци	Guliantsi	Geography	Bulgarian	N/A	PROPN
37	бенковската	benkovska	Geography	Bulgarian	FEM+DEF	ADJ (Proper)
38	знеполското	the znepolsko	Geography	Bulgarian	NEUT+DEF	ADJ (Proper)
39	алитуска	the alitusk	Geography	Bulgarian	FEM+DEF	ADJ (Proper)
40	отроковице	Otrokovice	Geography	Czech	N/A	PROPN
41	блекпулски	Blackpool	Geography	English	MASC	ADJ (Proper)
42	пфалцски	(pertaining to) Pfalz	Geography	German	MASC	ADJ (Proper)
43	сицилианска	Sicilian	Geography	Italian	FEM	ADJ
44	няманоро	Nyamanoro	Geography	Japan	N/A	PROPN
45	твърдишкото	(pertaining to) Tvarditsa	Geography	Place	NEUT+DEF	ADJ
46	можайският	Mojayska	Geography	Russian	MASC+DEF	ADJ (Proper)
47	саянската	of the Sayan (Mountains)	Geography	Russian	FEM+DEF	ADJ
48	лядунския	(of) Liaodong	Geography	Russian	MASC+DEF	ADJ
49	верхневилюйское	Verkhnevilyuysk	Geography	Russian	N/A	ADJ (Proper)
50	болградското	Bolgradski	Geography	Ukranian	NEUT	ADJ (Proper)
51	азраки	Azraqi	Georgraphy	Persian	N/A	PROPN
52	купчето	the small pile/bunch	Declension	N/A	NEUT+DIM+DEF	NOUN
53	естуарното	the estuarine	Declension	N/A	NEUT+DEF	ADJ
54	сметачната	the calculating	Declension	N/A	FEM+DEF	ADJ
55	просешката	the beggary	Declension	N/A	FEM+DEF	ADJ
56	трахити	trachytes	Declension	N/A	PL	NOUN
57	миллети	millets	Declension	N/A	PL	NOUN
58	ротердамци	inhabitants of Rotherdam	Declension	N/A	PL	NOUN
59	ресинтезът	the resynthesis	Declension	N/A	MASC+DEF	NOUN
60	флуксетинът	the Fluoxetine	Declension	N/A	MASC+DEF	NOUN
61	неосъзнаваното	the unconsciously	Declension	N/A	NEUT+DEF	ADJ
62	изсечена	set in stone, cut down	Declension	N/A	FEM	ADJ
63	сайгите	the saiga anthlopes	Declension	N/A	PL+DEF	NOUN
64	смърчови	(of) spruce	Declension	N/A	PL	ADJ
65	конецовидните	thread-like	Declension	N/A	PL+DEF	ADJ
66	предлози	prepositions	Declension	N/A	PL	NOUN
67	селяка	peasant	Declension	N/A	MASC+DEF	NOUN
68	ленни	land granted by Ottomans	Declension	N/A	PL	ADJ
69	екстрахепатичните	extrahepatic	Declension	N/A	PL+DEF	ADJ
70	пуническото	pertaining to Punics	Declension	N/A	NEUT+DEF	ADJ
71	владишкото	of the bishop	Declension	N/A	NEUT+DEF	ADJ
72	кратовския	(pertaining to) Kratovo	Declension	N/A	MASC+DEF	ADJ (Proper)
73	дудукът	duduk	Declension	N/A	MASC+DEF	NOUN
74	пастафории	pastophoria	Declension	Transliteration	PL	NOUN
75	създателю	creator	Declension	Vocative	MASC	NOUN
76	костурчанка	kosturchanka	Name	Inhabitants	FEM	PROPN
77	дъмбълдоров	(pertaining to) Dumbledore	Name	Person	MASC	ADJ (proper)
78	северо	Severo	Name	Person	N/A	PROPN
79	квинтерна	gittern	Target	N/A	FEM	NOUN
80	назалност	nasality	Target	N/A	FEM	NOUN
81	нелютив	mild (not hot)	Target	N/A	MASC	ADJ
82	биткойн	bitcoin	Target	Transliteration	MASC	NOUN
83	скакач	springbok	Target	Zoology	MASC	NOUN
84	гюйсът	the jack	Transliteration	Dutch	MASC+DEF	NOUN
85	дайнамикс	dynamics	Transliteration	English	N/A	PROPN
86	обложака	cover	Typo	Addition	FEM	NOUN
87	меджународната	the international	Typo	Character Swap	FEM+DEF	ADJ
88	палеографикатегория	paleography category	Typo	Concatenation	FEM	NOUN
89	юдеизма	the judaism	Typo	Declension	MASC+DEF	NOUN
90	наташната	the next	Typo	Omission	FEM+DEF	ADJ
91	широкоразпространеня	the widespread	Typo	Omission	MASC+DEF	ADJ
92	низхождение	descent	Typo	Omission	NEUT	NOUN
93	цитросуви	(of) citrus	Typo	Substitution	PL	ADJ
94	оковчателното	(of) the final result	Typo	Substitution	NEUT+DEF	ADJ
95	жинотните	the animals	Typo	Substitution	PL+DEF	NOUN
96	татраедър	tetrahedron	Typo	Substitution	MASC	NOUN
97	имплементзция	implementation	Typo	Substitution	FEM	NOUN
98	принадлежът	belong	Typo	Substitution	PL	VERB
99	оръсия	weapons	Typo	Substitution	PL	NOUN
100	предозонова	prednisone (therapy)	Typo	Transliteration	FEM	ADJ

Table A.2: Manual classification of 100 randomly sampled words from the tokenized Bulgarian Wikipedia Corpus **after** eliminating entries from the union of dictionaries. We use the UD part-of-speech tags from: `https://universaldependencies.org/u/pos/`. Table is summarized in Figure 3.

Index	Word	Valid
1	витал	Yes
2	втрху	Yes
3	ихрам	Yes
4	синут	Yes
5	съкър	Yes
6	витрал	Yes
7	маквис	Yes
8	пераун	Yes
9	почест	Yes
10	ревабш	Yes
11	рененг	Yes
12	живееяг	Yes
13	модулин	Yes
14	гутболни	Yes
15	джутсуту	Yes
16	камбоурн	Yes
17	убеждавт	Yes
18	читирима	Yes
19	антатната	Yes
20	вододпади	Yes
21	наблядава	Yes
22	присъствт	Yes
23	художникт	Yes
24	обстрикция	Yes
25	преостъпва	Yes
26	ассортимент	Yes
27	монодрамата	Yes
28	присътствал	Yes
29	революциятс	Yes
30	числиността	Yes
31	продолжавало	Yes
32	нараставащата	Yes
33	пристрелването	Yes
34	стандфордското	Yes
35	модерницираните	Yes
36	туид	No
37	течащ	No
38	тодас	No
39	шейдър	No
40	спомнящ	No
41	връчващо	No
42	кодеинът	No
43	напомняш	No
44	невиждащ	No
45	струващо	No
46	китобойци	No
47	влайковите	No
48	кварковото	No
49	радиошоуто	No
50	семинолско	No

Table A.3: Human validation of random sample of misspelling classifications.

Index	Word	Decomposition	Edit Distance	Valid Compound	Valid Decomposition
1	калейдоскопът	калейдоскоп|път	1	No	No
2	дроидчето	@|ридчето	2	No	No
3	вазодилатиращ	вазови|датиращ	2	Yes	No
4	паналбанската	пан|албанската	0	Yes	Yes
5	трудноподвижност	трудно|подвижност	0	Yes	Yes
6	узункьопрюйския	@|райския	9	No	No
7	крайгълните	крайгълен|ите	1	No	No
8	епископалианците	епископа|ливанците	1	No	No
9	фотостареенето	фото|стареенето	0	Yes	Yes
10	тескерета	тес|@/@|ета	6	No	No
11	видеообмен	видео|обмен	0	Yes	Yes
12	дефтерхането	дефтера|нето	1	Yes	Yes
13	класфицира	@|скицира	4	No	No
14	несатнтименталното	@|менталното	8	Yes	No
15	предпубертетна	пред|пубертетна	0	Yes	Yes
16	екплозивни	@|позивни	3	No	No
17	сложноустроени	сложно|устроени	0	Yes	Yes
18	миникомикси	мини|комикси	0	Yes	Yes
19	бромалгин	бром|олгин	1	Yes	Yes
20	хиподермата	хипо|дермата	0	Yes	Yes
21	зогисткия	зог|есткия	1	No	Yes
22	колаборанти	кол|лаборанти/кола|оранти	1	Yes	No
23	древноеврейските	древно|еврейските	0	Yes	Yes
24	щалупьоненщалупьонен	щало|@	16	No	No
25	нарамвали	@|вали/нара|@	5	No	No
26	друмевите	@|ите	6	No	No
27	екстрабукалната	@|калната	8	Yes	No
28	дзайбацу	дза|@	5	Yes	No
29	анасонлийките	анасон|@	7	No	No
30	петокласно	пето|класно	0	Yes	Yes
31	джазформация	джаз|формация	0	Yes	Yes
32	крайдунавски	край|дунавски	0	Yes	Yes
33	елабуцки	ела|@	5	No	No
34	ориксът	ори|кът	1	No	No
35	римокатолическа	римо|католическа	0	Yes	Yes
36	арондисмана	@|имана	6	No	No
37	истанбулчаникатегория	истанбулчани|категория	0	Yes	Yes
38	сподобиха	спо|добиха	0	Yes	Yes
39	прокомуникирана	прокоп|@	10	Yes	No
40	леополдините	леополд|дините	1	No	No
41	детройтът	детройт|@	2	No	Yes
42	шитл'ивица	@|вица	5	No	No
43	премъдростната	@|яростната	6	Yes	No
44	шестмоторни	шест|моторни	0	Yes	Yes
45	филмографията	фил|зографията/филм|зографията	1	Yes	Yes
46	средногъстата	средно|гъстата	0	Yes	Yes
47	безкуполен	без|куполен	0	Yes	Yes
48	гоцезелчевската	гоце|енчевската	2	Yes	Yes
49	епскоп	@|коп	3	No	No
50	лопатовиднозъб	лопатови|@	6	Yes	No

Table A.4: Human validation of random sample of compound analysis.

Index	Word	Human Trans.	Alg. Trans.	Human Type	Human Sub-Type	Alg. Type	Alg. Sub-Type	Features	POS
1	н-к	manage (abbreviated)	N/A	Abbreviation	N/A	Foreign	Russian	MASC	NOUN
2	полупансион	half board	semi board	Compound	N/A	Compound	N/A	MASC	NOUN
3	свръхелегантен	overly well dressed	svra|elegant	Compound	N/A	Compound	N/A	MASC	NOUN
4	по-нагъл	more impudent	N/A	Compound	N/A	Foreign	Russian	MASC	ADJ
5	търговско-промишлена	Industrial-and-retail	N/A	Compound	N/A	Foreign	Russian	FEM	ADJ
6	русокоси	blond	russian	Compound	N/A	Declension	Fuzzy	PL	ADJ
7	новооткритото	the newly found	openings	Compound	N/A	Declension	Fuzzy	NEUT+DEF	ADJ
8	мироопазващите	peace-keeping	peace|keeping	Compound	N/A	Declension	Simple	PL+DEF	ADJ
9	подплашил	slightly scared	tear off|beaten	Conjugation	N/A	Compound	N/A	MASC	PART
10	инспирирано	inspired	@|early	Conjugation	N/A	Compound	N/A	NEUT	PART
11	жужи	buzz	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
12	могли	could	N/A	Conjugation	N/A	Conjugation	N/A	N/A	PART
13	забиха	poke	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
14	карало	driven	N/A	Conjugation	N/A	Conjugation	N/A	NEUT	ADJ
15	плъзна	slide	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
16	изгонва	expels	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
17	изкъпем	take a bath	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
18	копнели	longing	N/A	Conjugation	N/A	Conjugation	N/A	PL	PART
19	работиш	work	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
20	сдобили	obtained	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
21	оставаме	remaining	remaining	Conjugation	N/A	Conjugation	N/A	N/A	VERB
22	отзовали	responded	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
23	обвиняват	accuse	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
24	познаваха	recognized	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
25	промъквал	sneaked	N/A	Conjugation	N/A	Conjugation	N/A	N/A	PART
26	дипломираш	graduate	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
27	започвайте	begin	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
28	проведохме	carried out	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
29	разчитайте	rely	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
30	поздравяват	greet	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
31	представяше	represented	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
32	претоварваш	overload	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
33	разстройваме	disturb	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
34	съсредоточиш	concentrate (mentally)	N/A	Conjugation	N/A	Conjugation	N/A	N/A	VERB
35	тръшна	fall abbruptly	trish	Conjugation	N/A	Declension	Fuzzy	N/A	VERB
36	наметна	drape over	document	Conjugation	N/A	Declension	Fuzzy	N/A	VERB
37	назначиха	appointed	N/A	Conjugation	N/A	Declension	Simple	N/A	VERB
38	озъртат	look around	N/A	Conjugation	N/A	N/A	N/A	N/A	VERB
39	поведох	lead	N/A	Conjugation	N/A	Proper	Likely	N/A	VERB
40	Дувър	Dover	N/A	Geography	English	Proper	Likely	N/A	PROPN
41	Козро	Kozro	N/A	Geography	Russian	Proper	Likely	N/A	PROPN
42	Ридсдейл	Reedsdale	N/A	Geography	English	Proper	Likely	N/A	PROPN
43	Апенините	Appennini	N/A	Geography	Italian	Proper	Likely	PL	PROPN
44	далавери	deals	given|friends	Declension	N/A	Compound	N/A	FEM+PL	NOUN
45	манталитетът	the mentality	mentality|@	Declension	N/A	Compound	N/A	MASC+DEF	NOUN
46	клъвки	beak	click	Declension	N/A	Declension	Fuzzy	FEM+PL	NOUN
47	божиите	godly	godly	Declension	N/A	Declension	Fuzzy	PL+DEF	ADJ
48	болките	the pains	pains	Declension	N/A	Declension	Simple	FEM+PL+DEF	NOUN
49	гърмежи	thunder	report	Declension	N/A	Declension	Simple	PL	NOUN
50	епохата	the epoch	N/A	Declension	N/A	Declension	Simple	FEM+DEF	NOUN
51	великите	The great	veliki	Declension	N/A	Declension	Simple	PL+DEF	ADJ
52	депутата	the congressman	congressman	Declension	N/A	Declension	Simple	MASC+DEF	NOUN
53	детската	the childish	toy	Declension	N/A	Declension	Simple	FEM+DEF	ADJ
54	повелите	the commands	entrusted	Declension	N/A	Declension	Fuzzy	FEM+PL+DEF	NOUN
55	клетвения	sworn	sworn	Declension	N/A	Declension	Simple	MASC+DEF	ADJ
56	пазарлъци	bargains	bargain	Declension	N/A	Declension	Fuzzy	MASC PL	NOUN
57	погребите	cellar, arms depot	entomb, bury	Declension	N/A	Declension	Fuzzy	MASC+PL+DEF	NOUN
58	случилото	occurred	occurred	Declension	N/A	Declension	Simple	NEUT+DEF	PART
59	чехкините	The Czech (females)	Check (female)	Declension	N/A	Declension	Simple	FEM+DEF	ADJ (Proper)
60	заловеният	captured	captured	Declension	N/A	Declension	Simple	MASC+DEF	ADJ
61	известните	famous	famous	Declension	N/A	Declension	Simple	PL+DEF	ADJ
62	момченцето	the little boy (demunitive)	little boy	Declension	N/A	Declension	Simple	NEUT+DEF	NOUN
63	отдалечила	distanced	N/A	Declension	N/A	Declension	Simple	FEM	PART
64	премиерите	the prime ministers	premiers	Declension	N/A	Declension	Simple	MASC+DEF	NOUN
65	софийската	Sofia	Sofia	Declension	N/A	Declension	Simple	FEM+DEF	ADJ
66	тексасците	the texans	texan	Declension	N/A	Declension	Fuzzy	PL+DEF	NOUN
67	еврофондове	european funds	eurofor	Declension	N/A	Declension	Fuzzy	MASC+PL	NOUN
68	изпратените	sent	sent	Declension	N/A	Declension	Simple	PL+DEF	PART
69	позиционните	positioning	position	Declension	N/A	Declension	Simple	PL+DEF	ADJ
70	съвестността	the conscience	conscience	Declension	N/A	Declension	Fuzzy	FEM+DEF	NOUN
71	холивудските	the hollywood	hollywood	Declension	N/A	Declension	N/A	PL+DEF	ADJ
72	необмислените	the thoughtless	thoughtless	Declension	N/A	Declension	Fuzzy	PL+DEF	ADJ
73	вестникарските	the newspaper	newspaper	Declension	N/A	Declension	Simple	PL+DEF	ADJ
74	изразходваните	consumed	consumed	Declension	N/A	Declension	Simple	PL+DEF	ADJ
75	социалдемократически	social democratic	socialdemocrat	Declension	N/A	Declension	Simple	PL	ADJ
76	мъжът	the man	N/A	Declension	N/A	N/A	N/A	MASC+DEF	NOUN
77	студът	the cold	N/A	Declension	N/A	N/A	N/A	MASC+DEF	NOUN
78	провинилите	the guilty	N/A	Declension	N/A	N/A	N/A	PL+DEF	ADJ
79	БВ	BV	N/A	N/A	N/A	Proper	Likely	N/A	N/A
80	Жега	Heat	N/A	Name	Movie	Proper	Likely	FEM	NOUN
81	Клио	Clio	N/A	Name	Car	Proper	Likely	N/A	PROPN
82	ПАНОВ	Panov	N/A	Name	Person	Proper	Likely	MASC	PROPN
83	Симон	Simon	N/A	Name	Person	Proper	Likely	N/A	PROPN
84	Чейни	Cheney	N/A	Name	Person	Proper	Likely	N/A	PROPN
85	Ганева	Ganeva	N/A	Name	Person	Proper	Likely	N/A	PROPN
86	Емилия	Emilia	N/A	Name	Person	Proper	Likely	FEM	PROPN
87	Трифон	Trifon	N/A	Name	Person	Proper	Likely	MASC	PROPN
88	Централ	Central	N/A	Name	Hotel	Proper	Likely	N/A	PROPN
89	литерер	Litteraire	N/A	Name	Newspaper	Proper	Funky	N/A	PROPN
90	Елизабет	Elizabeth	N/A	Name	Person	Proper	Likely	N/A	PROPN
91	Компанис	Companys	N/A	Name	Person	Proper	Likely	N/A	PROPN
92	Лизаразу	Lizarazu	N/A	Name	Person	Proper	Likely	N/A	PROPN
93	Талейран	Talleyrand	N/A	Name	Person	Proper	Likely	N/A	PROPN
94	Анастасия	Anastasia	N/A	Name	Person	Proper	Likely	FEM	PROPN
95	налудничаво	crazy	@|chavo	Target	N/A	Compound	N/A	NEUT	ADJ
96	олелия	commotion	yikes	Target	N/A	Declension	Fuzzy	FEM	NOUN
97	разведряване	détente	clearing	Target	N/A	Declension	Fuzzy	NEUT	NOUN
98	Земя	Earth	N/A	Target	N/A	N/A	N/A	FEM	NOUN
99	навръх	at the peak of	N/A	Target	N/A	N/A	N/A	N/A	ADV
100	Даунтаун	downtown	N/A	Transliteration	English	Proper	Likely	MASC	NOUN

Table A.5: Manual classification of 100 randomly sampled words after classifying all of the BulTreeBank corpus, in analogy with Table 1.