Orthographic databases and lexicons: introduction to the Special Issue

Lynne Cahill and Terry Joyce

The availability of linguistic databases for a variety of languages and a variety of linguistic levels is enabling for a number of areas of research. Within writing system and literacy research, databases of orthographic information *per se* as well as a variety of forms of phonological information at least, and potentially morphological, syntactic and semantic information may all be useful. The papers within this special issue address a range of databases for different languages and with a range of different types of information. They talk about both the development and uses of such databases.

Databases can allow researchers to easily and reliably account for a range of variables that may be important in experimental work on writing systems and literacy. The list of such variables is long, as discussed in Balota, Yap, Hutchison & Cortese (2012) and Yap & Balota (2015). Balota et al. list "word frequency, familiarity, age of acquisition, imageability, number of meanings, letter length, phoneme length, syllable length, number of morphemes, syntactic class, orthographic neighborhood, phonological neighborhood, frequency of orthographic and phonological neighborhoods, spelling-to-sound consistency, among many others" (2012: p. 90). Making available databases with information about a large range of such variables therefore aids experimental and theoretical research in these areas.

They can be used to provide large samples of data in order to develop theories about the relationship between (aspects of) pronunciation and spelling in one or more languages. They can also be the basis for the lexicons that can be used in NLP applications. Increasingly, databases and lexicons for such applications may be linked to ontologies to permit greater interoperability and consistency across applications. Huang, Calzolari, Gangemi, Lenci, Oltramari & Prévot (2010) and Oltramari, Vossen, Qin & Hovy (2013) discuss examples of work that combines ontologies and lexicons. They may also be the result of theoretical developments, where theories of interactions between levels of representation can be tested on large lexicons that incorporate a range of representations.

Technological developments allow for increasingly large and varied databases to be developed. This work is important to ensure that the most up-to-date resources are available for researchers. Balota et al (2012:96) discuss the problems of relying on old databases which may have been based on

smaller, less representative and out of date corpora. Specifically, they lament the reliance on Kučera and Francis (1967) norms, when others are available such as HAL (Hyperspace Analog to Language; Burgess & Livesay 1998), CELEX (Center for Lexical Information; Baayen, Piepenbrock, & van Rijn 1993), TASA (Touchstone Applied Science Associates; Zeno, Ivens, Millard, & Duvvuri 1995), and BNC (British National Corpus; Leech, Rayson, & Wilson, 2001). They compare results of experiments based on a selection of databases as evidence of the importance of using updated resources.

What do we mean by databases and lexicons and how do we distinguish the two? Nerbonne (1998) defines linguistic databases as having two crucial properties: they must be declarative and they must be consistent. By declarative he means that they must not be dependent on any particular software, but must be in a computationally neutral format that can be read and processed by any program. Thus we expect databases to be in a format such as ASCII (for maximum usability) or UNICODE. The question of consistency can mean different things for different linguistic levels. For orthographic databases this means that the spellings provided must reflect a consistent spelling system represented in a consistent way. Thus a decision to use either British or American spellings would be crucial for English and the decision about how to represent accents or diacritics, especially within an ASCII format database would be crucial for many writing systems.

Lexicons, on the other hand, may not follow these criteria. They may be developed as part of larger systems to process speech or text and therefore need to fulfil the specific requirements of those systems. Lexicons might also be developed in order to test theories of the representation of various levels of lexical information. The precise status of the lexicon in linguistic theory has varied from the early Chomskyan view of it as essentially a list of irregularities that cannot be accounted for by means of rules to the radical lexicalist view of the repository of most, if not all, linguistic information about words, their meanings, forms and even how they combine. It should be noted that no restriction was placed on the interpretation of these notions of “database” and “lexicon” and not all authors conform exactly to this distinction.

The Ninth Workshop on Written Language and Literacy was held at the University of Sussex in Brighton (UK) in September 2014. The event was attended by people from eight different countries on four different continents. There was an invited talk by Professor Viorica Marian from Northwestern University and around 15 other oral and poster presentations, covering a range of topics, languages and writing systems. Around half of the papers related to the workshop theme of “Orthographic Databases and Lexicons”. This special issue contains four papers initially presented at the workshop together with a further paper, relating to the theme, that was submitted after the workshop.

The workshop invited papers that addressed questions relating to the development and the use of databases and lexicons. The papers on the theme presented at the workshop covered databases of English, Dutch, German, Polish, Spanish, French, Japanese and Kabiye in a range of mono- and multi-lingual approaches. The five papers in this special issue address a similarly diverse range of languages and approaches.

The first paper, by Viorica Marian, *Orthographic and Phonological Neighborhood Databases across Multiple Languages*, discusses possible uses for the CLEARPOND databases. These databases (introduced in Marian, Bartolotti, Chabal & Shook (2012)) provide information about the orthographic and phonological neighbourhoods for English, French, Spanish, German and Dutch. The databases allow linguists to investigate neighbourhood effects both within and across languages and modalities. For example, is an English-French bilingual affected by phonological neighbours in French when reading or accessing a word in English? In order to investigate this, experimenters can use the CLEARPOND databases to find phonological neighbours in French for the English word(s) in question. A closer look at neighbourhood effects on lexical access reveals that not only orthographic, but also phonological neighbourhoods can influence visual lexical access both within and across languages.

The second paper, *Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system*, by Terry Joyce, Bor Hodošček and Hisashi Masuda describes a key part of an enterprise to develop a database of Japanese Lexical Properties. The part described in this paper is the development of the ontology at its core. The complex nature of the Japanese writing system makes the development of a reference database both very important and rather challenging. An ontology that encompasses the many and varied features of Japanese writing is likely to be applicable to many if not all other writing systems.

Johan Zuidema and Anneke Neijt's paper, *The BasisSpellingBank – a spelling database with knowledge stored as a lexicon of triplets*, reports on the development of a database with a novel theoretical approach to the representation of orthographic information for Dutch. Their triplet approach involves representing the relationship between phonemes and graphemes by means of triplets which each consist of a phoneme, a grapheme and a rule specifying how the relationship is defined, for example as a default or because of occurrence in a non-native word. The lexicon then specifies for each word the triplets that make up the spelling and its relationship to the pronunciation. In this paper the terms "lexicon" and "database" are used broadly interchangeably, illustrating the blurring of the boundaries between the two. The BSB could be seen as a database in that it fulfils Nerbonne's two requirements. However, it also incorporates the specific theory

underlying the triplet approach to the representation of linguistic (and specifically spelling) information in a way that is consistent with the idea of a lexicon.

The paper *STRESYL: An Italian Stress-in-Syllables database for reading research*, by Simone Sulpizio Giacomo Spinelli and Cristina Burani describes a database of Italian which focuses on the specific feature of stress in syllables. This is an example of a database which does not specifically include orthographic information, but which can be an important resource in research on orthography and reading. The importance of syllable and stress information in reading has been noted in studies such as Ferrand & New (2003) and Columbo & Sulpizio (2014) and the STRESYL database for Italian provides a useful resource for further research in this area.

Finally, the paper by Lynne Cahill, *What are the "phonemes" in phoneme-grapheme mappings? A perspective on the use of databases for lexicon development* takes a broader view, discussing the precise nature of phonological (and specifically "phonemic") representations available in databases that promise to be useful for research into phoneme-grapheme mappings. There are numerous databases that appear to include this level as well as the orthographic level of information, but a closer inspection reveals that there is no clear agreement about the exact level of "underlying" phonological representation and there are practical issues in obtaining this representation for large databases. These issues are illustrated by means of a case study examining the problems that arose when the CELEX lexical databases were used to develop lexicons of English, Dutch and German.

This collection of papers illustrates the varied types of database available and a range of ways in which they can be used. The increase in the availability of large sources of linguistic data, both in the form of corpora and other electronically available sources, promises to provide many more opportunities for researchers in writing systems and literacy research, as well as those working in linguistics, psycholinguistics and computational linguistics.

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database*. Philadelphia, PA : Linguistic Data Consortium, University of Pennsylvania.

Balota, David A., Melvin J. Yap, Keith A. Hutchison & Michael J. Cortese (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing?. In James S. Adelman (ed.) *Visual Word Recognition Volume 1: Models and methods, orthography and phonology* London: Taylor and Francis, pp. 90-115.

Burgess, C. & Livesay, K. (1998). The effect of corpus size in predicting RT in a basic word recognition task: Moving on from Kucera and Francis. *Behavior ResearchMethods, Instruments, & Computers*, *30*, 272‑277.

Colombo, L., & Sulpizio, S. (2015). When orthography is not enough: The effect of lexical stress in lexical decision. *Memory & Cognition, 43*, 811-824.

Ferrand, L., & New, B. (2003). Syllabic length effects in visual word recognition and naming. *Acta Psychologica*, *113*, 167-183.

Huang, Chu-Ren, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari & Laurent Prévot (Eds.). (2010). Ontology and the lexicon: A natural language processing perspective. (Studies in Natural Language Processing). Cambridge: Cambridge University Press.

Kučera, H. & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI:Brown University Press.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London:Longman.Marian, Viorica, James Bartolotti, Sarah Chabal & Anthony Shook (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. PloS one 7(8): e43230.

Nerbonne, John (1998). *Linguistic Databases.* CSLI (ISBN: 9781575860930)

Oltramari, Alessandro, Piek Vossen, Lu Qin & Hovy, Eduard. (2013). New trends of research in ontologies and lexical resources: Ideas, projects, systems. Springer.

Yap, Melvin J. & David A. Balota. (2015). Visual word recognition. In Alexander Pollatsek & Rebecca Treiman (Eds.), *The Oxford Handbook of Reading* (pp. 26-43). New York: Oxford University Press.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R . (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.