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.

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.
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 provides 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 lexis² 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-

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

²We 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 English 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, mis-spellings, compound words, and proper names like place names or people. We devise computational approaches to tackle these five major categories.

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 (Honlibal 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

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. 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 пър-фа́йлз ‘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-’.

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 (16%), Kazakh (6%), Chinese (6%), Albanian (4%), and single exemplars of Irish, Portuguese, and Moldovan.

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 Hajičková (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 ‘ъ’.
2. If the soft sign ‘ъ’ occurs in the middle of a
| Index | Word | Human Trans. | Alg. Trans. | Human Type | Sub-Type | Human Type | Alg. Type | Alg. Sub-Type | Features | POS |
|-------|------|--------------|-------------|------------|----------|------------|-----------|---------------|----------|-----|
| 1     | локомотив | locomotive | locomotive | Compound | – | – | – | – | – | FEM | NOUN |
| 2     | независим | independent | independent | Compound | – | – | – | – | – | ADJ | |
| 3     | операция | operation | operation | Compound | – | – | – | – | – | FEM | NOUN |
| 4     | битумно | bitumen | bitumen | Compound | – | – | – | – | – | PL | NOUN |
| 5     | светлочервено | light orange | light orange | Compound | – | – | – | – | – | MAS | ADJ |
| 6     | конфигурация | configuration | configuration | Compound | – | – | – | – | – | NEUT | NOUN |
| 7     | управление | management | management | Compound | – | – | – | – | – | PL | PART |
| 8     | астрофизика | astrophysics | astrophysics | Compound | – | – | – | – | – | – | NOUN |
| 9     | далекомер | distance meter | distance meter | Compound | – | – | – | – | – | Foreign | Russian | Russian | NOUN |
| 10    | машинка | machine | machine | Compound | – | – | – | – | – | – | NOUN |
| 11    | фенольная | phenol | phenol | Compound | – | – | – | – | – | PL | NOUN |
| 12    | география | geography | geography | Compound | – | – | – | – | – | FEM+DEF | ADJ (Proper) |
| 13    | неоспоримо | unassailable | unassailable | Compound | – | – | – | – | – | – | NOUN |
| 14    | звероферма | zoo farm | zoo farm | Compound | – | – | – | – | – | PL | PART |
| 15    | анатомия | anatomy | anatomy | Compound | – | – | – | – | – | – | NOUN |
| 16    | памятник | monument | monument | Compound | – | – | – | – | – | – | NOUN |
| 17    | действующий | acting | acting | Compound | – | – | – | – | – | PL | PART |
| 18    | дестващ | acting | acting | Compound | – | – | – | – | – | PL | PART |
| 19    | домантите | The Odomanti | The Odomanti | Compound | – | – | – | – | – | PL+DEF | ADJ (Proper) |
| 20    | същността | essence | essence | Compound | – | – | – | – | – | NEUT+DEF | ADJ |
| 21    | реакции | reactions | reactions | Compound | – | – | – | – | – | PL | NOUN |
| 22    | клавесиниста категория | akin to denudational day|odd person | Compound | – | – | – | – | – | – | NOUN |
| 23    | форматор | formatter | formatter | Compound | – | – | – | – | – | – | NOUN |
| 24    | футуризм | futurism | futurism | Compound | – | – | – | – | – | – | NOUN |
| 25    | фокус | focus | focus | Compound | – | – | – | – | – | PL | NOUN |
| 26    | дестинация | destination | destination | Compound | – | – | – | – | – | PL | NOUN |
| 27    | откриватели | discoverers | discoverers | Compound | – | – | – | – | – | PL | PART |
| 28    | откриватели | discoverers | discoverers | Compound | – | – | – | – | – | PL | PART |
| 29    | Памтивек (коло̀вите за античност) | Pamtivek (colloquial for ancient) | Pamtivek (colloquial for ancient) | Compound | – | – | – | – | – | – | NOUN |
| 30    | Рieseberg | Rieseberg | Rieseberg | Compound | – | – | – | – | – | – | NOUN |
| 31    | Olivier | Olivier | Olivier | Compound | – | – | – | – | – | – | NOUN |
| 32    | Евтахий | Evtahiy | Evtahiy | Compound | – | – | – | – | – | – | NOUN |
| 33    | Вигберт | Witbert | Witbert | Compound | – | – | – | – | – | – | NOUN |
| 34    | Бейтлър | Beightler | Beightler | Compound | – | – | – | – | – | – | NOUN |
| 35    | Халиду | Halidu | Halidu | Compound | – | – | – | – | – | – | NOUN |
| 36    | Пritsak | Pritsak | Pritsak | Compound | – | – | – | – | – | – | NOUN |
| 37    | Дарбес | Darbez | Darbez | Compound | – | – | – | – | – | – | NOUN |
| 38    | Пийбо | Pipto | Pipto | Compound | – | – | – | – | – | – | NOUN |
| 39    | Мусан | Musan | Musan | Compound | – | – | – | – | – | – | NOUN |
| 40    | Lopov | Lopov | Lopov | Compound | – | – | – | – | – | – | NOUN |
| 41    | Kakai | Kakai | Kakai | Compound | – | – | – | – | – | – | NOUN |
| 42    | ЗЕЛПО | ZELPO | ZELPO | Compound | – | – | – | – | – | – | NOUN |
| 43    | Юджи | Azel | Azel | Compound | – | – | – | – | – | – | NOUN |
| 44    | ЦТА | Central Tibet Administration | Central Tibet Administration | Compound | – | – | – | – | – | – | NOUN |
| 45    | Ancestors of Serbians | Ancestors of Serbians | Ancestors of Serbians | Compound | – | – | – | – | – | – | NOUN |
| 46    | Amorite | amorite | amorite | Compound | – | – | – | – | – | – | NOUN |
| 47    | Kodaikanal | kodaikanal | kodaikanal | Compound | – | – | – | – | – | – | NOUN |
| 48    | Шотландска мултиетническа кибернетизация | Scottish multiethnic cybernetics | Scottish multiethnic cybernetics | Compound | – | – | – | – | – | – | NOUN |
| 49    | тубуларен | tubular | tubular | Compound | – | – | – | – | – | – | NOUN |
| 50    | стагнет | stagnant | stagnant | Compound | – | – | – | – | – | – | NOUN |
| 51    | турбулентен | turbulent | turbulent | Compound | – | – | – | – | – | – | NOUN |
| 52    | аматер | amateur | amateur | Compound | – | – | – | – | – | – | NOUN |
| 53    | атрофия | atrophy | atrophy | Compound | – | – | – | – | – | – | NOUN |
| 54    | аграрно | agrarian | agrarian | Compound | – | – | – | – | – | – | NOUN |
| 55    | аррахноиден | arachnoid | arachnoid | Compound | – | – | – | – | – | – | NOUN |
| 56    | генеалогичен | genealogical | genealogical | Compound | – | – | – | – | – | – | NOUN |
| 57    | декалиметров | decametrical | decametrical | Compound | – | – | – | – | – | – | NOUN |
| 58    | сканерен | scanner | scanner | Compound | – | – | – | – | – | – | NOUN |
| 59    | информационен | informational | informational | Compound | – | – | – | – | – | – | NOUN |
| 60    | измерителен | measuring | measuring | Compound | – | – | – | – | – | – | NOUN |
| 61    | булев | boolean | boolean | Compound | – | – | – | – | – | – | NOUN |
| 62    | електронен | electronic | electronic | Compound | – | – | – | – | – | – | NOUN |
| 63    | библиографски | bibliographical | bibliographical | Compound | – | – | – | – | – | – | NOUN |
| 64    | виник | origin | origin | Compound | – | – | – | – | – | – | NOUN |
| 65    | винченциан | vincentian | vincentian | Compound | – | – | – | – | – | – | NOUN |
| 66    | аристократичен | aristocratic | aristocratic | Compound | – | – | – | – | – | – | NOUN |
| 67    | анехроничен | achronical | achronical | Compound | – | – | – | – | – | – | NOUN |

Table 1: Manual classification of 100 randomly sampled words after classifying all of Bulgarian Wikipedia.
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 $\ell$ of word form $\xi$:

$$p_\theta(\ell \mid \xi) \propto p_\theta(\xi \mid \ell) \pi(\ell).$$

In factoring this generative model, we use character 5-gram models as the language models $p_\theta(\xi \mid \ell)$. The Bulgarian model is trained on Bulgarian ParaMint 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). 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 $\xi$ to its citation form is equivalent to composing a finite-state acceptor representing $\xi$ with the transducer. If the two cannot compose (because $\xi$ is not in the domain of definition (i.e., input language) of the transducer), then we do not suppose that $\xi$ 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 *собе-но, 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-

---

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

---

10UniMorph is a collection of morphological lexica in 167 languages, annotated in a cross-linguistically consistent schema. Apertium is a rule-based machine translation system which includes a finite-state morphological analyzer and generator.
\[ d_{x,y}(i, j) = \min \begin{cases} 
0 & \text{if } i = j = 0, \\
 d_{x,y}(i - 1, j) + 1 & \text{if } i > 0, \\
 d_{x,y}(i, j - 1) + 1 & \text{if } j > 0, \\
 d_{x,y}(i - 1, j - 1) + 1(\mathbf{x}_i \neq \mathbf{y}_j) & \text{if } i, j > 0, \\
 d_{x,y}(i - 2, j - 2) + 1 & \text{if } i, j > 1 \text{ and } \mathbf{x}_i = \mathbf{y}_{j-1} \text{ and } \mathbf{x}_{i-1} = \mathbf{y}_j, 
\end{cases} \]

Figure 5: Recurrence relation the Damerau–Levenshtein distance between two strings \( x \) and \( y \). 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 \( \text{хиндуистките} \) ‘the Hinduistics’ and \( \text{хинду} \) ‘Hindu’:

\begin{align*}
\text{хинду} & \rightarrow \text{хиндуист} \quad \text{(nominal derivation)} \\
\text{хиндуист} & \rightarrow \text{хиндуистка} \quad \text{(diminutive feminine)} \\
\text{хиндуистка} & \rightarrow \text{хиндуистки} \quad \text{(plural)} \\
\text{хиндуистки} & \rightarrow \text{хиндуистките} \quad \text{(definite article)}
\end{align*}

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 \(-\mathbf{k}i\). 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 \( \text{мед ‘copper’} \rightarrow \text{медникар ‘coppersmith’} \) or \( \text{злато ‘gold’} \rightarrow \text{златар ‘goldsmith’} \) 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. *\text{lang}auge* instead of *\text{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. 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 \( \leq 2 \) (note that earlier we consider a compound to be valid if at least one component’s edit distance to a known word is \( \leq 2 \)). 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., джазiformация 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.

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.

\footnote{These values likely need to be adapted to new languages.}
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 Jágrrová, 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 Hajíč and Eva Hajíčková. 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-Shiuee 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 Sillavberg, Timofey Arkhangelskiy, Nataly Krizhanovskaya, Andrew Krizhanovskaya, Elena Klyuchko, Alexey Sorokin, John Mansfield, Valts Ernstruits, 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. McLlroy. 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 Ofazer. 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.

Kirk 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 Kirow, 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 Kirow, 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     | апостол | apostle | Compounded | Declension | N/A | ADJ |
| 2     | христос | Christ | Compounded | Declension | N/A | ADJ |
| 3     | Петродржавност | Petrodollar | Compounded | N/A | N/A | NOUN |
| 4     | кириак | Kyriakos | Compounded | N/A | N/A | NOUN |
| 5     | апостол | apostle | Compounded | N/A | N/A | NOUN |
| 6     | колофон | colophon | Compounded | N/A | N/A | NOUN |
| 7     | приложение | Appendix | Compounded | N/A | N/A | NOUN |
| 8     | интерпретация | Interpretation | Compounded | N/A | N/A | NOUN |
| 9     | таганка | Tagger | Compounded | N/A | N/A | NOUN |
| 10    | полиция | Police | Compounded | N/A | N/A | NOUN |
| 11    | брат | Brother | Compounded | N/A | N/A | NOUN |
| 12    | World of Che | World | Compounded | English | N/A | NOUN |
| 13    | ге | Ge | Compounded | English | N/A | NOUN |
| 14    | mi | Me | Compounded | English | N/A | NOUN |
| 15    | брать | Brother | Compounded | English | N/A | NOUN |
| 16    | стихотворение | Poem | Compounded | English | N/A | NOUN |
| 17    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 18    | таг | Tag | Compounded | English | N/A | NOUN |
| 19    | генетика | Genetics | Compounded | English | N/A | NOUN |
| 20    | обнародовать | Publish | Compound | N/A | N/A | NOUN |
| 21    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 22    | Проспект | Prospect | Compounded | English | N/A | NOUN |
| 23    | Учёный | Scientist | Compounded | English | N/A | NOUN |
| 24    | гиометрия | Geometry | Compounded | English | N/A | NOUN |
| 25    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 26    | генетика | Genetics | Compounded | English | N/A | NOUN |
| 27    | моностоль | Monostyle | Compounded | English | N/A | NOUN |
| 28    | гиометрия | Geometry | Compounded | English | N/A | NOUN |
| 29    | Пират | Pirate | Compounded | English | N/A | NOUN |
| 30    | изра | Kara | Compounded | English | N/A | NOUN |
| 31    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 32    | Критический | Critical | Compounded | English | N/A | NOUN |
| 33    | Трансляция | Translation | Compounded | English | N/A | NOUN |
| 34    | Эксперимент | Experiment | Compounded | English | N/A | NOUN |
| 35    | Бестселлер | Bestseller | Compounded | English | N/A | NOUN |
| 36    | Декларация | Declaration | Compounded | English | N/A | NOUN |
| 37    | изра | Kara | Compounded | English | N/A | NOUN |
| 38    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 39    | Платформа | Platform | Compounded | English | N/A | NOUN |
| 40    | Производственный | Production | Compounded | English | N/A | NOUN |
| 41    | экспортировать | Export | Compounded | English | N/A | NOUN |
| 42    | петрография | Petrography | Compounded | English | N/A | NOUN |
| 43    | химический | Chemical | Compounded | English | N/A | NOUN |
| 44    | экспортировать | Export | Compounded | English | N/A | NOUN |
| 45    | уничтожить | Destroy | Compounded | English | N/A | NOUN |
| 46    | хроматография | Chromatography | Compounded | English | N/A | NOUN |
| 47    | гравюра | Engraving | Compounded | English | N/A | NOUN |
| 48    | ткань | Cloth | Compounded | English | N/A | NOUN |
| 49    | абонент | Subscriber | Compounded | English | N/A | NOUN |
| 50    | петрография | Petrography | Compounded | English | N/A | NOUN |
| 51    | ткань | Cloth | Compounded | English | N/A | NOUN |
| 52    | yoga | Yoga | Compounded | English | N/A | NOUN |
| 53    | Tяжесть | Weight | Compounded | English | N/A | NOUN |
| 54    | маленький | Small | Compounded | English | N/A | NOUN |
| 55    | хозяин | Owner | Compounded | English | N/A | NOUN |
| 56    | Тяжёлый | Heavy | Compounded | English | N/A | NOUN |
| 57    | баритон | Baritone | Compounded | English | N/A | NOUN |
| 58    | плутоний | Plutonium | Compounded | English | N/A | NOUN |
| 59    | кармелит | Carmelite | Compounded | English | N/A | NOUN |
| 60    | квазиэкспериментальный | Quasieperimental | Compounded | English | N/A | NOUN |
| 61    | Мозаика | Mosaic | Compounded | English | N/A | NOUN |
| 62    | Кабина | Cab | Compounded | English | N/A | NOUN |
| 63    | Король | King | Compounded | English | N/A | NOUN |
| 64    | Грек | Greek | Compounded | English | N/A | NOUN |
| 65    | Аристократ | Aristocrat | Compounded | English | N/A | NOUN |
| 66    | Независимость | Independence | Compounded | English | N/A | NOUN |
| 67    | Легко | Easy | Compounded | English | N/A | NOUN |
| 68    | День | Day | Compounded | English | N/A | NOUN |
| 69    | Валното | Wave | Compounded | English | N/A | NOUN |
| 70    | ночь | Night | Compounded | English | N/A | NOUN |
| 71    | Рай | Heaven | Compounded | English | N/A | NOUN |
| 72    | Фронт | Front | Compounded | English | N/A | NOUN |
| 73    | Такт | Measure | Compounded | English | N/A | NOUN |
| 74    | Патруль | Patrol | Compounded | English | N/A | NOUN |
| 75    | Основные | Basic | Compounded | English | N/A | NOUN |
| 76    | Колония | Colony | Compounded | English | N/A | NOUN |
| 77    | Клиника | Hospital | Compounded | English | N/A | NOUN |
| 78    | 613 | 613 | Compounded | English | N/A | NOUN |
| 79    | Меди | Medicine | Compounded | English | N/A | NOUN |
| 80    | Труба | Tube | Compounded | English | N/A | NOUN |
| 81    | Трупная | Trunk | Compounded | English | N/A | NOUN |
| 82    | Основной | Main | Compounded | English | N/A | NOUN |
| 83    | Описание | Description | Compounded | English | N/A | NOUN |
| 84    | Сжатый | Tight | Compounded | English | N/A | NOUN |
| 85    | Паттерн | Pattern | Compounded | English | N/A | NOUN |
| 86    | Объект | Object | Compounded | English | N/A | NOUN |
| 87    | Мастер | Master | Compounded | English | N/A | NOUN |
| 88    | Термин | Term | Compounded | English | N/A | NOUN |
| 89    | асистент | Assistant | Compounded | English | N/A | NOUN |
| 90    | Медицинский | Medical | Compounded | English | N/A | NOUN |
| 91    | подверхность | Subsurface | Compounded | English | N/A | NOUN |
| 92    | Структура | Structure | Compounded | English | N/A | NOUN |
| 93    | 2-й Путешествие | 2nd Trip | Compounded | English | N/A | NOUN |
| 94    | прописанный/полученный | Official | Compounded | English | N/A | NOUN |
| 95    | Указание | Indication | Compounded | English | N/A | NOUN |
| 96    | Осветление | Enlightenment | Compounded | English | N/A | NOUN |
| 97    | повторять | Repeat | Compounded | English | N/A | NOUN |
| 98    | цветной | Colour | Compounded | English | N/A | NOUN |
| 99    | братство | Brotherhood | Compounded | English | N/A | NOUN |
| 100   | Библиотека | Library | Compounded | English | N/A | NOUN |

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.
Table A.2: Manual classification of 100 randomly sampled words from the tokenized Bulgarian Wikipedia Corpus

| Index | Word                                      | Type          | Sub-Type          | Features | POS  |
|-------|-------------------------------------------|---------------|-------------------|----------|------|
| 1     | oat producer                              | Compound      | N/A               | MASC     | NOUN |
| 2     | non-pornographic                          | Compound      | N/A               | MASC     | AD   |
| 3     | fast growing                              | Compound      | N/A               | PL       | AD   |
| 4     | eye declawing                             | Compound      | N/A               | PL       | AD   |
| 5     | the reforming                             | Compound      | N/A               | MASC+DEF | NOUN |
| 6     | the meta cognitive                        | Compound      | N/A               | MASC+DEF | NOUN |
| 7     | of three galaxies                         | Compound      | N/A               | FEM+DEF  | AD   |
| 8     | coach of German naval team                 | Compound      | N/A               | MASC     | AD   |
| 9     | synesthesia                               | Compound      | N/A               | MASC     | NOUN |
| 10    | (pertaining to) aerospace                 | Compound      | N/A               | MASC     | AD   |
| 11    | post-American                            | Compound      | N/A               | MASC+DEF | NOUN |
| 12    | shelling                                  | Compound      | N/A               | MASC     | NOUN |
| 13    | initial                                   | Compound      | N/A               | MASC+DEF | AD   |
| 14    | relevant                                  | Compound      | N/A               | MASC     | AD   |
| 15    | the newly built                           | Compound      | N/A               | MASC+DEF | AD   |
| 16    | primary basis                             | Compound      | N/A               | PL       | AD   |
| 17    | the life guard                            | Compound      | N/A               | PL+DEF   | AD   |
| 18    | drought tolerant                          | Compound      | N/A               | PL+DEF   | AD   |
| 19    | the outflow                               | Compound      | N/A               | MASC+DEF | AD   |
| 20    | the outflow                               | Compound      | N/A               | MASC+DEF | AD   |
| 21    | the tied rope                             | Compound      | N/A               | MASC+DEF | AD   |
| 22    | nailed                                    | Compound      | N/A               | MASC+DEF | AD   |
| 23    | investigating, inquiring                  | Compound      | N/A               | MASC+DEF | AD   |
| 24    | go around, circumvent                      | Compound      | N/A               | MASC+DEF | AD   |
| 25    | moved around                              | Compound      | N/A               | MASC+DEF | AD   |
| 26    | called, natural                           | Compound      | N/A               | MASC+DEF | AD   |
| 27    | the ancient                               | Compound      | N/A               | MASC+DEF | AD   |
| 28    | the international                         | Compound      | N/A               | MASC+DEF | AD   |
| 29    | the synonym                               | Compound      | N/A               | MASC+DEF | AD   |
| 30    | the barrier                               | Compound      | N/A               | MASC+DEF | AD   |
| 31    | the minor branch                          | Compound      | N/A               | MASC+DEF | AD   |
| 32    | the same                                  | Compound      | N/A               | MASC+DEF | AD   |
| 33    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 34    | the boggy                                 | Compound      | N/A               | MASC+DEF | AD   |
| 35    | the tidy rope                             | Compound      | N/A               | MASC+DEF | AD   |
| 36    | the soft web                              | Compound      | N/A               | MASC+DEF | AD   |
| 37    | the estuarine                             | Compound      | N/A               | MASC+DEF | AD   |
| 38    | the resynthesis                           | Compound      | N/A               | MASC+DEF | AD   |
| 39    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 40    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 41    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 42    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 43    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 44    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 45    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 46    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 47    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 48    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 49    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 50    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 51    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 52    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 53    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 54    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 55    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 56    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 57    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 58    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 59    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 60    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 61    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 62    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 63    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 64    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 65    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 66    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 67    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 68    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 69    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 70    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 71    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 72    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 73    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 74    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 75    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 76    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 77    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 78    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 79    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 80    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 81    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 82    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 83    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |
| 84    | the calculating                           | Compound      | N/A               | MASC+DEF | AD   |

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     | нераун     | Yes   |
| 9     | почест     | Yes   |
| 10    | ревабш     | Yes   |
| 11    | редент     | Yes   |
| 12    | живееэг    | Yes   |
| 13    | модулин    | Yes   |
| 14    | гутбонии   | Yes   |
| 15    | джутсуу    | Yes   |
| 16    | камбоури   | Yes   |
| 17    | убеждайт   | Yes   |
| 18    | читирима   | Yes   |
| 19    | антатията | Yes   |
| 20    | водщади    | Yes   |
| 21    | наблядава  | Yes   |
| 22    | присъстит  | Yes   |
| 23    | художникт  | Yes   |
| 24    | обстрцция  | Yes   |
| 25    | преостъпва  | Yes   |
| 26    | асортимент | Yes   |
| 27    | мододрацата| Yes   |
| 28    | присъстввал | Yes   |
| 29    | революцияте| Yes   |
| 30    | числиността| Yes   |
| 31    | продлжавало| Yes   |
| 32    | нароставашата| Yes |
| 33    | пристрлването| Yes|
| 34    | стандрордското| Yes|
| 35    | модерншираниите| Yes|
| 36    | тунд        | No    |
| 37    | течащ      | No    |
| 38    | тодас      | No    |
| 39    | шейдър     | No    |
| 40    | сполнищ   | No    |
| 41    | връчацо    | No    |
| 42    | коденищт  | No    |
| 43    | наполнищ  | No    |
| 44    | невиждац  | No    |
| 45    | струвацо  | No    |
| 46    | китобойци | No    |
| 47    | влайковите| No    |
| 48    | кварковото| No    |
| 49    | радиошуото| No    |
| 50    | семиновско | 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     | дрончетодо           | @|ридчетодо | 2              | No             | No                  |
| 3     | вазодилатирац         | валови|датирац     | 2              | Yes            | No                  |
| 4     | паналбанската         | пан|албанската | 0              | Yes            | Yes                 |
| 5     | трудноподвижност      | трудно|подвижност   | 0              | Yes            | Yes                 |
| 6     | узуньорийскся         | @|райськия 9 | No            | No             | No                  |
| 7     | крайтълните          | крайтълен|ите   | 1              | No             | No                  |
| 8     | епископалиицице       | епископа|лиицице   | 1              | No             | No                  |
| 9     | фотофотареноето      | фото|фотареноето | 0              | Yes            | Yes                 |
| 10    | тескетата            | тес|та | 6              | No             | No                  |
| 11    | видеоноблен            | видео|облен    | 0              | Yes            | Yes                 |
| 12    | дефертрането         | деферт|рането | 1              | Yes            | Yes                 |
| 13    | клашицица            | @|кишициа | 4              | No             | No                  |
| 14    | несатитименталното    | @|менталното 8 | Yes        | No             | No                  |
| 15    | преднубертетна        | пред|нубертетна | 0              | Yes            | Yes                 |
| 16    | экспозиция            | @|позиция 3 | No            | No             | No                  |
| 17    | сложноустроение       | сложно|устроение | 0              | Yes            | Yes                 |
| 18    | минимоникиси          | мини|комикиси | 0              | Yes            | Yes                 |
| 19    | бромалги              | бром|алги    | 1              | Yes            | Yes                 |
| 20    | хиподермата           | хипо|дермата | 0              | Yes            | Yes                 |
| 21    | эозиткия              | эоз|есткия | 1              | No             | Yes                 |
| 22    | колаборанти           | кол|лаборанти|кола|оранти | 1              | Yes            | No                  |
| 23    | древноеврейските      | древ|еврейските | 0              | Yes            | Yes                 |
| 24    | цалалунопицалунопиен  | цало|@ | 16             | No             | No                  |
| 25    | нарамвали             | @|вали|нара|@ | 5              | No             | No                  |
| 26    | друмевите             | @|ите | 6              | No             | No                  |
| 27    | екстрабукалиата       | @|алата | 8              | Yes            | No                  |
| 28    | дзвайзацу             | да|@ | 5              | Yes            | No                  |
| 29    | анааспийките          | анаас|@ | 7              | No             | No                  |
| 30    | петоокласно          | пето|классо | 0              | Yes            | Yes                 |
| 31    | джязформация          | джз|формация | 0              | Yes            | Yes                 |
| 32    | крайдунавски          | край|дунавски | 0              | Yes            | Yes                 |
| 33    | елабуцики            | ела|@ | 5              | No             | No                  |
| 34    | орискът              | орик|ът | 1              | No             | No                  |
| 35    | римокатолическа       | римо|католическа | 0              | Yes            | Yes                 |
| 36    | арондисмана           | ард|мана | 6              | No             | No                  |
| 37    | истанбулчаннокатегория | истанбулчанн|категория | 0              | Yes            | Yes                 |
| 38    | спокобиха             | спо|добиха | 0              | Yes            | Yes                 |
| 39    | прокомуннирана        | прокол|@ | 10             | Yes            | No                  |
| 40    | леонополннините       | леонол|днините | 1              | No             | No                  |
| 41    | детройтът             | детройт|@ | 2              | No             | Yes                 |
| 42    | шипл’ивацита         | @|вица | 5              | No             | No                  |
| 43    | пръмдюстната          | @|ростната | 6              | Yes            | No                  |
| 44    | шестмоторни           | шест|моторни | 0              | Yes            | Yes                 |
| 45    | филмографията         | филм|ографията|филм|зографията | 1              | Yes            | Yes                 |
| 46    | средновистата         | сред|виста | 0              | Yes            | Yes                 |
| 47    | безкуполен           | без|куполен | 0              | Yes            | Yes                 |
| 48    | гоцеелчевската        | гое|елчевската | 2              | Yes            | Yes                 |
| 49    | епископ               | @|коп | 3              | No             | No                  |
| 50    | лопатовиднюхъб       | лопатов|юхъб | 6              | Yes            | No                  |

Table A.4: Human validation of random sample of compound analysis.
Table A.5: Manual classification of 100 randomly sampled words after classifying all of the BulTreeBank corpus, in analogy with Table 1.

| Index Word | Human Trans. | Alg. Trans. | Human Type | Alg. Type | Alg. Sub-Type | Features | POS |
|-------------|---------------|-------------|------------|-----------|---------------|----------|-----|
| miroopazvate | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| novoоткритото | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| русокоси | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| по-нагъл | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| svрхелегантен | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Daunttaun | downtown | N/A | N/A | N/A | N/A | N/A | N/A |
| blond russian | Compound | N/A | N/A | N/A | N/A | N/A | N/A |
| more impudent | Compound | N/A | N/A | N/A | N/A | N/A | N/A |
| manage (abbreviated) | Abbreviation | N/A | N/A | N/A | N/A | N/A | N/A |
| at the peak of | Target | N/A | N/A | N/A | N/A | N/A | N/A |
| Земя | Olелия | Anastasия | Talleyrand | N/A | Name | Person | Proper | Likely | N/A | PROPN | 94 |
| talейран | Lиза | Companis | Elizabeth | N/A | Name | Person | Proper | Likely | N/A | PROPN | 91 |
| Litteraire | Name | Newspaper | Proper | Funky | N/A | PROPN | 90 |
| Central | Name | Hotel | Proper | Likely | MASC | PROPN | 89 |
| Trifon | Name | Person | Proper | Likely | MASC | PROPN | 88 |
| Emilia | Name | Person | Proper | Likely | FEM | PROPN | 87 |
| Ganeva | Name | Person | Proper | Likely | N/A | PROPN | 86 |
| Simon | Name | Person | Proper | Likely | MASC | PROPN | 84 |
| Panov | Name | Person | Proper | Likely | MASC | PROPN | 83 |
| Heat | Name | Movie | Proper | Likely | FEM | NOUN | 81 |
| провинилите | the cold | Declension | N/A | N/A | N/A | MASC+DEF | NOUN | 78 |
| студът | the man | Declension | N/A | N/A | N/A | MASC+DEF | NOUN | 77 |
| социалдемократически | изразходваните | необмислените | холивудските | съвестността | позиционните | эврофондове | тексасците | отдалечила | момченцето | заловеният | The Czech (females) | Check (female) | Declension | N/A | Declension | Simple | FEM+DEF | ADJ (Proper) | 60 |
| occurred | occurred | Declension | N/A | Declension | Simple | NEUT+DEF | PART | 59 |
| sworn | sworn | Declension | N/A | Declension | Simple | MASC+DEF | ADJ | 56 |
| the commands entrusted | Declension | Fuzzy | FEM+PL+DEF | NOUN | 55 |
| the childish toy | Declension | Simple | FEM+DEF | ADJ | 54 |
| The great | Declension | Simple | PL | ADJ | 52 |
| epoхата | thunder report | Declension | N/A | Declension | Fuzzy | PL | NOUN | 50 |
| болките | божиите | beak click | Declension | N/A | Compound | N/A | FEM+PL | NOUN | 47 |
| манталитетът | deals given | friends Declension | N/A | Compound | N/A | FEM+PL | NOUN | 45 |
| Reedsdale | Geography | English | Proper | Likely | N/A | PROPN | 43 |
| Kozro | Geography | Russian | Proper | Likely | N/A | PROPN | 42 |
| Dover | Geography | English | Proper | Likely | N/A | PROPN | 41 |
| Bos | Geography | Russian | Proper | Likely | N/A | PROPN | 40 |
| Поно | Geography | English | Proper | Likely | N/A | PROPN | 39 |
| Broken | Geography | English | Proper | Likely | N/A | PROPN | 38 |
| hungarian | bargains | bargain Declension | N/A | Declension | Simple | MASC PL | NOUN | 57 |
| Beatrice | Authors | English | Proper | Likely | MASC | PROPN | 56 |
| the commands | commanded | Declension | N/A | Declension | Simple | MASC+DEF | ADJ | 55 |
| the man | Declension | N/A | N/A | N/A | N/A | PROPN | 54 |
| the commands | commanded | Declension | N/A | Declension | Simple | MASC+DEF | ADJ | 53 |
| the house | Declension | N/A | N/A | N/A | N/A | PROPN | 52 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 51 |
| the congress | congressmen Declension | N/A | Declension | Simple | MASC+DEF | ADJ | 50 |
| the congress | congressmen Declension | N/A | Declension | Simple | MASC+DEF | ADJ | 49 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 48 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 47 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 46 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 45 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 44 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 43 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 42 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 41 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 40 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 39 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 38 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 37 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 36 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 35 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 34 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 33 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 32 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 31 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 30 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 29 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 28 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 27 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 26 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 25 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 24 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 23 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 22 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 21 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 20 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 19 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 18 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 17 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 16 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 15 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 14 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 13 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 12 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 11 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 10 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 9 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 8 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 7 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 6 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 5 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 4 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 3 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 2 |
| the depression | Declension | N/A | N/A | N/A | N/A | PROPN | 1 |