Redundancy and Productivity in the Speech Technology Lexicon - Can We Do Better?

Susan Fitt and Korin Richmond

Centre for Speech Technology Research
University of Edinburgh, UK
sue@inf.ed.ac.uk, korin@inf.ed.ac.uk

Abstract

Current lexica for speech technology typically contain much redundancy, while omitting useful information. A comparison with lexica in other media and for other purposes is instructive, as it highlights some features we may borrow for text-to-speech and speech recognition lexica.

We describe some aspects of the new lexicon we are producing, Combilex, whose structure and implementation is specifically designed to reduce redundancy and improve the representation of productive elements of English. Most importantly, many English words are predictable derivations of baseforms, or compounds. Storing the lexicon as a combination of baseforms and derivational rules speeds up lexicon development, and improves coverage and maintainability.

Index Terms: dictionary, lexicon, pronunciation, English accents, productivity, derivation, redundancy, relational database

1. Introduction

"The lexicon is like a prison – it contains only the lawless, and the only thing that its inmates have in common is lawlessness." [1] p3. This is a neat summary of what we need from a lexicon; however, it does not correspond to what we find in lexica for text-to-speech (TTS) and automatic speech recognition (ASR). While printed dictionaries organise words by lexeme, and may show derivations only partially specified (e.g. omitting parts of the orthography or pronunciation), speech technology pronunciation lexica have traditionally been lists of fully-specified entries, with no explicit relationship between words belonging to the same lexeme, and much redundant information.

This paper will examine some of the differences in structure between different lexica, and the advantages and disadvantages of these structures for speech technology lexica. We will then outline the structure being created for our new speech technology lexicon, Combilex, and show how this has advantages over other TTS and ASR dictionaries.

2. Comparison of lexica

We will compare the structure of a selection of typical dictionaries by summarising entries for a single word, *clos*e, focusing on the representation of pronunciation. We will first look at pronunciation dictionaries for speech technology, and then compare these to printed and on-line dictionaries.

These different media have traditionally had different functions. Most TTS/ASR lexica contain limited information. They are really only needed for pronunciations, and at their simplest this is all they are, a list of headwords and associated pronunciations, with different pronunciations given for homographs which are not also homophones (e.g. *record* as a noun and *record* as a verb). More detailed lexica may contain free variants, and also parts of speech, and some systems, such as Angie [2] contain sub-word information such as phonological structure.

Printed (monolingual) dictionaries have been used mainly to describe the semantics of words, and often their usage and etymology. Pronunciation is included in most, but not all, such dictionaries. Homographs are given different entries, whether or not they have different pronunciations; as we will see though, one orthography can be split in several different ways. Links between words are shown by including derived words, and sometimes compounds, under the headword, and by including pointers to related entries.

Web-based dictionaries are able to include all this and more, for example by providing direct links to related words. They do not have the space restrictions of printed material, and because they also have search facilities, it is possible for the user to perform searches on partial strings and come up with a series of matches, or even suggested corrections for misspellings.

In our examples below, pronunciations are given in square brackets using each dictionary's symbol set; some of these are UK English, some are US, and some show both. The abbreviation *sem.* means semantic information is given in the dictionary, but this information is not detailed here, and example usage and etymology are omitted.

2.1. Speech technology lexica

2.1.1. CMU

Although there exist much richer speech technology lexica (e.g. ANGIE [2] and CELEX [3]), CMU [4] is still widely used as it is free, so it is worth looking at here.

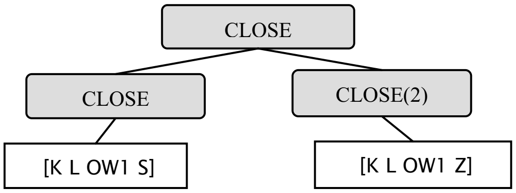

Figure 1: *Representation of* close*: CMU pronouncing dictionary*

This is the simplest possible structure for a pronunciation lexicon. Only pronunciations are listed, with no part of speech, semantic or other differentiating information. We have no way of knowing which variant to use in what circumstances, nor whether these are conditioned variants (dependent on part of speech or semantics, as is the case for *close*) or free variants (dependent only on speaker choice, as is the case for *economic*, which is given the same structure as *close*). Derivations, inflections and compounds are all listed as separate headwords, and collocations and phrases are not listed at all.

2.1.2. Unisyn

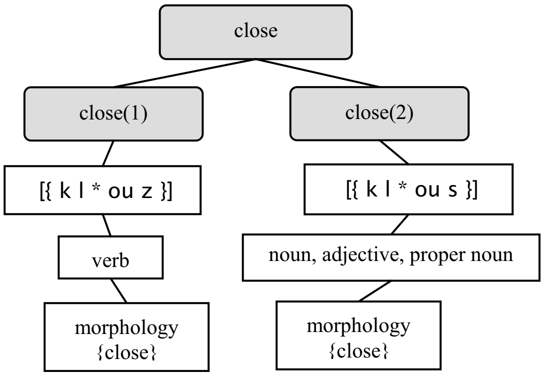

Figure 2: *Representation of* close*: Unisyn multi-accent speech technology lexicon*

In Unisyn [5], the division into entries is made by pronunciation. Part of speech is given, and where this does not differentiate homographs, semantic information is also given. Like CMU, this lexicon does not explicitly differentiate between conditioned variants and free variants, although this information can be inferred by comparing the part-of-speech and semantic fields: if for a given spelling these two fields differ, the headwords are conditioned variants; if they are the same, the headwords are free variants.

There are no explicit links to derivations, compounds etc., though the morphological field is used in conjunction with the morpheme boundaries {} shown on the pronunciation to implicitly link related entries such as *closer, close-up*. (Note that the part of speech *noun*, meaning *the finish*, is missing here from *close*(1)).

2.2. Printed lexica

2.2.1. The Chambers Dictionary

Chambers [6] is a traditional UK-English dictionary with a rich structure and detailed information for each entry. However, the pronunciations of derived words, compounds etc. are usually underspecified, with only the stress being shown (see Figure 1). For words whose derivations are unpredictably pronounced, or cannot easily be guessed from the pronunciation of the headword (e.g. *mice*, listed under *mouse*), a full pronunciation is shown; for the rest, native speaker knowledge is needed to ascertain the complete pronunciation. Although suffixed words (e.g. *closure*) are listed within the headword entry, prefixed words (e.g. *reclose*) are listed alphabetically under the prefix.

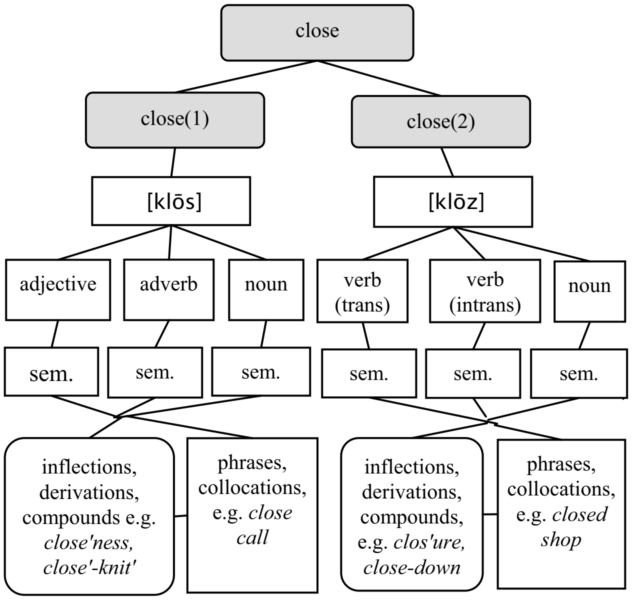
close
close(1)
close(2)
[klōs]
[klōz]
adjective
adverb
noun
verb (trans)
verb (intrans)
noun
sem.
sem.
sem.
sem.
sem.
sem.
inflections, derivations, compounds e.g. close'ness, close'-knit'
phrases, collocations, e.g. close call
inflections, derivations, compounds, e.g. clos'ure, close-down
phrases, collocations, e.g. closed shop

Figure 3: *Representation of* close*: Chambers*

2.2.2. Longman Pronunciation Dictionary

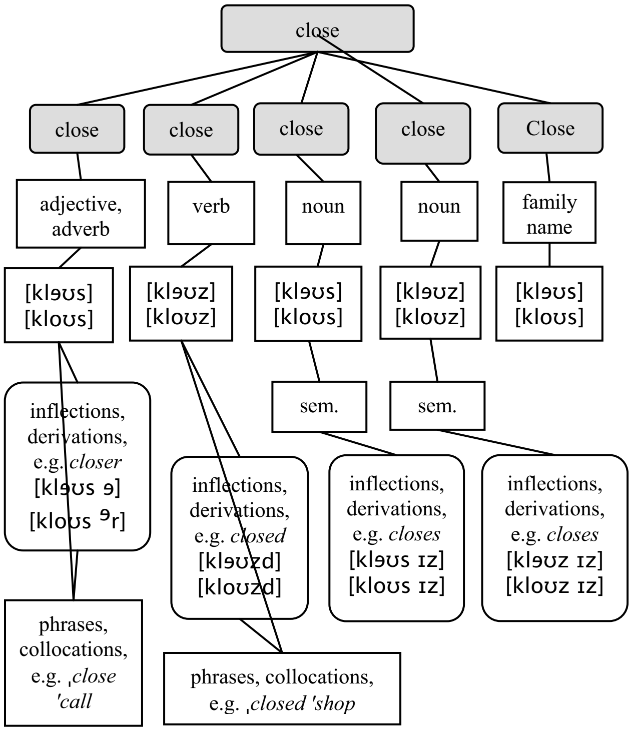
close
close
close
close
close
Close
adjective, adverb
verb
noun
noun
family name
[kləʊs] [kloʊs]
[kləʊz] [kloʊz]
[kləʊs] [kloʊs]
[kləʊz] [kloʊz]
[kləʊs] [kloʊs]
sem.
sem.
inflections, derivations, e.g. closer [kləʊs ə] [kloʊs ᵊr]
inflections, derivations, e.g. closed [kləʊzd] [kloʊzd]
inflections, derivations, e.g. closes [kləʊs ɪz] [kloʊs ɪz]
inflections, derivations, e.g. closes [kləʊz ɪz] [kloʊz ɪz]
phrases, collocations, e.g. ˌclose 'call
phrases, collocations, e.g. ˌclosed 'shop

Figure 4: *Representation of* close*: Longman*

The Longman Pronunciation Dictionary [7], which covers UK and US pronunciations, has a very different type of structure from Chambers, with five entries to Chambers' two. One of these is a personal proper noun, a type generally not included in Chambers; the other four are reorganisations of Chambers' two. Chambers has split these two headwords by pronunciation, although the meanings are related. Longman generally follows this rule, but in this case we have two identical parts of speech (noun) which have different pronunciations, and these have been given separate entries with semantic information. Semantics is generally omitted from Longman unless it differentiates entries. Inflections and derivations are mostly given fully-specified pronunciations, while phrases are given only stress marks. This lexicon is aimed in part at non-native speakers, who may not understand all the affix pronunciation rules which native speakers can be expected to know.

While inflections, derivations and phrases are listed under these headwords, compounds such as *close-knit* are given separate, fully specified entries. Like Chambers, prefixed words such as *reclose* are listed under the prefix.

2.3. Web dictionaries

HTML with hyperlinks and facilities such as electronic searches offers the richest structural possibilities.

2.3.1. Merriam-Webster OnLine

We are not in a position here to analyse the structures which produce the Merriam-Webster website, only the resulting webpage [8]. A search on this page for *close* produces the following.

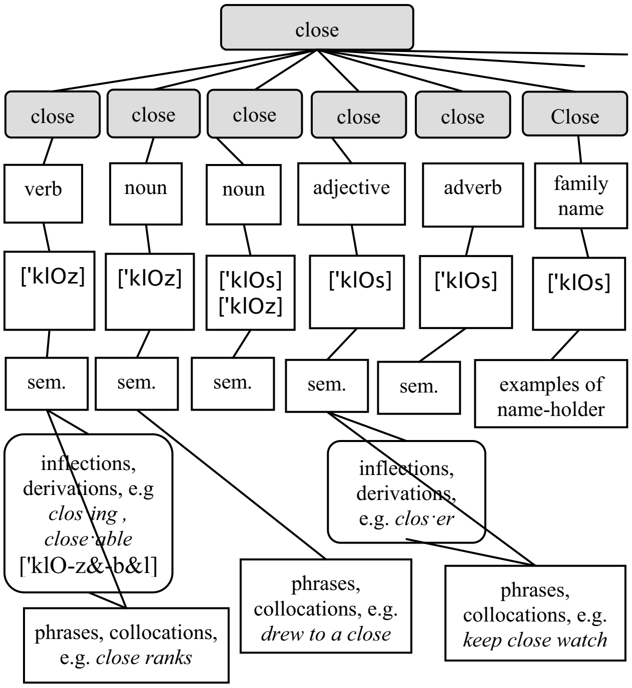
close
close
close
close
close
close
Close
verb
noun
noun
adjective
adverb
family name
['klOz]
['klOz]
['klOs] ['klOz]
['klOs]
['klOs]
['klOs]
sem.
sem.
sem.
sem.
sem.
examples of name-holder
inflections, derivations, e.g clos·ing , close·able ['klO-z&-b&l]
inflections, derivations, e.g. clos·er
phrases, collocations, e.g. close ranks
phrases, collocations, e.g. drew to a close
phrases, collocations, e.g. keep close watch

Figure 5: *Representation of* close*: Merriam-Webster OnLine*

In addition to the above information, the online Merriam-Webster contains synonyms, links to audio files for listening to fully-specified pronunciations, further phrasal entries listed directly under *close*, and several other features.

As for pronunciation, some inflections and derivations are merely given syllabification, e.g. *clos·ing*; some are given full transcriptions with links to audio files. The nouns are lacking any plurals, possibly under the assumption that speakers can construct these themselves. Like Longman, compounds are listed as separate entries, with full pronunciation. Merriam-Webster gives US pronunciations, but notes the UK pronunciation where this is unpredictably different.

3. Discussion of desirable lexical features

A lexicon for speech technology should contain, at a minimum:

- orthographies
- pronunciations for these orthographies
- part of speech associated with these pronunciations, to be used for disambiguation
- semantic information where part of speech is insufficient to distinguish homographs.

However, this specification allows for a lot of redundancy, as is shown in CMU and Unisyn. Chambers and Merriam-Webster give full pronunciations for headwords, but assume that readers can perform obvious inflections and compounding themselves. Longman's, whose focus is pronunciation, does not assume this, and includes this (to native-speakers) redundant information.

Space and search time are not issues for modern speech technology lexica; they may contain large numbers of fully-specified entries without loss of performance. However, for the lexicon maintainer, a fully-specified lexicon is a hindrance. If, for example, we find a new compound such as *close-work* that we wish to add to the lexicon, we must determine which version of *close* to use, look up also the pronunciation given for *work* if we are unsure of the symbols used, and add stress, part of speech, and any other necessary information.

Surely it would be useful if the speech technology lexicon contained more of the knowledge that we expect from native-speakers and is implied by the underspecification in Chambers and Merriam-Webster, so that it could do some of this work for us. If we were adding the part of speech *noun* under *close*(1) in Unisyn, it would be nice if the lexicon would create for us a corresponding plural noun, and a pronunciation for it.

To illustrate the potential benefits, let us examine more closely the types of words found in English lexica. The Unisyn lexicon has morphological markings, so we can easily arrive at an approximation of the percentage of different types of words, for example free roots, inflected words, compounds and so on. This analysis is shown in Figure 6, where we can see that a very large number of entries are potentially derivable from other entries. Free roots and proper names taken together comprise only 24% of entries. A full 63% consist of inflections and derivations (with one or more prefixes or suffixes), while simple compounds (compounds made of free roots, e.g. *close-up*) make up 6%, and complex compounds (which include inflections or derivations, e.g. *close-ups, hummingbird*) make up a further 8%. We have not examined variant spellings here, such as *analyse-analyze*; these will be a further category of derivable words, with some examples taken from all other categories.

Of course, some of these derivable words will have pronunciations which are not predictable from the respective pronunciations of the roots and affixes, but many are predictable. As an example, 12301 entries, or 10.4% of the

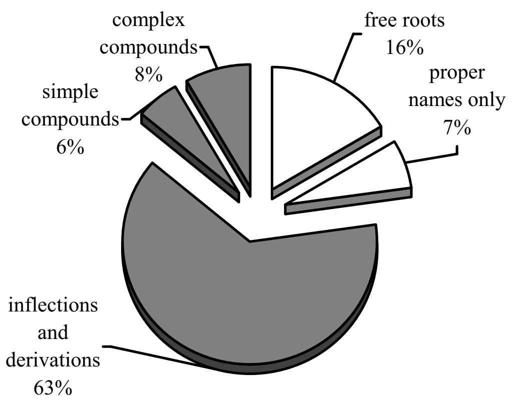

Figure 6: *Different types of word in Unisyn; white segments are underivable, grey segments are potentially derivable. Word total 118,374.*

lexicon, comprise free roots combined with *-s* endings and no other affixes (these are plural nouns or third person singular verbs); very nearly all of these have regular pronunciations. A further 5192 (4.4%) are *-ed* endings (past tense verbs), also overwhelmingly regular.

4. Current work

4.1. Underspecification and productivity

Underspecifying derivable entries and deriving the full pronunciations by rule is highly advantageous in both lexicon-building and lexicon maintenance. Even such information-rich lexica as CELEX [3] hard-code pronunciations for these words.

We are developing a new speech technology lexicon, Combilex, which amongst other features will do just this. Orthographies, pronunciations and other features are stored in a relational database; use of a database structure rather than a flat-text file enables cross-references to be implemented explicitly. Free roots are fully-specified for pronunciation and other features, except where they are simply variant spellings. Likewise, proper names are fully specified unless they have the same orthography and pronunciation as free roots in the dictionary, in which case they are underspecified and a pointer to the relevant free root provides the missing information.

Derived words and compounds are underspecified. There are two possible approaches to this. One is to list the breakdown for each derived word in the lexicon, with pointers at the component parts. The second approach is to omit them from the dictionary altogether, allow a morpheme analyser to break down words not in the dictionary, then look up the component parts and generate an entry. Either option needs compositional rules for dealing with stress and phonology; for example, the rule for *-s* (pronunciation [ɪz]) includes:

[.ɪz] → [z] / [vowel or voiced stop/lateral/nasal
or labiodental fricative] _
[.ɪz] → [s] / [voiceless stop] _
[.ɪz] → [.ɪz] / [affricate or alveolar/palato-
alveolar fricative] _

Close [kləʊs], with a final voiceless alveolar fricative [s], would become [kləʊs.ɪz], while [kləʊz], with a voiced alveolar fricative, would become [kləʊz.ɪz].

The first approach we suggested, listing breakdowns for complex words, requires more lexical entries, though as they are underspecified this should not be onerous. It does give us more control over what is happening in the lexicon; for example although *man* is both a verb and a singular noun, we need to allow *mans* as a third person verb but not a plural noun.

The second option, omitting them from the dictionary, has the advantage of simplicity, and we may in any case want the capacity to perform morphological breakdowns to generate pronunciations for OOV words and random neologisms such as *misunderestimated.* We are therefore planning to investigate this approach, but we will need to block certain derivations, such as *mans* as a plural noun.

4.2. Disambiguation

Another feature novel to speech technology lexica is the representation of collocations, which we are including in Combilex. This will not be extensive, but is provided to aid disambiguation. For example, *terrible* is always pronounced [’tɛ.ɹɪ.bəl] or [’tɛ.ɹə.bəl], except in the phrase *enfant terrible*, where it is [tɛ’ɹiː.blə]. This approach is also useful for multi-word proper names, which in speech technology lexica are usually split into single words, sometimes unhelpfully; for example, *Baton Rouge* (Louisiana, US) is [ˌbæ.tn̩ ˈɹuʒ], although *baton* is generally [bə’tɑn] in US English.

5. Conclusions

We have examined various different approaches to the representation of pronunciation, and shown how avoiding redundancy in our lexicon will also bring productivity. We need therefore imprison in the lexicon only the lawless; we can encode in our system the knowledge that the native speaker brings to a printed dictionary, and use this to set the lawful free.

6. Acknowledgements

We would like to thank Scottish Enterprise, who are funding this work. On completion in early 2007 Combilex will be available under commercial licence.

7. References

[1] Di Sciullo, Anna Maria, and Williams, Edwin (1987). *On the definition of word.* Cambridge, MA: MIT Press.

[2] Seneff, Stephanie (1998). The use of linguistic hierarchies in speech understanding. ICSLP keynote speech.

[3] CELEX. http://www.ru.nl/celex/

[4] CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[5] Unisyn multi-accent lexicon: http://www.cstr.ed.ac.uk/projects/unisyn/

[6] *The Chambers Century Dictionary* (2003). Edinburgh: Chambers Harrap.

[7] Wells, J.C. (2000). *Longman Pronunciation Dictionary.* Harlow: Longman.

[8] Merriam-Webster OnLine. http://www.m-w.com/cgi-bin/dictionary.