CityLex: a free multi-source English digital lexicon

1 Introduction

Lexical databases consist of word lists paired with information such as morphosyntactic annotations, pronunciation transcriptions, word frequencies, and/or aggregated human ratings for properties such as age of acquisition, arousal, concreteness, dominance, familiarity, imageability, or valence [1–5]. Such databases are commonly used for stimulus design in behavioral psycholinguistics. They may also provide important covariates for propensity score matching, statistical analysis, and computational cognitive modeling. These databases are also used to develop speech and language technologies. Systems for part-of-speech tagging, syntactic parsing, text simplification, automatic speech recognition, text-to-speech synthesis, and machine translation, for example, exploit features found found in lexical databases.

2 Proposal

The resources needed to build high-quality lexical databases—morphological analyzers, digital pronunciation dictionaries, and text corpora used to compute frequency norms—are all freely available for dozens of languages, and researchers with experience in computing can convert these resources into a database without great difficulty. However, lexical databases are available for only a handful of the world's languages, and of these, only a few are freely available to the public. We are only aware of one free English lexical database: the MRC Psycholinguistic Database [6, 7] published in 1981. Furthermore, virtually all lexical databases, whether free or proprietary, are "static" in the sense that they are distributed as pre-generated spreadsheets, and rarely ever corrected, updated, or otherwise improved. Yet such improvements are often called for. For instance, Steiner [8] provides software to automatically correct deficiencies in the German morphosyntactic annotations of the proprietary CELEX database [9]. Pronunciations may change; for instance, older American English speakers tend to pronounce the word *ration* so that it rhymes with *nation*, whereas younger speakers tend to rhyme it with *fashion*. Word frequencies may change rapidly as well, often in response to cultural change [e.g., 10]; for instance, the word *slavery* greatly increased in print frequency during the American Civil War and later civil rights struggles [11, 12]. Such changes may have consequences for psycholinguists; for instance, some studies find that frequency norms collected from older and smaller corpora are poorer predictors of speakers' familiarity with words than newer, larger resources [13, 14]. While one cannot place a "expiration date" on lexical databases, the inevitability of language change ensures any static resource will become obsolete.

We propose to refine, release, and evaluate a free lexical database for English called CityLex. CityLex is compiled from both free and proprietary resources providing morphosyntactic analyses, phonemic transcriptions, and word frequencies. It is also "dynamic" in the sense that it includes not only the database, but software which can be used to (re)generate the database as the resources it is derived from are improved. This design allows CityLex to be free for all users and

uses, while exploiting both free resources whose licensing terms forbid redistribution of derivatives as well as specific proprietary resources for users who hold the relevant licenses.

3 Design

CityLex is built by extracting the aforementioned lexical features from various proprietary and free resources, and then merging them into a single schematic database using the outer join operation. CityLex is therefore constructed from the union—rather than the intersection—of multiple resource, and may contain multiple pronunciations, morphosyntactic analyses, or frequencies for a given entry. This is in contrast to proprietary databases, some of which have complete and unique records for each word (i.e., its morphosyntactic analysis, its pronunciation, and its frequency). At the time of writing, the "alpha" version of CityLex provides partial or full lexical entries for over 326,000 words, roughly twice as many entries as found in the largest proprietary resources for English. Since CityLex contains multiple sources of truth for certain types of annotations, efforts are made to provide *translations* between annotation sources; some of these translations are described below. CityLex uses a SQL database as its back-end, but will also provide the ability to export data to tab-separated values (TSV) files. Such files can be read by spreadsheet applications such as Microsoft Excel and Google Sheets, and by statistical computing environments such as R or SAS. Rather than requiring users to install any local software, CityLex will also be accessible via a web browser. Using this interface, users will be able to download a current CityLex database or to issue simple queries for words with particular linguistic properties.

4 Data sources

The following major sources are currently included in CityLex.[1]

4.1 Morphosyntactic annotations

Morphosyntactic annotations potentially include a wide variety of information ranging from the word's lemma (i.e., citation form), morphological tags (i.e., feature bundles indicating the word's part of speech and the presence of properties like plurality or past tense), and segmentations into roots and affixes. For users who hold a license, the CELEX lexicon can be used to the lemma and morphosyntactic tags for each word. CityLex provides similar, but free, morphosyntactic annotations extracted from UDLexicons [15] and UniMorph [16], and word segmentations extracted from the English Lexicon project [17]. Morphosyntactic features are automatically translated from other sources into the UniMorph tagset [18].

4.2 Pronunciations

Pronunciations are extracted from CELEX (where available), the CMU Pronouncing Dictionary, NETTalk [19], the proprietary CALLHOME lexicon (where available), and both British and American versions of WikiPron [20]. Pronunciations will be provided in the International Phonetic Alpha-

[1]This list is not exhaustive, and other sources can be quickly incorporated as desired.

```
elp_morph_sp: "{god}{father}"
udlexicons_morph { lemma: "godfather" features: "N;SG" }
unimorph_morph { lemma: "godfather" features: "V;NFIN" }
wikipron_uk_pron: "g ɒ d f ɑː ð ə"
wikipron_us_pron: "g ɑ d f ɑ ð ɚ"
subtlex_uk_freq: 486
subtlex_us_freq: 278
```

Figure 1: The CityLex entry for the word *godfather*, with some features and translations omitted for space.

bet (IPA) and X-SAMPA, a deterministic ASCII encoding of the IPA. Supervised neural sequence-to-sequence models similar to those used for grapheme-to-phoneme conversion will be used to convert non-IPA transcriptions to the IPA, though observed and automatically generated pronunciations are distinguished in the resulting metadata.

4.3 Frequencies

Frequency norms are extracted from CELEX (where available), SUBTLEX-UK [21], and SUBTLEX-US [22]. Frequencies will be provided as raw frequencies, and as two scaled measurements: frequency per million words (fpmw) and "Zipf scales" [21].

5 Preliminary results

Figure 1 provides a sample CityLex entry, with several different sources represented.

Table 1 gives the number of entries obtained from each of the aforementioned sources, including CELEX. While the widely-used CELEX data is a substantial resource in all three categories, it is not the largest resource in any one category: the UniMorph morphosyntactic annotations, CMUDict pronunciations, and the two SUBTLEX frequency norms each exceed the size of CELEX in their respective category. Table 2 reports lexical overlap between CELEX and the various non-proprietary resources used by CELEX. Whereas there is a large overlap between CELEX and other resources, these resources may contain information for many thousands of words not found in CELEX. For example, there are over 90,000 words not found in CELEX but which are provided a morphosyntactic analysis by at least one free data sources.

6 Approach

An "alpha" version of CityLex has been prepared by the PI. Funds are requested as wages for graduate RAs, for open-access fees, and for web hosting. The RAs will implement data ingestion, translations, the database back-end, and the web front-end. Best practices in software development, including continuous integration testing, will be used. It is anticipated that the RAs will gain valuable experience in web development and resource curation.

	# entries
Morphosyntactic annotations	
CELEX	89,059
ELP	68,623
UDLexicons	66,976
UniMorph	115,523
Pronunciations	
CELEX	73,351
CMUDict	133,852
WikiPron-UK	52,995
WikiPron-US	49,132
Frequencies	
CELEX	72,628
SUBTLEX-UK	160,022
SUBTLEX-US	74,286

Table 1: The number of CityLex entries obtained from each data source.

7 Outcomes

Three major outcomes are anticipated:

- The creation of a "production release" of the CityLex library, licensed for public use and released via the Python Package Index (PyPI).
- The launch of a web front-end allowing CityLex databases to be generated and queried in a web browser (without any software being installed).
- The preparation of a manuscript for submission to the *Behavior Research Methods* journal.

If time and funding allows, "stretch" outcome will include the addition of human ratings like valence, and a study of automatic imputation for missing values [e.g., 23].

References

[1] H. Stadthagen-Gonzalez and C. J. Davis. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods* 38.4 (2006). [2] M. J. Cortese and J. Schock. Imageability and age of acquisition effects in disyllabic word recognition. *Quarterly Journal of Experimental Psychology* 66.5 (2008). [3] J. Schock et al. Age of acquisition estimates for 3,000 disyllabic words. *Behavior Research Methods* 44.4 (2012). [4] A. B. Warriner et al. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45.4 (2013). [5] G. C. Scott et al. The Glasgow norms: ratings of 5,500 words on nine scales. *Behavior Research Methods* 51.3 (2019). [6] M. Coltheart. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology Section A* 33.4 (1981). [7] M. Wilson. The MRC Psycholinguistic Database: machine readable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* 20.1 (1988). [8] P. Steiner.

	$\lvert C \cap L \rvert$	$\lvert C - L \rvert$	$\lvert L - C \rvert$
Morphosyntactic analyses			
ELP	61,075	10,948	7,548
UDLexicons	33,928	38,095	22,688
UniMorph	27,775	44,248	63,856
Pronunciations			
CMUDict	41,233	31,395	83,840
WikiPron-UK	23,688	48,940	23,688
WikiPron-US	22,803	19,051	49,825
Frequencies			
SUBTLEX-UK	47,539	7,240	112,483
SUBTLEX-US	43,977	10,802	30,309

Table 2: Overlap between CELEX (C) and CityLex (L) sources; C: CELEX; $|C \cap L|$: number of shared items; $|C - L|$: number of items present in CELEX but not in L; $|L - C|$: number of items present in L but not in CELEX.

Refurbishing a morphological database for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2016. [9] R. H. Baayen et al. CELEX2. LDC96L14. 1996. [10] J.-B. Michel et al. Quantitative analysis of culture using millions of digitized books. *Science* 331 (2011). [11] W. L. Hamilton et al. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016. [12] A. Rosenfeld and K. Erk. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018. [13] J. D. Zevin and M. S. Seidenberg. Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language* 47.1 (2002). [14] D. A. Balota et al. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General* 133.2 (2004). [15] B. Sagot. A multilingual collection of CoNLL-U-compatible morphological lexicons. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. 2018. [16] C. Kirov et al. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2016. [17] D. A. Balota et al. The English Lexicon Project. *Behavior Research Methods* 39.3 (2007). [18] A. D. McCarthy et al. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. 2018. [19] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems* 1 (1987). [20] J. L. Lee et al. Massively multilingual pronunciation mining with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020. [21] W. J. B. van Heuven et al. SUBTLEX-UK: a new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* 67.6 (2014). [22] M. Brysbaert and B. New. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41.4 (2009). [23] G. H. Paetzold and L. Specia. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.