Word class flexibility: A deep contextualized approach

Bai Li[1,4], **Guillaume Thomas**[2], **Yang Xu**[1,3,4], **Frank Rudzicz**[1,4,5]

[1] Department of Computer Science, University of Toronto, Toronto, Canada
[2] Department of Linguistics, University of Toronto, Toronto, Canada
[3] Cognitive Science Program, University of Toronto, Toronto, Canada
[4] Vector Institute for Artificial Intelligence, Toronto, Canada
[5] St. Michael's Hospital, Toronto, Canada
`{bai, yangxu, frank}@cs.toronto.edu`
`guillaume.thomas@utoronto.ca`

Abstract

Word class flexibility refers to the phenomenon whereby a single word form is used across different grammatical categories. Extensive work in linguistic typology has sought to characterize word class flexibility across languages, but quantifying this phenomenon accurately and at scale has been fraught with difficulties. We propose a principled methodology to explore regularity in word class flexibility. Our method builds on recent work in contextualized word embeddings to quantify semantic shift between word classes (e.g., noun-to-verb, verb-to-noun), and we apply this method to 37 languages[1]. We find that contextualized embeddings not only capture human judgment of class variation within words in English, but also uncover shared tendencies in class flexibility across languages. Specifically, we find greater semantic variation when flexible lemmas are used in their dominant word class, supporting the view that word class flexibility is a directional process. Our work highlights the utility of deep contextualized models in linguistic typology.

1 Introduction

In natural languages, lexical items can often be used in multiple word classes without overt changes in word form. For instance, the word *buru* in Mundari can be used as a noun to denote 'mountain', or as a verb to denote 'to heap up' (Evans and Osada, 2005). Known as word class flexibility, this phenomenon is considered one of the most challenging topics in linguistic typology (Evans and Levinson, 2009). We present a computational methodology to quantify the regularity in word class flexibility across languages.

There is an extensive literature on how languages vary in word class flexibility, either directly (Hengeveld, 1992; Vogel and Comrie, 2000; Van Lier and Rijkhoff, 2013) or through related notions such as word class conversion (with zero-derivation) (Vonen, 1994; Don, 2003; Bauer and Valera, 2005a; Manova, 2011; Ştekauer et al., 2012). However, existing studies tend to rely on analyses of small sets of lexical items that may not be representative of word class flexibility in the broad lexicon. Critically lacking are systematic analyses of word class flexibility across many languages, and existing typological studies have only focused on qualitative comparisons of word class systems.

We take to our knowledge the first step towards computational quantification of word class flexibility in 37 languages, taken from the Universal Dependencies project (Zeman et al., 2019). We focus on lexical items that can be used both as nouns and as verbs, i.e., noun-verb flexibility. This choice is motivated by the fact that the distinction between nouns and verbs is the most stable in word class systems across languages: if a language makes any distinction between word classes at all, it will likely be a distinction between nouns and verbs (Hengeveld, 1992; Evans, 2000; Croft, 2003). However, our understanding of cross-linguistic regularity in noun-verb flexibility is impoverished.

We operationalize word class flexibility as a property of lemmas. We define a lemma as flexible if some of its occurrences are tagged as nouns and others as verbs. Flexible lemmas are sorted into noun dominant lemmas, which occur more frequently as nouns, and verb dominant lemmas that occur more frequently as verbs. Our methodology builds on contextualized word embedding models (e.g., ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019)) to quantify semantic shift be-

---

[1]Code and data to reproduce the experimental findings are available at: `https://github.com/SPOClab-ca/word-class-flexibility`.

tween grammatical classes of a lemma, within a single language. This methodology can also help quantify metrics of flexibility in the lexicon across languages.

We use our methodology to address one of the most fundamental questions in the study of word class flexibility: should this phenomenon be analyzed as a directional word-formation process similar to derivation, or as a form of underspecification? Derived words are commonly argued to have a lower frequency of use and a narrower range in meaning compared to their base (Marchand, 1964; Iacobini, 2000). If word class flexibility is a directional process, we should expect that flexible lemmas are subject to more semantic variation in their dominant word class than in their less frequent class. We also test the claim that noun-to-verb flexibility involves more semantic shift than verb-to-noun flexibility. While previous work has explored these questions, it remains challenging to quantify semantic shift and semantic variation, particularly across different languages.

We present a novel probing task that reveals the ability of deep contextualized models to capture semantic information across word classes. Our utilization of deep contextual models predicts human judgment on the spectrum of noun-verb flexible usages including homonymy (unrelated senses), polysemy (different but related senses), and word class flexibility. We find that BERT outperforms ELMo and non-contextual word embeddings, and that the upper layers of BERT capture the most semantic information, which resonates with existing probing studies (Tenney et al., 2019).

2 Related work and assumptions

2.1 Types of flexibility

The phenomenon of word class flexibility has been analyzed in different ways. One way is to assume the existence of underspecified word classes. For instance, Hengeveld (2013) claims that basic lexical items in Mundari belong to a single class of *contentives* that can be used to perform all the functions associated with nouns, verbs, adjectives or adverbs in a language like English. Alternatively, word class flexibility can be analyzed as a form of conversion, i.e., as a relation between words that have the same form and closely related senses but different word classes, such as *a fish* and *to fish* in English (Adams, 1973). Conversion has been analyzed as a derivational process that

relates different lexemes (Jespersen, 1924; Marchand, 1969; Quirk et al., 1985), or as a property of lexemes whose word class is underspecified (Farell, 2001; Barner and Bale, 2002). We use word class flexibility as a general term that subsumes these different notions. This allows us to assess whether there is evidence that word class flexibility should be characterized as a directional word formation process, rather than as a form of underspecification.

2.2 Homonymy and polysemy

Word class flexibility has often been analyzed in terms of homonomy and polysemy (Valera and Ruz, 2020). Homonymy is a relation between lexemes that share the same word form but are not semantically related (Cruse, 1986, p.80). Homonyms may differ in word class, such as *ring* 'a small circular band' and *ring* 'make a clear resonant or vibrating sound.' Polysemy is defined as a relation between different senses of a single lexeme (*ibid.*). Insofar as the nominal and verbal uses of flexible lexical items are semantically related, one may argue that word class flexibility is similar to polysemy, and must be distinguished from homonymy. In practice, homonymy and polysemy exist on a continuum, so it is difficult to apply a consistent criterion to differentiate them (Tuggy, 1993). As a consequence, we will not attempt to tease homonymy apart from word class flexibility.

Regarding morphology, word class flexibility excludes pairs of lexical items that are related by overt derivational affixes, such as *to act/an actor*. In such cases, word class alternations can be attributed to the presence of a derivational affix, and are therefore part of regular morphology. In contrast, we allow tokens of flexible lexical items to differ in inflectional morphology.

2.3 Directionality of class conversion

Word class flexibility can be analyzed either as a static relation between nominal and verbal uses of a single lexeme, or as a word formation process related to derivation. The merits of each analysis have been extensively debated in the literature on conversion (see e.g., Farell, 2001; Don, 2005). One of the objectives of our study is to show that deep contextualized language models can be used to help resolve this debate. A hallmark of derivational processes is their directionality. Direction of derivation can be established using several synchronic criteria, among which are the principles

that a derived form tends to have a lower frequency of use and a smaller range of senses than its base (Marchand, 1964; Iacobini, 2000). In languages where word class flexibility is a derivational process, one should therefore expect greater semantic variation when flexible lemmas are used in their dominant word class—an important issue that we verify with our methodology.

Semantic variation has been operationalized in several ways. Kisselew et al. (2016) uses an entropy-based metric, while Balteiro (2007) and Bram (2011) measure semantic variation by counting the number of different noun and verb senses in a dictionary. The latter study found that the more frequent word class has greater semantic variation at a rate above random chance. Here we propose a novel metric based on contextual word embeddings to compare the amount of semantic variation of flexible lemmas in their dominant and non-dominant grammatical classes. Differing from existing methods, our metric is validated explicitly on human judgements of semantic similarity, and can be applied to many languages without the need for dictionary resources.

2.4 Asymmetry in semantic shift

If word class flexibility is a directional process, a natural question is whether derived verbs stand in the same semantic relation to their base as derived nouns. The literature on conversion suggests that there might be significant differences between these two directions of derivation. In English, verbs that are derived from nouns by conversion have been argued to describe events that include the noun's denotation as a participant (e.g. *hammer*, 'to hit something with a hammer') or as a spatio-temporal circumstance (*winter* 'to spend the winter somewhere'). Clark and Clark (1979) argue that the semantic relations between denominal verbs and their base are so varied that they cannot be given a unified description. In comparison, when the base of conversion is a verb, the derived noun most frequently denotes an event of the sort described by the verb (e.g. *throw* 'the act of throwing something'), or the result of such an act (e.g. *release* 'state of being set free') (Jespersen, 1942; Marchand, 1969; Cetnarowska, 1993). This has led some authors to suggest that verb to noun conversion in English involves less semantic shift than noun to verb conversion (Bauer, 2005, p.22). Here we consider a new metric of semantic shift based

on contextual embeddings, and we use this metric to test the hypothesis that the expected semantic shift involved in word class flexibility is greater for noun dominant lexical items (as compared to verb dominant lexical items) in our sample of languages. As we will show, this proposal is consistent with the empirical observation that verb-to-noun conversion is statistically more salient than noun-to-verb conversion.

2.5 Contextualized language models

Deep contextualized language models take a sequence of natural language tokens and produce a sequence of context-sensitive embeddings for each token. These embeddings can be used in a variety of downstream tasks and have achieved state-of-the-art performance in many of them. There are many models that generate contextual embeddings, generally trained with unsupervised learning using a large corpus. In particular, ELMo (Peters et al., 2018) uses a left-to-right and a right-to-left LSTM trained to minimize perplexity across a large corpus. To generate contextual embeddings, it feeds the sentence through both LSTMs and concatenates the left-to-right and right-to-left LSTM states. BERT (Devlin et al., 2019) uses 12 layers of the Transformer module (Vaswani et al., 2017) and is pre-trained on a large corpus using two tasks: masked language modeling to predict randomly masked tokens from context, and next sentence prediction to predict whether two sentences are contiguous in the original text.

Both ELMo and BERT can be adapted to non-English languages without modification. The authors of BERT trained multilingual BERT (mBERT) by concatenating Wikipedia for 104 languages. There are models designed specifically for multilingual situations: XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020) are similar to BERT, but include an additional pre-training objective that leverages parallel text.

Typically, BERT is used in combination with task-specific modules and the parameters fine-tuned using domain data. Here we use contextual embeddings without fine-tuning. Probing experiments revealed that BERT embeddings contain semantic information beyond static embeddings, especially in the upper layers (Tenney et al., 2019), and this information is demonstrably useful for word sense disambiguation (Wiedemann et al., 2019; Hadiwinoto et al., 2019).

3 Identification of word class flexibility

3.1 Definitions

A lemma is *flexible* if it can be used both as a noun and as a verb. To reduce noise, we require each lemma to appear at least 10 times and at least 5% of the time as the minority class to be considered flexible. The *inflectional paradigm* of a lemma is the set of words that have the lemma.

A flexible lemma is *noun (verb) dominant* if it occurs more often as a noun (verb) than as a verb (noun). This is merely an empirical property of a lemma: we do not claim that the base POS should be determined by frequency. The *noun (verb) flexibility* of a language is the proportion of noun (verb) dominant lemmas that are flexible.

3.2 Datasets and preprocessing

Our experiments require corpora containing part-of-speech annotations. For English, we use the British National Corpus (BNC), consisting of 100M words of written and spoken English from a variety of sources (Leech, 1992). Root lemmas and POS tags are provided, and were generated automatically using the CLAWS4 tagger (Leech et al., 1994). For our experiments, we use BNC-baby, a subset of BNC containing 4M words.

For other languages, we use the Universal Dependencies (UD) treebanks of over 70 languages, annotated with lemmatizations, POS tags, and dependency information (Zeman et al., 2019). We concatenate the treebanks for each language and use the languages that have at least 100k tokens.

The UD treebanks are too small for our contextualized experiments and are not matched for content and style, so we supplement them with Wikipedia text[2]. For each language, we randomly sample 10M tokens from Wikipedia; we then use UDPipe 1.2 (Straka and Straková, 2017) to tokenize the text and generate POS tags for every token. We do not use the lemmas provided by UD-Pipe, but instead use the lemma merging algorithm to group lemmas.

3.3 Lemma merging algorithm

The UD corpus provides lemma annotations for each word, but these lemmas are insufficient for our purposes because they do not always capture instances of flexibility. In some languages, nouns

---

[2]We use Wikiextractor to extract text from Wikimedia dumps: https://github.com/attardi/wikiextractor.

Language	Nouns	Verbs	Noun flexibility	Verb flexibility
Arabic	1517	299	0.076	0.221
Bulgarian	786	343	0.039	0.047
Catalan	1680	590	0.039	0.147
Chinese	1325	634	0.125	0.391
Croatian	1031	370	0.042	0.062
Danish	324	216	0.108	0.269
Dutch	958	441	0.077	0.188
English	1700	600	0.248	0.472
Estonian	1949	592	0.032	0.115
Finnish	1523	631	0.028	0.136
French	1844	649	0.062	0.257
Galician	802	334	0.031	0.135
German	4239	1706	0.049	0.229
Hebrew	850	315	0.111	0.321
Indonesian	572	243	0.052	0.128
Italian	2227	770	0.067	0.256
Japanese	1105	417	0.178	0.566
Korean	1890	1003	0.026	0.048
Latin	1090	885	0.056	0.122
Norwegian	1951	636	0.072	0.259
Old Russian	527	416	0.034	0.060
Polish	2054	1084	0.069	0.427
Portuguese	1711	638	0.037	0.185
Romanian	1809	740	0.060	0.151
Slovenian	746	316	0.068	0.123
Spanish	2637	873	0.046	0.202
Swedish	784	384	0.038	0.109
Excluded Languages				
Ancient Greek	1098	1022	0.015	0.026
Basque	650	247	0.020	0.105
Czech	5468	2063	0.004	0.011
Hindi	1364	133	0.019	0.135
Latvian	1159	603	0.022	0.061
Persian	1125	47	0.010	0.234
Russian	3909	1760	0.005	0.024
Slovak	488	281	0.006	0.011
Ukrainian	659	238	0.006	0.029
Urdu	722	51	0.018	0.216

Table 1: Noun and verb flexibility for 37 languages with at least 100k tokens in the UD corpus. We include the 27 languages with over 2.5% noun and verb flexibility; 10 languages are excluded from further analysis.

and verbs are lemmatized to different forms by convention. For example, in French, the word *voyage* can be used as a verb (*il voyage* 'he travels') or as a noun (*un voyage* 'a trip'). However, verbs are lemmatized to the infinitive *voyager*, whereas nouns are lemmatized to the singular form *voyage*. Since the noun and verb lemmas are different, it is not easy to identify them as having the same stem.

The different lemmatization conventions of French and English reflect a more substantial linguistic difference. French has a stem-based morphology, in which stems tend to occur with an inflectional ending. By contrast, English has a word-based morphology, where stems are commonly used as free forms (Kastovsky, 2006). This dif-

ference is relevant to the definition of word class flexibility: in stem-based systems, flexible items are stems that may not be attested as free forms (Bauer and Valera, 2005b, p.14).

We propose a heuristic algorithm to capture stem-based flexibility as well as word-based flexibility. The key observation is that the inflectional paradigms of the noun and verb forms often have some words in common (such is the case for *voyager*). Thus, we merge any two lemmas whose inflectional paradigms have a nonempty intersection. This is implemented with a single pass through the corpus, using the union-find data structure: for every word, we call UNION on the inflected form and the lemmatized form.

Using this heuristic, we can identify cases of flexibility that do not share the same lemma in the UD corpus (Table 1). This method is not perfect, and is unable to identify cases of stem-based flexibility where the inflectional paradigms don't intersect, for example in French, *chant* 'song' and *chants* 'songs' are not valid inflections of the verb *chanter* 'to sing'. There are also false positives that cause two unrelated lemmas to be merged if their inflectional paradigms intersect, for example, *avions* (plural form of *avion* 'airplane') happens to have the same form as *avions* (first person plural imperfect form of *avoir* 'to have').

4 Methodology and evaluation

4.1 Probing test of contextualized model

Deep contextual embeddings can capture a variety of information other than semantics, which can introduce noise into our results, for example: the lexicographic form of a word, syntactic position, etc. In order to compare different contextual language models on how well they capture semantic information, we perform a probing test of how accurate the models can capture human judgements of word sense similarity.

We begin with a list of the 138 most frequent flexible words in the BNC corpus. Some of these words are flexible (e.g., *work*), while others are homonyms (e.g., *bear*). For each lemma, we get five human annotators from Mechanical Turk to make a sentence using the word as a noun, then make a sentence using the word as a verb, then rate the similarity of the noun and verb senses on a scale from 0 to 2. The sentences are used for quality assurance, so that ratings are removed if the sentences are nonsensical. We will call the av-
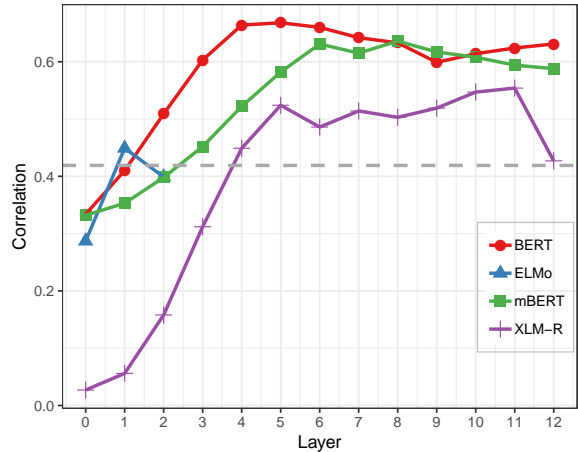

Figure 1: Spearman correlations between human and model similarity scores for ELMo, BERT, mBERT, and XLM-R. The dashed line is the baseline using static GloVe embeddings.

erage human rating for each word the *human similarity score*.

Next, we evaluate each layer of ELMo, BERT, mBERT, and XLM-R[3] on correlation with the human similarity score. That is, we compute the mean of the contextual vectors for all noun instances of the given word in the BNC corpus, the mean across all verb instances, then compute the cosine distance between the two mean vectors as the model's similarity score. Finally, we evaluate the Spearman correlation of the human and model's similarity scores for 138 words: this score measures the model's ability to gauge the level of semantic similarity between noun and verb senses, compared to human judgements.

For a baseline, we do the same procedure using non-contextual GloVe embeddings (Pennington et al., 2014). Note that while all instances of the same word have a static embedding, different words that share the same lemma still have different embeddings (e.g., *work* and *works*), so that the baseline is not trivial.

The correlations are shown in Figure 1. BERT and mBERT are better than ELMo and XLM-R at capturing semantic information, in all transformer models, the correlation increases for each layer up until layer 4 or so, and after this point, adding more layers neither improves nor degrades the performance. Thus, unless otherwise noted, we use the final layers of each model for down-

---

[3]We use the models 'bert-base-uncased', 'bert-base-multilingual-cased', and 'xlm-roberta-base' from Wolf et al. (2019).
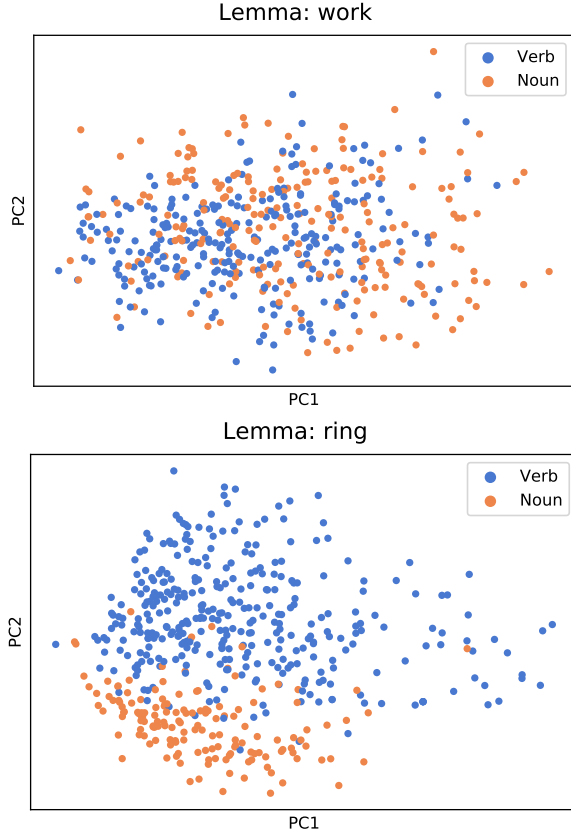
Figure 2: PCA plot of BERT embeddings for the lemmas "work" (high similarity between noun and verb senses) and "ring" (low similarity).

stream tasks.

Figure 2 illustrates the contextual distributions for two lemmas on the opposite ends of the noun-verb similarity spectrum: *work* (human similarity score: 2) and *ring* (human similarity score: 0). We apply PCA to the BERT embeddings of all instances of each lemma in the BNC corpus. For *work*, the noun and verb senses are very similar and the distributions have high overlap. In contrast, for *ring*, the most common noun sense ('a circular object') is etymologically and semantically unrelated to the most common verb sense ('to produce a resonant sound'), and accordingly, their distributions have very little overlap.

4.2 Three contextual metrics

We define three metrics based on contextual embeddings to measure various semantic aspects of word class flexibility. We start by generating contextual embeddings for each occurrence of every flexible lemma. For each lemma l, let $E_{n,l}$ and $E_{v,l}$ be the set of contextual embeddings for noun and verb instances of l. We define the *prototype noun vector* $\mathbf{p}_{n,l}$ of a lemma l as the mean of em-

beddings across noun instances, and the *noun variation* $V_{n,l}$ as the mean Euclidean distance from each noun instance to the noun vector:

$$\mathbf{p}_{n,l} = \frac{1}{|E_{n,l}|} \sum_{\mathbf{x} \in E_{n,l}} \mathbf{x} \quad (1)$$

$$V_{n,l} = \frac{1}{|E_{n,l}|} \sum_{\mathbf{x} \in E_{n,l}} ||\mathbf{x} - \mathbf{p}_{n,l}|| \quad (2)$$

The *prototype verb vector* $\mathbf{p}_{v,l}$ and *verb variation* $V_{v,l}$ for a lemma l are defined similarly:

$$\mathbf{p}_{v,l} = \frac{1}{|E_{v,l}|} \sum_{\mathbf{x} \in E_{v,l}} \mathbf{x} \quad (3)$$

$$V_{v,l} = \frac{1}{|E_{v,l}|} \sum_{\mathbf{x} \in E_{v,l}} ||\mathbf{x} - \mathbf{p}_{v,l}|| \quad (4)$$

Lemmas are included if they appear at least 30 times as nouns and 30 times as verbs. To avoid biasing the variation metric towards the majority class, we downsample the majority class to be of equal size as the minority class before computing the variation. The method does not filter out pairs of lemmas that are arguably homonyms rather than flexible (section 2.2); we choose to include all of these instances rather than set an arbitrary cutoff threshold.

We now define language-level metrics to measure the asymmetries hypothesized in sections 2.3 and 2.4. The *noun-to-verb shift (NVS)* is the average cosine distance between the prototype noun and verb vectors for noun dominant lemmas, and the *verb-to-noun shift (VNS)* likewise for verb dominant lemmas:

$$NVS = 1 - \mathbb{E}_{l \text{ noun-dominant}}[\cos(\mathbf{p}_{n,l}, \mathbf{p}_{v,l})] \quad (5)$$

$$VNS = 1 - \mathbb{E}_{l \text{ verb-dominant}}[\cos(\mathbf{p}_{n,l}, \mathbf{p}_{v,l})] \quad (6)$$

We define the *noun (verb) variation* of a language as the average of noun (verb) variations across all lemmas. Finally, define the *majority variation* of a language as the average of the variation of the dominant POS class, and the *minority variation* as the average variation of the smaller POS class, across all lemmas.

5 Results

5.1 Identifying flexible lemmas

Of the 37 languages in UD with at least 100k tokens; in 27 of them, at least 2.5% of verb and noun lemmas are flexible, which we take to indicate that

Language	N→V shift	V→N shift	Noun variation	Verb variation	Majority variation	Minority variation
Arabic	0.098	0.109	8.268	8.672***	8.762***	8.178
Bulgarian	0.146	0.136	8.267	8.409	8.334	8.341
Catalan	0.165	0.169	8.165	8.799***	8.720***	8.244
Chinese	0.072	0.070	7.024	7.212***	7.170***	7.067
Croatian	0.093	0.144**	8.149	8.109	8.219**	8.037
Danish	0.103	0.110	8.245	8.338	8.438***	8.146
Dutch	0.146	0.174	7.716	8.786***	8.354*	8.148
English	0.175*	0.160	8.035	8.624***	8.390***	8.268
Estonian	0.105	0.103	7.800	7.902	8.022**	7.679
Finnish	0.100	0.114	7.972	7.854	8.181***	7.644
French	0.212	0.204	8.189	9.472***	9.082***	8.578
Galician	0.111	0.117	7.922	8.340***	8.137	8.127
German	0.382	0.355	8.078	9.758***	9.096**	8.740
Hebrew	0.121	0.130	8.096	9.116***	8.574	8.638
Indonesian	0.034	0.048	7.100	7.076	7.076	7.101
Italian	0.207	0.184	8.520	9.345***	9.149***	8.716
Japanese	0.061	0.057	7.419***	7.173	7.309	7.283
Latin	0.092	0.139***	7.920***	7.710	7.905***	7.724
Norwegian	0.133	0.132	8.112	8.336***	8.332***	8.116
Polish	0.090	0.080	8.318	8.751***	8.670***	8.399
Portuguese	0.186	0.155	7.907	8.921***	8.642***	8.187
Romanian	0.175	0.145	8.682	8.658	8.934***	8.406
Slovenian	0.093	0.113	8.046	7.983	8.177***	7.853
Spanish	0.235	0.214	7.898	8.961***	8.691***	8.168
Swedish	0.088	0.082	8.262*	8.147	8.328***	8.081
Overall	**1 of 3**	**2 of 3**	**3 of 17**	**14 of 17**	**20 of 20**	**0 of 20**

Table 2: Semantic metrics for 25 languages, computed using mBERT and 10M tokens of Wikipedia text for each language. Asterisks denote significance at $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. For the "Overall" row, we count the languages with a significant tendency towards one direction, out of the number of languages with statistical significance towards either direction (with $p < 0.05$ treated as significant).

word class flexibility exists in the language (Table 1). The lemma merging algorithm is crucial for identifying word class flexibility: only 6 of the 37 languages pass the 2.5% flexibility threshold using the default lemma annotations provided in UD[4]. Languages differ in their prevalence of word class flexibility, but every language in our sample has higher verb flexibility than noun flexibility.

5.2 Asymmetry in semantic metrics

Table 2 shows the values of the three metrics, computed using mBERT and Wikipedia data for 25 languages[5]. For testing significance, we use the unpaired Student's t-test to compare N-V versus V-N shift, and the paired Student's t-test for the other two metrics[6]. The key findings are as follows:

1. **Asymmetry in semantic shift.** In English, N-V shift is greater than V-N shift, in agreement with Bauer (2005). However, this pattern does not hold in general: there is no significant difference in either direction in most languages, and two languages exhibit a difference in the opposite direction as English.

---

[4]Chinese, Danish, English, Hebrew, Indonesian, and Japanese pass the flexibility threshold without the lemma merging algorithm.

[5]We exclude 2 of the 27 languages that we identify word class flexibility. Old Russian was excluded because it is not supported by mBERT; Korean is excluded because the lemma annotations deviate from the standard UD format.

[6]We do not apply the Bonferroni correction for multiple comparisons, because we make claims for trends across all languages, and not for any specific languages.

Dataset	Model	N→V shift	V→N shift	Noun variation	Verb variation	Majority variation	Minority variation
BNC	ELMo	0.389*	0.357	20.261	20.455	20.329	20.388
	BERT	0.122*	0.112	9.015	9.074	9.100***	8.989
	mBERT	0.189*	0.169	7.211	8.401***	7.875**	7.717
	XLM-R	0.004	0.005	2.058	2.374***	2.262	2.170
Wikipedia	ELMo	0.339***	0.330	22.556	22.521	22.463	22.614*
	BERT	0.120***	0.100	9.218***	8.944	9.118**	9.044
	mBERT	0.175*	0.160	8.035	8.624***	8.390***	8.268
	XLM-R	0.004**	0.003	1.966	1.954	1.946	1.974

Table 3: Comparison of semantic models on BNC and Wikipedia datasets (English), computed using several different language models. Asterisks denote significance at $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

2. **Asymmetry in semantic variation between noun and verb usages.** Of the 17 languages with a statistically significant difference in noun versus verb variation, 14 of them have greater verb variation than noun variation.

3. **Asymmetry in semantic variation between majority and minority classes.** All of the 20 languages with a statistically significant difference in majority and minority variation have greater majority variation.

5.3 Model robustness

Next, we assess the robustness of our metrics with respect to choices of corpus and language model. Robustness is desirable because it gives confidence that our models capture true linguistic tendencies, rather than artifacts of our datasets or the models themselves. We compute the three semantic metrics on the BNC and Wikipedia datasets, using all 4 contextual language models: ELMo, BERT, mBERT, and XLM-R. Table 3 summarizes the results from this experiment.

We find that in almost every case where there is a statistically significant difference, all models agree on the direction of the difference. One exception is that noun variation is greater when computed using Wikipedia data than when using the BNC corpus. Wikipedia has many instances of nouns used in technical senses (e.g., *ring* is a technical term in mathematics and chemistry), whereas similar nonfiction text is less common in the BNC corpus.

6 Discussion

6.1 Frequency asymmetry

Every language in our sample has verb flexibility greater than noun flexibility. The reasons for this asymmetry are unclear, but may be due to semantic differences between nouns and verbs. We note that every language in our sample has more noun lemmas than verb lemmas, a pattern that was also attested by Polinsky (2012), although this does not provide an explanation of the observed phenomenon. We leave further exploration of the flexibility asymmetry to future work.

6.2 Implications for theories of flexibility

There is a strong cross-linguistic tendency for the majority word class of a flexible lemma to exhibit more semantic variation than the minority class. In other words, the frequency and semantic variation criteria of determining the base of a conversion pair agree more than at chance. This supports the analysis of word class flexibility as a directional process of conversion, as opposed to underspecification (section 2.3)[7]. Within a flexible lemma, verbs exhibit more semantic variation than nouns. It is attested across many languages that nouns are more physically salient, while verbs have more complex event and argument structure, and are harder for children to acquire than nouns (Gentner, 1982; Imai et al., 2008). Thus, verbs are expected to have greater semantic variation than nouns, which our results confirm. More importantly, for our purposes, this metric serves as a control for the previous metric. Flexible lemmas are more likely to be noun-dominant than verb-dominant, so could the majority and minority variation simply be proxies for noun and verb variation, respectively? In fact, we observe greater verb than noun variation, so this cannot be the case.

---

[7]Since 18 of the 25 languages for which semantic metrics were calculated are Indo-European, it is unclear whether these results generalize to non-Indo-European languages.

Finally, as suggested by Bauer (2005), we find evidence in English that N-V flexibility involves more semantic shift than V-N flexibility, and the pattern is consistent across multiple models and datasets (Table 3). However, this pattern is idiosyncratic to English and not a cross-linguistic tendency. It is thus instructive to analyze multiple languages in studying word class flexibility, as one can easily be misled by English-based analyses.

7 Conclusion

We use contextual language models to examine shared tendencies in word class flexibility across languages. We find that the majority class often exhibits more semantic variation than the minority class, supporting the view that word class flexibility is a directional process. We also find that in English, noun-to-verb flexibility is associated with more semantic shift than verb-to-noun flexibility, but this is not the case for most languages.

Our probing task reveals that the upper layers of BERT contextual embeddings best reflect human judgment of semantic similarity. We obtain similar results in different datasets and language models in English that support the robustness of our method. This work demonstrates the utility of deep contextualized models in linguistic typology, especially for characterizing cross-linguistic semantic phenomena that are otherwise difficult to quantify.

Acknowledgments

References

Valerie Adams. 1973. *An Introduction to Modern English Word Formation*. Longman, London.

Isabel Balteiro. 2007. *The directionality of conversion in English: A dia-synchronic study*, volume 59 of *Linguistics Insights*. Peter Lang.

David Barner and Alan Bale. 2002. No nouns, no verbs: psycholinguistic arguments in favor of lexical underspecification. *Lingua*, 112:771–791.

Laurie Bauer. 2005. Conversion and the notion of lexical category. In Laurie Bauer and Salvador Valera, editors, *Approaches to Conversion/Zero-derivation*, pages 19–30. Waxmann, Münster.

Laurie Bauer and Salvador Valera, editors. 2005a. *Approaches to Conversion/Zero-derivation*. Waxmann, Münster.

Laurie Bauer and Salvador Valera. 2005b. Conversion or zero-derivation: an introduction. In Laurie Bauer and Salvador Valera, editors, *Approaches to Conversion/Zero-derivation*, pages 7–18. Waxmann, Münster.

Barli Bram. 2011. *Major total conversion in English: The question of directionality*. Ph.D. thesis, Victoria University of Wellington.

Bożena Cetnarowska. 1993. *The Syntax, Semantics and Derivation of Bare Nominalizations*. Uniwersytet Śląski, Katowice.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

William Croft. 2003. *Typology and Universals*, 2 edition. Cambridge University Press, Cambridge.

David Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

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

Jan Don. 2003. A note on conversion in Dutch and German. *Linguistics in the Netherlands*, pages 33–44.

Jan Don. 2005. On conversion, relisting and zero-derivation: A comment on Rochelle Lieber: English word-formation processes. *SKASE Journal of Theoretical Linguistics*, 2(2):2–16.

Nicholas Evans. 2000. Word classes in the world's languages. In Geert Booij, Christian Lehmann, and Joachim Mugdan, editors, *Morphologie/morphology: An international handbook on inflection and word-formation*, pages 708–732. de Gruyter, Berlin.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32:429–492.

Nicholas Evans and Toshiki Osada. 2005. Mundari: The myth of a language without word classes. *Linguistic Typology*, 9(3):351–390.

Patrick Farell. 2001. Functional shift as category underspecification. *English Language and Linguistics*, 5(1):109–130.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5300–5309.

Kees Hengeveld. 1992. *Non-verbal predication: theory, typology, diachrony*. Mouton de Gruyter, Berlin, New York.

Kees Hengeveld. 2013. Parts-of-speech systems as a basic typological determinant. In Eva Van Lier and Jan Rijkhoff, editors, *Flexible word classes: a typological study of underspecified parts-of-speech*, pages 31–55. Oxford University Press, Oxford.

Claudio Iacobini. 2000. Base and direction of derivation. In Geert Booij, Christian Lehmann, and Joachim Mugdan, editors, *Morphology: An International Handbook on Inflection and Word-Formation*, pages 865–876. Walter de Gruyter, Berlin.

Mutsumi Imai, Lianjing Li, Etsuko Haryu, Hiroyuki Okada, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Jun Shigematsu. 2008. Novel noun and verb learning in Chinese-, English-, and Japanese-speaking children. *Child development*, 79(4):979–1000.

Otto Jespersen. 1924. *The Philosophy of Grammar*. Allen & Unwin., London.

Otto Jespersen. 1942. *A Modern English Grammar. On Historical Principles. Part VI Morphology*. Allen & Unwin., London.

Dieter Kastovsky. 2006. Typological changes in derivational morphology. In Ans van Kemenade and Bettelou Los, editors, *The Handbook of the History of English*, pages 151–176. Blackwell Publishing.

Max Kisselew, Laura Rimell, Alexis Palmer, and Sebastian Padó. 2016. Predicting the direction of derivation in English conversion. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–98.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Geoffrey Neil Leech. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research*, 28:1–13.

Stela Manova. 2011. *Understanding Morphological Rules*. Springer.

Hans Marchand. 1964. A set of criteria of derivational relationship between words unmarked by derivational morphemes. *Indogermanische Forschungen*, 69:10–19.

Hans Marchand. 1969. *The Categories and Types of Present-day English Word-formation*. Beck, München.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Maria Polinsky. 2012. Headedness, again. *Theories of Everything. In Honor of Ed Keenan. Los Angeles: UCLA Department of Linguistics*, pages 348–359.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Pavol Štekauer, Salvador Valera, and Lívia Körtvélyessy. 2012. *Word-Formation in the Worlds Languages*. Cambridge University Press.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290.

Salvador Valera and Alba E. Ruz. 2020. Conversion in English: homonymy, polysemy and paronymy. *English Language and Linguistics*, pages 1–24.

Eva Van Lier and Jan Rijkhoff, editors. 2013. *Flexible word classes: a typological study of underspecified parts-of-speech*. Oxford University Press Oxford.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petra Vogel and Bernard Comrie, editors. 2000. *Approaches to the Typology of Word Classes*. Mouton de Gruyter, Berlin and New York.

Arnfinn Muruvik Vonen. 1994. Multifunctionality and morphology in Tokelau and English. *Nordic Journal of Linguistic*, 17:155–178.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *Conference on Natural Language Processing (KONVENS) 2019*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Word	$|N|$	$|V|$	Sim	Word	$|N|$	$|V|$	Sim	Word	$|N|$	$|V|$	Sim
aim	137	98	2.0	change	889	858	1.6	force	470	188	0.8
answer	480	335	2.0	claim	222	239	1.6	grant	108	87	0.8
attempt	302	214	2.0	cut	92	488	1.6	note	287	361	0.8
care	403	249	2.0	demand	169	142	1.6	sense	536	88	0.8
control	519	179	2.0	design	246	153	1.6	tear	124	89	0.8
cost	234	192	2.0	experience	522	150	1.6	account	337	122	0.6
count	143	220	2.0	hope	114	571	1.6	act	644	268	0.6
damage	270	82	2.0	increase	252	399	1.6	back	764	88	0.6
dance	81	97	2.0	judge	80	96	1.6	face	1185	281	0.6
doubt	261	132	2.0	limit	125	134	1.6	hold	130	1251	0.6
drink	456	315	2.0	load	230	87	1.6	land	393	123	0.6
end	1171	244	2.0	offer	93	489	1.6	lift	100	165	0.6
escape	95	111	2.0	rise	164	283	1.6	matter	572	294	0.6
estimate	96	118	2.0	smoke	128	100	1.6	order	841	133	0.6
fear	209	99	2.0	start	159	1269	1.6	place	1643	341	0.6
glance	101	161	2.0	step	401	167	1.6	press	130	188	0.6
help	200	897	2.0	study	1037	211	1.6	roll	135	201	0.6
influence	204	150	2.0	support	290	292	1.6	sort	1613	216	0.6
lack	194	107	2.0	trust	90	126	1.6	fire	444	89	0.4
link	147	176	2.0	waste	103	98	1.6	form	1272	354	0.4
love	495	573	2.0	work	1665	1593	1.6	notice	115	387	0.4
move	131	1272	2.0	base	109	378	1.4	play	185	1093	0.4
name	960	112	2.0	cover	137	399	1.4	turn	226	1566	0.4
need	587	2350	2.0	plant	591	82	1.4	wave	402	120	0.4
phone	382	238	2.0	run	152	999	1.4	cross	102	215	0.2
plan	321	161	2.0	stress	159	106	1.4	deal	191	315	0.2
question	1285	96	2.0	approach	409	175	1.2	hand	1765	127	0.2
rain	182	92	2.0	cause	237	530	1.2	present	219	353	0.2
result	752	206	2.0	match	110	123	1.2	set	387	652	0.2
return	138	441	2.0	miss	320	410	1.2	share	104	232	0.2
search	215	163	2.0	process	720	91	1.2	sign	284	121	0.2
sleep	171	291	2.0	shift	96	104	1.2	suit	162	108	0.2
smell	141	149	2.0	show	132	1843	1.2	wind	189	82	0.2
smile	211	422	2.0	sound	313	496	1.2	address	257	148	0.0
talk	119	1302	2.0	dress	191	196	1.0	bear	110	394	0.0
use	791	2801	2.0	lead	107	716	1.0	head	1355	96	0.0
view	811	102	2.0	light	669	124	1.0	mind	736	620	0.0
visit	136	203	2.0	look	699	5893	1.0	park	179	105	0.0
vote	124	93	2.0	mark	562	198	1.0	point	1534	469	0.0
walk	144	914	2.0	measure	226	223	1.0	ring	185	387	0.0
dream	254	107	1.8	rest	414	132	1.0	square	225	82	0.0
record	1057	276	1.8	tie	82	112	1.0	state	471	156	0.0
report	313	331	1.8	break	117	519	0.8	stick	109	294	0.0
test	273	126	1.8	charge	392	115	0.8	store	95	158	0.0
touch	145	271	1.8	drive	88	476	0.8	train	224	94	0.0
call	209	1558	1.6	focus	92	168	0.8	watch	119	940	0.0

Table 4: 138 flexible words in English (top in BNC corpus) and human similarity scores, average of 5 ratings.