The Design of Online Lexicons

Burke, Sean Michael (`sburke@cpan.org`). 1998. *The Design of Online Lexicons*. Master's thesis: Northwestern University, Evanston, IL.

Introduction

This work is an introduction to topics in the design of online lexicons.

While online lexicons have been a technical possibility since the days of the first wide−area computer networks in the 1970s (Cerf and Kahn 1974) and have existed in some form since at least the early 1980s (Unknown ?1983; Curry 1990, 1996; Mayer 1996), it is only with the popularization of the World−Wide Web in the mid−1990s that significant work in producing online lexicons has begun.

This work, first and foremost, is an attempt to apply and extend aspects of lexicographic theory in the light of the possibilities and demands of online media, so that the theories of the past can be put to use in producing better online lexicons. Secondarily, I hope to point out the advantages for lexicography which online media have over print media. In this discussion of online lexicons, I will first introduce the reader to what I mean when I say "lexicon" and "online".

"Lexicons"

By "lexicons" I mean works, *made for use by humans* (although not necessarily exclusively so) which are about words, the main content of which is divided into articles ("entries") each of which is about a word or group of related words.

This formulation of "lexicon" includes:

- standard definitional dictionaries such as Merriam−Webster (1963) or Larousse (1971).
- bilingual dictionaries (although such a work as an English−French/French−English dictionary is in fact two lexicons bound in a single volume),
- thesauruses,
- phonetic dictionaries like rhyming dictionaries or pronouncing dictionaries
- orthographic dictionaries like shorthand dictionaries, secretaries' dictionaries of hard−to−spell words, or crossword puzzle dictionaries. (Although these are unusual in that the entry for a given word consists generally of just the headword itself.)
- More encyclopedic dictionaries like: ethnographic dictionaries (e.g., Franciscan Fathers 1910) or dictionaries of specialized fields of knowledge (e.g., Howe 1994).

I do not address the issue how encyclopedic a work can be and still be a "lexicon".

The reader with a background in Natural Language Processing (NLP) should be aware that my usage of "lexicon" here has nothing to do with the distinction between "lexicon" and "dictionary" in NLP, as explained in *Electric Words* (Wilks, Slator, and Guthrie 1996:6):

> In this book we continue with the now conventional usage of "lexicon" to mean a set of formalized entries, to be used with a set of [natural language processing −−SB] computer programs, and keep "dictionary" to mean a physical printed text giving lexical information, including meaning descriptions."

This distinction sees only on the one hand, dictionaries in print form for use by humans, and on the other, electronic−media databases for use by NLP applications. It leaves no place for electronic−media databases for use by humans, and so is not a distinction useful to this work. Moreover, the NLP sense of "lexicon" is incompatible with the meaning this word has in lexicography in general, where it usually refers to dictionaries which are atypical in form or content, e.g., *The Analytical Lexicon of Navajo* (Young & Morgan 1992), *A Concise Hopi and English Lexicon* (Albert 1985), *A Lexicon of New Red Sandstone Stratigraphy* (Taylor 1988), and so on. My use of "lexicon" is based on lexicographic usage, not NLP usage.

"Online"

In "Lexicomputing and the Dictionary of the Future", Dodd (1989) made these comments about the distribution media for lexicons:

> It is clear that we are not far from the point at which the dictionary will cease to be merely a *product* such as a book, or a somewhat more sophisticated substitute for a book, for example, a CD−ROM, which remains as fixed in its contents as a book is, and will become a *service*. This implies that instead of multiple identical copies of a dictionary, sold to users, there would be a single version of a database, from which clients of the dictionary services obtained the information they required, much as professionals of various sorts already get abstracts and similar data "on−line". [Dodd 1989:87, emphasis in the original]

Dodd's sense of an "on−line" "sevice" is exactly what I mean by an online resource, specifically an online lexicon. (The reader may find interesting the fact that while Dodd's comments sound hypothetical, an online lexicon system, the Internet `webster` had already been available via Internet/ARPANet since the mid 1980s; under this system, users would run simple client programs (called `webster(1)` or, later, Xwebster) to query, thru the Internet, one of the remote servers for definitons or spellings (Unknown ?1983; Curry 1990, 1996; Mayer 1996; Faith and Martin 1997)).

To rephrase and expand Dodd's conception of "online", I say that if a lexicon is online, it exists not on each user's computer (nor even on a CDROM accessed thru a local network), but instead it is served, across a network, from the lexicographer's computer. While much of my discussion, such as the section on macrostructure, applies to CDROM lexicons about as well as to online lexicons, the sections on interstructure and on editorial issues are relevant to online lexicons but have little applicability to CDROM lexicons.

Macrostructure

I use the word "macrostructure" here to refer to way the lexicon is set up such that users can enter the lexicon and find the desired headword. I do *not* use it to mean a *physical* structure of the medium of the lexicon (although in print lexicons, layout of the bound volume is an artifact of the method of access, as is discussed below), but instead the *procedural* structure of the how the user goes about accessing entries.

In this section I will discuss how macrostructure works essentially differently in online lexicons as compared to print lexicons.

Macrostucture & Indexing in Print Lexicons

In print lexicons, the entries are organized according to one macrostructure. For a user to find anything in the lexicon, he must learn the rules which underly this macrostructure.

For a user to find "patchouli", for example, in a conventional English dictionary, he must understand the English conception of alphabetical order –– e.g., he must know the order of the alphabet, he must know that the sort starts from the left edge of the word, and not the right (except in rhyming dictionaries); if he is a Spanish speaker he must unlearn the Spanish convention of treating "ch" as a letter between "c" and "d"; and so on. The user must then open the volume (assuming it's a one–volume lexicon). Recognizing that page numberings start on the left end of the volume, the user must then use runners at the top of the page to narrow in on the "p" section, then to find the beginning of it to find the "pa" words, and so on until he finds the left–justified boldface headword "patchouli".

The skills for navigating the macrostructure of conventional English lexicons are generally learned in the elementary grades, and it is basically a simple system –– one need only know the spelling of the word, and the straightforward rules for alphabetical sorting in English.

However, consider if a user wanted to find a word based on criteria other than its exact spelling. Suppose one wanted words which rhymed with "enroll", or which referred to a shade of red, or which were reflexes of the Latin etymon "capere", or which were pronounced /si:z/, or which were seven letters long, or which ended in "–ate".

In that case, the macrostructure of the conventional English lexicon is inadequate, and the user must use a dictionary with a specially adapted macrostructure (for example, a rhyming dictionary, a crossword–puzzle dictionary, or an etymological dictionary); or the user may find this information in an index in the back of the conventionally structured dictionary.

It is a basic fact about print lexicons that they have exactly one macrostructure –– no more and no fewer. A lexicon can't be devoid of macrostructure, or else it would be an unsorted wordlist, useless as a reference work. And the only way I can conceive of a print lexicon having two (or more) macrostructures is if the lexicon simply repeated its entire contents twice, once in the first macrostructure, and then again in the second (in a different order). Presumably this would be a extravagant waste of ink and paper. If one did want to have the utility of two macrostructures in a print lexicon, one would presumably reduce one of the macrostructures to an index, i.e., by replacing the full entries with references to the entries in the other lexicon. This "reduced" macrostructure isn't a proper macrostructure anymore, since it consists of references, instead of entries; it is just an index.

Indices and macrostructures are not on an equal footing. To navigate a macrostructure to get to an entry, the user is not obliged to know anything about any indices; but to use an index, he must know how to navigate the index *and then* how to follow up the index's references into the macrostructure; and to do that follow–up, the user must know the macrostructure's ordering rules.

For example, consider *The Pinyin Chinese–English Dictionary* (Jingrong 1979) as I would use it to discover the meaning of a Mandarin word. The macrostructure is that of an alphabetical sort of the Mandarin words, as represented in the Pinyin orthography (standard Romanization for Mandarin), and headwords are in Pinyin, followed by the Chinese ideogram. However, there is a large index which indexes entries by the graphic form of their ideogram, and gives the Pinyin spelling of each. If I want to find the meaning of a Mandarin word which I know to write in Pinyin as "yù", I go right to the main part of the lexicon and find "yù" in the macrostructure. In this macrostructure I use the rules of alphabeticization which, in this lexicon, are the same as for English alphabeticization. If, instead, I were going to the dictionary to find the meaning of an unfamiliar ideogram, I could *not* look up the ideogram directly to discover its meaning, as I could with the Pinyin spelling. Instead, I have to consult the ideogram index to find out the Pinyin spelling of ideogram, and then look that up in the macrostructure. In either of these cases, I must know alphabetical order, since I always end up in the Pinyin macrostructure. In short, the macrostructure is like Rome: all roads, or references, lead to it.

Macrostucture in Online Lexicons

A print lexicon is a fixed, physical artifact, and the macrostructure is mapped onto the storage medium of that artifact–– i.e., the start of an English language lexicon is at the *physical* left end of the volume, the end is at the *physical* right end, and the middle is physically inbetween.

Online lexicons do indeed exist as physical objects; their data is encoded in a material object which exists in a specific place. But the nature of digital media has made irrelevant all details of where information is stored, or in what sequence. An online lexicon is not *perceived* as a physical object any more than a movie or a video game is, even though all of these are stored and accessed only through physical objects. An online lexicon, like any online resource, is perceived as data presented in whatever way the interface chooses to present it –– the implication being that the user may be able to reconfigure his interface to display the entries differently. In this way, an online lexicon is essentially dynamic, whereas a print lexicon is inherently static.

The reader may find my use of "macrostructure" unusual, since in other works on lexicographic theory (e.g., Landau 1984), it *refers* to the designed arrangement of entries in the physical medium of the lexicon. However, I see the physical structure as being merely an artifact of the steps the user is meant to follow in getting to entries, and I instead use "macrostructure" to refer to the these steps, to this plan of action; this sense happens to *imply* the physical structure of the print lexicon –– but it has no such implication with online lexicons, given the lack of essential physicality, as discussed above. Since this lack of physicality does not, in my experience, disorient users or keep them from learning a given online lexicon's macrostructure, I cannot help but conclude that the physical artifacts of macrostructure in print dictionaries are not an *essential* design feature of lexicons in general.

So, viewing macrostructure as the procedures that the user is meant to follow in getting to the entries he wants, we arrive at the basic novelty of online macrostructure: there are as many macrostructures in a given lexicon as there are search methods that the programmers and lexicographers have provided. Dodd (1989:88), in referring to "routes" (synonymous with what I call "macrostructures"), says:

> "In a truly dynamic dictionary, it should be possible to gain access to an entry by means of any of the pieces of information composing it. Potential routes are thus limited only to the frontiers of what is contained in the dictionary, combined with possible manipulations or intersections of these items of data."

This is a tall order, but it is a goal that designers of online lexicons should try to meet. At every stage of the design of the lexicon, designers should ask "is there *another* way I can make this lexicon searchable? Is there another way to link to the entries?" Of course, making a lexicon searchable by "any piece of information" in entries is feasable only where that information is not merely present in entries, but is also systematically coded in a form amenable to search routines. For example, in an English dictionary, if argument structures of verbs being defined are not explicitly stated, but instead are merely demonstrated in example sentences (as is often the case in English dictionaries), then it will be difficult if not impossible to write a search routine so that users can search for verbs having particular argument structures. In that case, it would probably be simpler to edit all the verbal entries in the lexicon to have an explicit formalization of their argument structure, in a form usable by search routine.

It has to be decided on a lexicon–by–lexicon basis what aspects of the content of entries is worth encoding so as to be searchable. But in existing print lexicons, lexicographers have shown what kinds of information they consider important enough to enshrine as an aspect of the macrostructure (e.g., the spelling of a word); or important enough to consistently declare in entries (as with part–of–speech, etymology, etc.); or important enough to compile into indices. These kinds of information are exactly the kinds of information that lexicographers of online lexicons should consider making accessible as macrostructures. To wit:

- Users should be able to access entries by simply searching for headwords matching a string they type in. This is the most obvious macrostructure in online lexicons, and I know of no online lexicon where this is not the main macrostructure. This macrostructure provides the functionality of the primary macrostructure of print lexicons, headword lookup.
- Users should be able to search for entries which are of a certain part of speech, or of a certain subcategory of a part of speech (e.g., transitive verbs). This macrostructure provides the functionality of part−of−speech indices in analytical lexicons (e.g., the noun indices in Young and Morgan 1992).
- Users should be able to search based on etymology or morphological composition. E.g., users should be able to search an English lexicon for all reflexes of a particular Anglo−Saxon word, or to find all loanwords from Malay, or to find all headwords based on the suffix "−osis". This macrostructure provides the functionality of etymological dictionaries such as Weekley (1952) as well as rarer morphological wordbooks like Marchand (1960).
- Users should be able to search based on what register or dialect words belong to; e.g., to search for words which are literary, or are slang, or are vulgar, or are exclusive to Scots English, et cetera. This macrostructure provides, and greatly expands upon, dictionaries of slang, regionalisms, or other particular speech registers.
- Users should be able to search based on the semantic field of a particular word (for any conceivable lexicographic formulation of the concept "semantic field"). This macrostructure provides the functionality of such pedagogically useful topical dictionaries such as Kick and Henry (1988). This macrostructure is particularly well implemented in WordNet (Cognitive Science Laboratory at Princeton U. 1995).
- Users should be able to search on aspects of phonological content of headwords. In a simple case, this could take the form of being able to search for words which rhyme with a given input word, or which have the same metrical pattern, and as such this macrostructure would provide the functionaly of rhyming dictionaries. Moreover, with the introduction of even a simple search language such as regular expressions (Friedl 1997), it becomes possible for users to formulate quite complex queries, such as to search for all headwords which, for examples, are disyllabic, begin with "n", and contain no "t"s or "d"s.
- In the case of languages with ideographic or pseudo−ideographic writing systems, users should be able to search on aspects of the graphic form of headwords, whether this takes the form of straightforward composition (e.g., in searching for all Chinese glyphs based on a particular graphic radical), or of higher−level characteristics (e.g., in searching a lexicon of Egyptian hieroglyphs for all which ideograms which depict animals). This macrostructure would provide (and could greatly expand upon) the functionality found in glyph−composition indexes such as are found in Jingrong (1979).

Of course, in an online lexicon with a well developed and powerful search system, one should be able to compose queries consisting of various criteria from each of the above macrostructures, such that one could, for example, search a Chinese lexicon for words which belonging to the literary register of Chinese, whose glyphs contain a given graphic radical, but which do not start with "b".

I emphasize that what I call "macrostructures" and what Dodd calls "routes" (see above) need not be seen merely in terms of the process of submitting a query to the online lexicon and having it return a list of matching entries. If the coding for a given entry represents it as, say, belonging to the semantic field "kinship terms", it is probably because the lexicographer expects users to formulate queries searching for kinship terms. However, in the case of lexicons which are heavily hypertextual, e.g., Lachler, McElwain, and Burke (1995), the datum that a given entry belongs to semantic field "kinship terms" is represented, when that entry is displayed, as a hyperlink to a list of all other words belonging to that semantic field. (Or, similarly, in a Chinese dictionary, this hyperlinking can be to all other words based on a given graphic radical; or in an English lexicon, to words which are cognates of a given entry, and so on with all the macrostructures discussed above.) Strictly speaking, such hyperlinking adds nothing to the *content* of the lexicon; however, as an interface feature, it shows users that, regardless of which macrostructure they used to get to the entry in question, that entry is similar to other entries in various other dimensions accessible thru other macrostructures. For naîve users, this provides a painless way to start exploring the various macrostructures of a given online lexicons. For more advanced users, it allows for the kind of half−structured browsing that so often leads one to stumble on the kinds of correlations that are the raw material of lexical research.

Fuzzy Matching & Stemming

I anticipate that the primary macrostructure for online lexicons will continue to be variations on the general theme of headword lookup, where a user enters a search key and expects to see any headwords containing that search key.

However, significant extensions to this basic "substring match" algorithm can be made. First off, "fuzzy matching" can be incorporated into the matching algorithm. That is, instead of merely looking for headwords which exactly match the user's query, the "fuzzy match" algorithm will be able to match headwords which *approximately* match the user's query. This feature is now used in spellcheckers to identify misspelled words and to suggest corrections. The fuzzy matching, integrated into a lexicon's lookup routines would be able, for example, to tell a user searching an English lexicon for an entry for "perogative" that there is no such word, but that "prerogative" is likely to be what he was after. This feature is present in the Internet `webster` (Unknown ?1983).

Fuzzy matching algorithms could extend from repairing spelling mistakes native speakers make, to repairing spelling mistakes common to non−natives who are likely to be using the lexicon. For example, Sherman Wilcox has included in his *Multimedia Dictionary of American Sign Language* (See Wilcox et al 1994) a fuzzy matching algorithm which (among other things) corrects for kinds of misperceptions of signs that non−Signers most often make. The details of the implementation of fuzzy matching algorithms depend on the language in question, as well as the kinds errors that potential users are likely to make.

In a similar vein, lookup routines should be able to accept orthographic variance which is not objectively incorrect. For example, a user searching a German online lexicon for the word "hoeren" should be redirected to the entry for "hören" without being accused of bad spelling. A dictionary of Arabic should be able to accept vowelled or unvowelled input; a Mongolian lexicon should accept Cyrillic or Old Script input; and so on.

The second significant extension to the matching algorithm is the integration of a stemmer algorithm. "Stemmer" here refers to an algorithm which can take an occurring (declined, conjugated, etc.) form of a word and return its headword form. Dodd (1989:89−90) says:

> Where such a morphological analyzer might be of most worth would be in languages with initial mutations, such as Welsh, Cornish, and Breton. These mutations lead to extreme difficulty in alphabetically ordered books in those cases where the language [i.e., in deriving noncanonical forms −−SMB] respells the words affected, leaving no indication of the original form. [...] It would also be of great value in languages with considerable morphological marking and numerous irregular forms. Such complexities can lead to major problems in a normal printed dictionary, obliged to be of much greater bulk than otherwise, through including the varying forms at least as cross−references to the normal headword, with no certainty of success. Some examples from Welsh are relevant:
>
> - *nghestyll*, headword *castell* "castle";
> - *ddeurudd*, headword *grudd* "cheek";
>   [...]
>
> −− all of these are mutated variants of plural nouns. [...]

In other words, stemmers ("morphological analyzers" as Dodd calls them) can solve one of the most difficult problems found in lexicography –– namely, how to make lexicons of languages where morphology is not just something that happens at the end of words. Dodd mentions the solution of listing all derived forms, but in many languages this is impractical. The alternate solution involves either choosing some form as the "canonical" headword form (Zgusta 1971:120–1), or, as Young & Morgan (1992), using roots as headwords. Whether it's a canonical occurring form or a root which ends up being the headword in the macrostructure of a given lexicon, a potentially huge amount of linguistic and metalinguistic knowledge is required of the user.

To use the Welsh example, a user wanting to find "ddeurudd" in a Welsh dictionary must be familiar with the morphological and morphophonemic processes which have been brought to bear on "ddeurudd"; he must be familiar with the analysis of "ddeurudd" as a mutated form of one word from a paradigm of words which are all differently inflected forms of a common base; and he must know that a certain word among that paradigm, "grudd", is what the dictionary in front of him uses as a headword form. If the user does not have such metalinguistic skills (which may require knowledge of extremely complex analyses of the morphology of the language) as well as knowledge of possibly arbitrary and unintuitive decisions that were made in the organization of the given dictionary, he will be unable to find words in the dictionary, even though he may be fluent as well as literate. In the particular case of Native American languages, it's generally unrealistic to expect users to possess such metalinguistic knowledge.

But a stemmer absolves the user of having to possess such knowledge. Once a stemmer has been integrated into the lookup algorithm, the user no longer has to learn to produce citation forms; he can feed any occurring form into the search box, because the stemmer will deduce the citation form and direct him to the appropriate entry.

It may not be easy to write a stemmer for a given language. It is likely to be quite difficult for languages with complex phonologies or morphophonologies (such as Yawelmani or Mingo) or difficult writing systems (such as Hebrew or Tibetan). However difficult it may be to develop smart stemmers, it is worthwhile, since it will make the lexicons usable by (and less frustrating to) people who are not fluent with the principles of what is and isn't a canonical form for the given languages.

The implementational details of stemmer algorithms are beyond the scope of this document. Anyone wanting to develop a stemmer algorithm for a particular language would profit from a reading of Sproat (1992), and especially from looking at the algorithms used in existing stemmers for languages typologically similar to the one in mind. A word of warning is necessary, though: whereas I use "stemmer" in the electronic dictionary sense of the word, to mean an algorithm that takes an existing form (from a user's query) as input and returns its *headword* form, the term is also used in much of the literature on computational morphology in a different sense, to refer to algorithms which take an existing form and return an *abstract stem.* The distinction is crucial in two ways: if the headwords in a given lexicon *are* abstract stems, the stemmer's formalization of stems needs to agree with the formalization used by the lexicographer in composing headwords. But more importantly, if the headwords in a given lexicon *are not* abstract stems, the algorithm needed for a stemmer (in the electronic dictionary sense) may have little or no relation whatsoever to the one needed for a stemmer (in the computational morphology sense), and in fact may be degrees of magnitude more complex.

Consider, for example, the case of Navajo verbal morphology. Two largest dictionaries of Navajo, Young & Morgan (1987) and Young & Morgan (1992), both use the same analysis of verbal morphosemantics, namely, that Navajo verbs consist of a word–final monosyllabic root, which provides the core meaning of the verb form, and a prefix complex, which modulates the meaning. Roots are slightly modified versions of an underlying form, conventionally called the "stem" (Lachler 1997). For example:

adi'ní
adi' – ní
IMPERFECTIVE – thunder.DURATIVE
"Thunder is rumbling."
(where "ní" is the durative root for the stem "nih", meaning "thunder rumbling")

Where Young & Morgan (1987) and Young & Morgan (1992) differ notably is in their macrostructural treatment of verbs. In Young & Morgan (1987), verbs are made into entries based on a canonical existing form, with that form being the headword. A user looking for a definition for "adi’ní" would simply look for a headword "adi’ní". In Young and Morgan (1992), however, verbs are arranged into entries by stem, with subsections for the different prefix complexes. A definition for "adi’ní", therefore, would be under the headword "nih", in the subsection for "adi’". (For sake of simplicity, I am ignoring Young and Morgan’s further analysis of the prefix complex.)

In the case of an electronic dictionary organized like Young & Morgan (1992), a stemmer that would get the user to the appropriate verbal entry would consist of merely an algorithm to identify the last syllable of the user’s query, account for minor root/stem alternation, and look for that headword. That is, in this case, a stemmer (in the computational morphology sense of the word) works fine as a stemmer (in the electronic dictionary sense of the word), since headwords and abstract stems are synonymous in Young & Morgan (1992).

However, a stemmer for an electronic dictionary organized like Young & Morgan (1987) would need to be much more complex; it would have to go from one occurring form (the one in the user’s query) to another (the canonical form used as the headword). This would require that the algorithm model at least a significant subset of the morphology of Navajo prefix complexes. In fact, given the incredible complexity of precisely this aspect of Navajo morphology (consider Kari’s book−length treatment (1976) of the subject), writing such an algorithm would be a major undertaking, requiring, say, a lookup table containing correspondences of tens of thousands of possible prefix complexes (in all combinations of object and subject person and number, in all tenses, etc.) to their canonical forms; or, alternately, a comparable number of lines of program code to model the morphonology underlying the generation of the these complexes.

While Navajo is quite an extreme case as far as the problems in stemmer design, comparable issues are to be found in constructing stemmers for languages such as Arabic, where we also find complex nonconcatenative morphology, very different formalization of headwords in different dictionaries (see Haywood 1965), and different formalizations of stem structure in existing stemmer algorithms. (And this is to say nothing of the special difficulties presented by complexity and variability of the Arabic writing system.)

As daunting a task as stemmer design may be for some languages, it is precisely such languages which most need stemmers in the lookup routines for their electronic dictionaries; if it’s difficult for a lexicographer−programmer to write a stemmer for a such a language, then it’s certainly harder still for a user (especially a metalinguistically naïve one) to use a dictionary which lacks a stemmer in its lookup routine.

Multiword Queries

Compared to the task of developing fuzzy matching routines and stemmers for single word queries, it is relatively simple to then add functionality to the lookup routine to handle multi−word lexical items, such as compounds or idioms. This solves (or obviates) a longstanding lexicographic problem: where in a dictionary should one define, for example, "North Star"? In the entry for "north"? In the entry for "star"? In an entry of its own? Whatever principled solution a particular dictionary settles on for dealing with multi−word lexical items such as "North Star", it will be arbitrary. However, in an online lexicon, the lookup routine can and should be designed so as to know the right place to look when the user runs a search on "North Star".

Microstructure & The Content of Entries

Microstructure is the way that the content of each entry is organized. This section discusses the implications that new online media have for what microstructures are possible, as well as what new kinds of content are possible.

Density in the Microstructure of Print Dictionaries

Here is a definition of the noun "dog" from Merriam-Webster (1963:246):

> **dog** \'do.g\ *n, often attrib* [ME, fr. OE *docga*] **1a:** a highly variable carnivorous domesticated mammal (*Canis familiaris*) prob. descended from the common wolf; broadly : any animal of the dog family (*Canidae*) to which this mammal belongs **b:** a male dog **2a:** a worthless fellow : **b:** CHAP, FELLOW <a gay ~> **3a:** any of various usu. simple mechanical devices for holding, gripping, or fastening consisting of a spike, rod, or bar **3b:** ANDIRON **4a:** SUN DOG **4b:** WATER DOG **4c:** FOGBOW **5:** affected stylishness or dignity **6** *cap* : either of the constellations Canis Major or Canis Minor **7** *pl, slang* : FEET **8** *slang* : something inferior of its kind **9** *pl* : RUIN <go to the ~s> **10** *cap* : any of various American Indian peoples – **dog.like** \'do.-.gli-k\ *adj*

As dictionary users, we may be so used to this format that we overlook its most distinguishing characteristic: it is extremely (some would say unreadably) dense. Specifically:

The layout for the definition for "dog" has no whitespace to speak of -- e.g., the various senses and subsenses run together instead of each being a new paragraph. Whitespace helps comprehension by having the divisions in the layout parallel the divisions in logical structure; this is why we have the concept of "paragraph" which at once conveys a division in thought and in layout.

There is extreme use of abbreviations in the above entry. Looking for what I would most prototypically call abbreviations -- i.e., bits of typography that I expand to full words when I read aloud -- I count these eleven: "n ME fr. OE attrib. prob. ~ usu. pl cap adj". The reader should note that none of these abbreviations are in wide use outside of lexicographic or perhaps linguistic work.

However, the delimiters of the various subsections can be considered to be abbreviations of sorts: a boldface number and/or letter, and colon, such as "3a:", are quasi-abbreviations for "New sense, number three-A". When a "slang" or "pl" or "cap" comes between the letter/number and the colon, these mean "New sense, which is slang,..." or "New sense, which occurs only in the plural..." or "New sense, which is written with an initial capital letter...", and so on. Similar quasi-abbreviations are: "\", used to bracket pronunciations; "[" and "]", used to bracket etymologies; and allcaps (e.g., "ANDIRON"), used to mean that the word in allcaps has an entry headword in the dictionary which the user is advised to find and read. All of these abbreviations and quasi-abbreviations must be understood if the user is to completely understand all the information that this definition seeks to convey.

Moreover, there is no use of metalanguage, as it is called in Rey-Debove (1971:43-52). Using metalanguage, instead of writing "**dog** \'do.g\", the lexicographer would write "the word **dog** is pronounced as \'do.g\". This tells the reader whether he should consider "**dog**" to mean the word *dog* (orthographically?), the sound of the word *dog*, the referent of the word *dog*, or the entry *dog* (as a cross-reference? in this dictionary or elsewhere?). Sophisticated dictionary users can grow accustomed to inferring which of these is meant, and several cues are given by typography. However, none of the typographic cues (e.g., allcaps for cross-references entries) are part of the general typographic conventions associated with the English language, and they are arbitrary and unintuitive; and the process of inference which users must rely on to understand these abbreviations is unreliable, as inference always is.

Consider then what the entry for "dog" might look like if whitespace were used, if abbreviations were expanded, and metalanguage were used:

> **dog**
>
> The word "dog" is pronounced as \'do.g\
>
> This entry defines the word "dog" when used as a noun.

The word "dog" occurs in Middle English, and is derived from the Old English *docga*

Senses:

Sense 1a: in this sense, "dog" refers to a highly variable carnivorous domesticated mammal whose scientific name is *Canis familiaris*. This animal is probably descended from the common wolf.
In a broader sense, this can refer to any of the dog family (whose scientific name is *Canidae*) to which the domesticated dog belongs.

Sense 1b: The word "dog" in this sense refers to a male dog.

Sense 2a: The word "dog" in this sense is synonymous with "a worthless fellow".

Sense 2b: The word "dog" in this sense is synonymous with the words "chap" or "fellow". (We recommend reading the entries for these words in this dictionary.) An example usage of this sense is the phrase "a gay dog".

Sense 3a: The word "dog" in this sense refers to any of the various, usually mechanical, devices for holding, gripping, or fastening consisting of a spike, rod, or bar.

Sense 3b: The word "dog" in this sense is synonymous with the word "andiron", which there is an entry for in this dictionary.

Sense 4a: The word "dog" in this sense is synonymous with the phrase "sun dog", which there is an entry for in this dictionary.

Sense 4b: The word "dog" in this sense is synonymous with the phrase "water dog", which there is an entry for in this dictionary.

Sense 4c: The word "dog" in this sense is synonymous with the word "fogbow", which there is an entry for in this dictionary.

Sense 5: the word "dog" in this sense refers to affected stylishness or dignity

Sense 6: This sense is always written capitalized. "Dog" in this sense refers to either of the constellations Canis Major or Canis Minor

Sense 7: This sense is found only in slang, and occurs only in the plural. In this sense, "dogs" means "feet", which there is an entry for in this dictionary.

Sense 8: The word "dog" in this sense refers to something inferior of its kind.

Sense 9: This sense is found only in the plural in the expression "go to the dogs", which means "be ruined". There is an entry for "ruin" in this dictionary.

Sense 10: This sense is always written capitalized. In this sense, "Dog" can refer to any of the various American Indian peoples.

The word "dog" has the derivative **doglike**, which is an adjective, and which is pronounced \'do.−.gli−k\

I believe that this more clearly conveys exactly the same information as in the original, dense definition. Why, then, isn't this format, or something like it, used for definitions in print dictionaries? Obviously, because this format takes up a huge amount of space when printed on printed on paper. Landau (1984, especially pages 248−250) discusses how length of entries, the number of entries, and printing factors such as point size, line spacing, and whitespace must all be very carefully controlled, lest the lexicographic labors of years or decades end up producing a dictionary which is twice as large, heavy, and expensive as initially planned −− and therefore at least twice as unsellable.

> Almost every criticism made of dictionaries comes down at bottom to the lexicographer's need to save space. The elements of style that so baffle and infuriate some readers are not maintained for playful or malicious reasons or from the factotum's unthinking observance of traditional practice. They save space. Every decision a lexicographer makes affects the proportion of space his dictionary will allot to each component. It is perfectly fair for critics to question his judgement, but they must realize that the length of a dictionary is finite, and as large as it may appear to them, it is never large enough for the lexicographer. [Landau 1984:87]

Any dictionary which formats an entry for "dog" as I've done above would be clearer and more comprehensible than the denser ways of formatting it, but I estimate that it would be at least five times as large. Merriam−Webster (1963) is already an immense volume, and as such already has a limited market; multiplying its size, weight, and cost by five would make it undesirable to what market Webster's Seventh already has. And so print dictionaries must use dense formatting because of the need to save page space, with consequences for readability.

The Microstructure of Online Lexicons

While space is scarce in print lexicons, it is an abundant resource in online lexicons. This is because digital storage media are extremely efficient for storing immense amounts of text information. At the time of this writing, CDROMs can store about 660 million bytes of information; in comparison, the MSWord source files for the *Analytical Lexicon of Navajo* (Young & Morgan 1992) take up up about 10 million bytes. In hardcopy, Young & Morgan (1992) is about 1500 pages; this gives us a conversion rate of about one million bytes to 150 pages of dense type. This means that a redaction of Young & Morgan (1992) could increase the size of the lexicon by a factor of sixty−six, producing a lexicon equivalent to 99,000 pages of dense type, and would, in digital form, *still fit on one compact disk.* In short, space is not at a premium with text in online lexicons.

What are the ramifications of this new luxury of space? In terms of microstructure, it means that my verbose and unabbreviated entry for "dog", above, may be preferable to the dense and obscurely abbreviated Webster's 7th formatting. At the very least, online versions of print dictionaries no longer have any compelling reason to use abreviations; abbreviations should be expanded. This is a trivial task which can even be performed as part of the interface routines which display entries. For example, to expand all instances of "n." to "noun" in a given entry can be done in a single line of code in a PERL program:

```
$entry =~ s/\bn\./noun/g;
```

Despite the ease with which abbreviations can be automatically expanded, most electronic lexicons still do not provide for this. The Internet `webster` (Unknown ?1983) for example, is still thick with abbreviations, a reflection of the fact that is merely a keying in of a print dictionary (i.e., Merriam−Webster 1963). *Webster New World Dictionary, Third Edition, with the American Concise Encyclopedia on Power CD* (ZCI Publishing 1995) and *The American Heritage Talking Dictionary, Third Edition* (American Heritage 1994) are just as filled with abbreviations as their print sources. *Merriam−Webster's Collegiate Dictionary, Deluxe Electronic Edition* (Merriam−Webster 1994) seems to expand its abbreviations, but beyond this, entries are exactly as they would appear in print.

Just as abbreviations can be automatically expanded, so can formatting codes be easily changed. This may be a more complex task, depending on the markup language the lexicon is coded in, but it is feasible in most cases. For example, in the experimental electronic version of Young & Morgan (1992) which I have produced, the Hypertext Markup Language tag "<P>" is added after every instance of the tag which ends every list of subentries. This <P> tag adds whitespace to the entries to make the divisions between sections clearer.

New Textual Content in Online Lexicons

The issues of microstructure I've discussed have merely addressed the issue of how to format what content is already there. I'll now turn the discussion to what new content can exist in online lexicons which is not commonly found in paper lexicons.

Full Paradigms

In languages with morphology more complex than that of English, information about the inflectional behavior of regularly behaved lexemes is often conveyed in an abbreviated form. In Latin dictionaries, for example, the headword "mos" (a noun meaning "character") will be followed by "moris". To users familiar with the Latin declension system as well as with conventions of Latin dictionaries, this signifies that the stem is "mor−" and that it is declined as a regular class three noun. However, such abbreviated ways of signaling the inflectional pattern, as concise as they are, are not intuitive, and take some practice to learn. But these abbreviated formats are used, because of a need to save space.

In a user−friendly online lexicon of such a morphologically complex language, it would be useful to the non−expert user to offer a more expanded sample of the inflection of the headword in question. In Latin, for example, there are only two grammatical numbers (i.e., singular and plural) and, for most nouns, five cases; so the entire declensional possibilities of "mos/moris" can be shown in a small table. In a language where the number of possible inflected forms of a root is orders of magnitude larger than the ten forms of most Latin nouns, it would still be pedagogically useful to represent at least the most frequently used forms and have the rest be viewable if the user desires them.

These forms need not even be coded in the underlying structure of the dictionary; instead, if the morphology of the language can be modeled in the programming of the lexicon's interface, then it can be left up to the programming to determine how, for example, "mos/moris" is to be declined, and to display these forms to the user as a part of the routine which retrieves entries from the lexical database.

Example Sentences

Beyond inflectional examples, it would no doubt be useful to give more example sentences than are common in print dictionaries. For example, I've never seen "dog" used as in sense 5, above, ("affected stylishness or dignity"), and having read the definition for sense 5, I don't feel I understand it well enough to attempt to use it, or even to confidently recognize it if I saw it. With at least one example, the usage of sense 5 would be clearer.

Necessary "Encyclopedic" Information

Robinson (1954:56) writes "a lexical definition could nearly always be truer by being longer". This is demonstrable in the definition for dog, above, in either formatting. Consider just the first sense: "a highly variable carnivorous domesticated mammal (*Canis familiaris*) prob. descended from the common wolf". This definition fails to convey some basic and salient facts about dogs which differentiate them from other domestic animals: They are smaller than oxen, they are larger than mice. Unlike cats, they are not useful for controlling the rodent population, but they can be used as guard animals. Unlike cattle, dogs are not generally not raised for their meat, pelts, or fur. Unlike rabbits, when faced with a stranger they may bark or even bite. Dogs have large sharp teeth and their bite can leave a severe wound. Dogs can be infected with rabies and in that case can transmit rabies through biting. More generally, dogs are quadrupeds, and are mammals (and so have the characteristics of mammals, such as reproducing sexually and bearing live young), and they can't fly or climb trees, but they can swim, and so on.

More information would be useful for interpretation of some of the metaphorical uses of "dog" the user might encounter: dogs are often considered exceedingly loyal (as typified in the expression "a dog is a man's best friend"); they are sometimes considered ugly or dirty animals (cf. calling an unattractive woman a "dog", or in the saying "a dog's life"), and so on. Wierzbicka (1985:169−171) lists literally dozens of other attributes (which she calles "formulae") which are not merely true about dogs, but which are necessary to an understanding of what a dog is, for the purpose of making sense of the word when it is heard. She then makes these crucial comments, well worth repeating in full:

> The definitions [for "dog" and other animal−words −−SB] proposed here state the semantic competence of native speakers, which a language learner must acquire. Hunn (197[6]:24) has insisted that statements which formulate native speakers' semantic knowledge are not to be called 'definitions' but 'descriptions'. I do not want to argue about terminology. I understand, of course, that the length of my formulae makes them look different from conventional definitions. I would insist, however, that whatever they are called, they explicate the linguistic competence that native speakers of English have and that they are, therefore, a necessary part of a complete description of English. They differ fundamentally from language−independent knowledge *about* animals that compendia such as the *Encyclopaedia Brittanica* seek to state. In any case the idea that there is some theoretically defensible model of conventional definitions, short and yet accurately reflecting a word's use, is a characteristic illusion of specialists in other disciplines. [Wierzbicka 1985:171]

The pragmatically−minded reader might at this point wonder why one would need to bother defining "dog" at all. In fact, there is some historical precedent here:

> Al−Fîrûzâbâdî [Majd al−Dîn Muhammad ibn Ya'qûb Al−Fîrûzâbâdî, a fourteenth and fifteenth century (AD) Arabic lexicographer, author of *al−Qâmûs al−Muhît* −−SB] used five letters as abbreviations: [the first being] the letter mîm, meaning "ma'rûf" (known), to avoid defining such common words as palm, bee, house, horse, and so on; previous lexicographers have frequently either given no definition, or written "ma'rûf" in full. Sometimes they had used some meaningless formula, such as "man−− the singular of men"! [Haywood 1965:86]

Depending on the purposes a given dictionary is expected to serve, this approach of simply leaving out some basic words may be wise. For example, a dictionary of electrical engineering probably should not be expected to contain a definition for "electricity" useful to the layman, since having and using such a dictionary presupposes that user has enough knowledge of electrical engineering that he doesn't need to be told what electricity is.

However, if a lexicon *is* going to bother to compose a definition for "dog" in its most basic sense, as Merriam−Webster (1963) does, and if it has effectively no limitations on space or layout, as is the case with online lexicons, then there's no good reason why it shouldn't give salient background information along the lines of at least some of Wierzbicka's "formulae", about what one needs to know about dogs to make sense of uses of the English word "dog". This may seem pointless for as ubiquitous and well known a word (and referent) as "dog", but for less common words, it is necessary. I will use "spittoon" as an example here.

"Spittoon" is uncommon enough of a word that it might send many people to the dictionary. "Spittoon" is defined in Merriam−Webster (1963:844) this way:

> **spit.toon** \spi−'tu:n, sp*−\ *n.* [spit + −oon (as in balloon)] : a receptacle for spit −− called also *cuspidor*

This definition says nothing untrue; it does say what spittoons are for (as opposed to, for example, the definition for "talc" which says nothing about the salient uses of it). However, in light of Robinson's adage (1954:56, as above), let's consider how this entry could be truer by being longer.

First, to be aware of the meanings and associations that "spittoon" has when used, a reader must know that spittoons were formerly quite common, as it was once quite common to chew tobacco. Moreover, the reader must know (or should now be told) that in the twentieth century, the habit of chewing tobacco has become rare, so that a spittoon is now considered to be a quaint artifact of the everyday life of another time, like inkwells, or wooden steamer trunks −− and, as such, they are more likely to be used as decorative bric−a−brac which are *not* to be spit into.

This is not to suggest that lexicographers working on Merriam–Webster (1963) were oblivious to these facts about spittoons; but instead that they had to suppress them for reasons of brevity, which was more necessary than completeness. However, in online lexicons, brevity is no longer as crucial, leaving completeness the prime virtue in definitions.

One may object that such "completeness" would be an exercise in pointless verbosity. This may well be the case with a monolingual dictionary like Merriam–Webster (1963), where the intended users are of the same culture as the one the language is traditionally spoken in. (In fact, I genuinely wonder who looks up "dog" in a monolingual English dictionary and what they expect to find.) However, when producing a lexicon of a language whose expected audience includes a good number of people from a different culture than the culture of speakers of that language (as is the case with Young & Morgan 1992, for example), one cannot assume that Wierzbicka–style "formulae" relevant to the language and culture in question are old news to the users of the lexicon. Failing to state at least the more informative formulae (e.g., that a given word refers to a plant that is widely known for its curative properties, or usefulness as a spice, etc.; or that a given culture considers dogs as pests, not pets) can result in unrevealing entries which are at best cryptic (as in the all–too–common case of lexicons of Native language where a Native word is glossed merely with a Linnean genus and species name, often with no hint of even whether it is a plant or animal), and at worst inviting cultural misunderstanding (e.g., failing to note that the referent of a given word is not considered a polite topic of conversation in the culture in question).

Multimedia in Online Lexicons

Suppose that I have found a definition for "spittoon" which was based on the above definition from Merriam–Webster (1963), but which has been amended to inform me, the naïve user, of the salient historical facts mentioned above. I would still have no idea what one looked like. For all I know, spittoons could be lacquered wood boxes mounted on exterior walls, just so long as they are/were customarily spit into. Suppose then that we amend the definition to include the fact that spittoons are made of unpainted metal, about a foot high and a foot round, with a wide brim, and are (or at least were) typically kept indoors, on the floor. This would make for an optimally useful textual entry for "spittoon".

But in practical terms, if I have read this textual entry, could I recognize a spittoon if I saw one, or would I mistake it for an empty flowerpot or the like? Illustrations are very useful here; simply including a photograph of a typical–looking spittoon, sitting on the floor, would be very instructive as a supplement to (or even a replacement for) a written description of the shape and size of a spittoon.

Of course, illustrations or photographs are by no means new to online dictionaries. However, consider Svensén's warnings to makers of print dictionaries: "The use of colours [in illustrations] other than black is an expensive process, which should be considered only when it is absolutely necessary." (1993:170) In online media, however, it is just as easy to embed a color image in an entry as it is to embed a black and white one, and this involves no special production costs or difficulties beyond that of procuring a suitable photograph or illustration. As Svensén notes, color illustrations and color photos are indispensible for conveying the meaning of color words, and in differentiating some kinds of plants and animals (e.g., limes from lemons, or weasels from minks). And illustrations in general are useful for conveying the appearance of the referent where this is especially salient, as it is in distinguishing breeds of dogs, species of trees, types of chess pieces, architectural terms, and so on; or in conveying the names of the various parts of a thing (e.g., labeling the parts of a flowering plant).

The media possibilities of print dictionaries are confined to text plus illustrations (whether line–drawings, photographs, maps, or diagrams) for these are about all that is possible with print. (Presumably a pop–up book or scratch–and–sniff dictionary is not feasible or especially desirable.) However, any number of media types can be used in online lexicons, notably sound clips and even short video clips.

The most obvious use of these multimedia capabilities is to convey the pronunciation of entries, instead of through the awkward symbology print dictionaries use. *The American Heritage Talking Dictionary, Third Edition* (American Heritage 1994) is an example of a dictionary which implements sound clips for this purpose.

But in addition to sound clips of word pronunciation, for some words there may be call for clips of the sound that the referent makes. Consider this entry from the Internet `webster` version (Unknown ?1983) of Merriam−Webster (1963):

> **ci.ca.da** \s*−'ka−d−*, −'ka.d−\ n [NL, genus name, fr. L, *cicada*] : any of a family (*Cicadidae*) of homopterous insects with a stout body, wide blunt head, and large transparent wings.

None of the facts in this definition are as salient as as the sound that cicadas make. (In fact, this is an undeniably bad definition because it fails to mention that they make any sound at all.) In an online lexicon, it would be simple to embed a sound clip that would enable users to hear the sound of cicadas, because knowing this sound is a crucial part of the linguistic competence (in the sense Wierzbicka uses this term) necessary to knowing what, in practical real world terms, the word "cicada" means.

Customizability

An aspect of online media in general which is especially relevant to our discussion of microstructure is customizability. That is to say, the lexicon server can customize entries for each user, in any number of ways. For example, the *Dicionário da Língua Portuguesa Online* (Priberam Informática 1996−) allows users to set their preferences for how they want entries to be formatted and what parts of the entry to include or exclude.

For example, depending on the preferences one chooses, the server will deliver the entry for "cão" (meaning "dog") with abbreviations expanded, and with no etymology, as here:

> cão
>
> **1.**
>
> *substantivo masculino*
>
> > *(zoologia)* mamífero carnívoro, da família dos Canídeos, domesticado desde a antiguidade e de origem discutível, representado por numerosas raças das mais diversas utilidades;
> >
> > *(Brasil)* cachorro;
> >
> > peça de percussão, nas armas de fogo portáteis;
> >
> > pedra saliente, nas paredes, para suster balcões;
> >
> > *(astronomia)* constelação austral (nesta acepção, grafa−se com inicial maiúscula);
> >
> > *(popular)* calote;
> >
> > homem de maus fígados;
> >
> > homem desprezível.
>
> **2.**
>
> *substantivo masculino*
>
> > príncipe ou chefe asiático;
> >
> > mercado ou estalagem no Oriente.

3.

adjectivo
branco. Feminino singular *cã*; feminino plural *cãs*; masculino plural *cãos*.

Or, alternately, abbreviations can be left intact, and the etymology can be provided:

cão

1.

s. m.
(zool.) mamífero carnívoro, da fam. dos Canídeos, domesticado desde a antiguidade e de origem discutível, representado por numerosas raças das mais diversas utilidades;

(Bras.) cachorro;

peça de percussão, nas armas de fogo portáteis;

pedra saliente, nas paredes, para suster balcões;

(astr.) constelação austral (nesta acepção, grafa−se com inicial maiúscula);

(pop.) calote;

homem de maus fígados;

homem desprezível.

(Do lat. *cane*−, "cão")

2.

s. m.
príncipe ou chefe asiático;

mercado ou estalagem no Oriente.

(Do tártaro *khán*, "príncipe; senhor")

3.

adj.
branco.

(Do lat. *canu*−, "branco")

Fem. sing. *cã*; fem. pl. *cãs*; masc. pl. *cãos*.

By using configurability options like this, the same dictionary can be made to serve varied audiences. For example, while I was producing an experimental online version of Young & Morgan (1992) during the summer of 1996, several Navajo teachers I consulted with expressed the view that the etymology paragraphs that start many of the entries in Young & Morgan (1992) should be suppressed for beginning and intermediate students using the online lexicon −− for whom the etymologies would be at best useless, and at worst confusing –– but that they should be viewable to advanced students, teachers, linguists, and other sophisticated users.

Similarly, one could suppress senses of a definition which are obsolete or which belong to jargon (as with "dog" in the sense 3a and 3b, above).

Besides optionally suppressing information for classes of users unlikely to find it useful, there is the important possibility of differently ordering the information in an entry. When ordering senses in an entry, almost all print dictionaries fall into two groups: those that order on historical principles (the oldest meaning coming first), and those that put the most "important" meaning first, where importance is generally based on considerations of frequency in common usage (Svensén 1993:213). Ordering based on historical principles is useful if the user needs information about sense development or about which sense to expect in a centuries–old text; but the historical ordering is likely to confuse many users, who naturally expect the most useful information to be at the start of the entry. An apt solution for online lexicons is to specify both orderings (historical and importance–based) in the coding for entries, and have it be configurable for each user which ordering he wants the senses to appear in on the screen.

Slate (1989, 1997) points out that customizability can be extended to the presentation of the morphological analysis within the lexicon. While a morphological analysis of a headword in a morphologically complex language may be indispensible to learners or linguists of the language, native speakers of the language may find it at best self–evident and at worst distracting.

Similarly, the phonological and phonetic representations of headwords or examples, if present in the lexicon, could be searchable and presentable as the user wishes. A user should be able to view the phonological form of a given entry at varied levels of realization, or possibly according to varied analyses. For example, a user searching for the occurrence of a phoneme/phone pattern in the content of an entry should be able to specify whether this pattern should be sought in an abstract phonological form of the headwords or examples (and if so, in whose analyses), or in more fleshed–out phonetic realizations of them. The amount of programming necessary to model phonological/phonetic rules and to convert between different analyses may be simple, or may be monumental, depending on the complexity of the analyses in question; but once such programming is in place, an interested user may be able to easily answer such questions as (to use a Mingo example) "What verb roots have, in their underlying form, an /u/ which, in the surface form of that root conjugated in the first–person–singular optative, would immediately precede a stressed vowel?" The ability to formulate and answer such questions will no doubt greatly aid phonological/phonetic research.

Configurability could even extend to issues of writing systems. For languages where there is consistent variance in spelling (as between American and British spelling; or in cases where two writing systems are involved, as with Mongolian, which can written in Cyrillic or in Old Script), it should be possible to model this variance in the programming for the interface, such that the same entry could be displayed to the user in the writing system of his choice. Such automatic transliteration would be especially welcome in the case of Native American languages, where the number of varied writing systems for each language has thus far greatly complicated even the most basic tasks of linguistic research.

Interstructure

If we use "macrostructure" to denote the ways lexicographers intend for the user to get to individual entries, what do we use to denote how an entry links to resources outside the lexicon? For this purpose I will adopt the word "interstructure" to denote the way in which a lexicon’s structure intergrates itself into resources external to the lexicon.

Print lexicons are generally "stand–alone" works. Rare is the lexicon that routinely refers the user to other works. The chief reason that a mainstream dictionary’s entry for, say, "fluoride" doesn’t refer the user to works on chemistry, dentistry, or whatnot, is simply because the average user cannot be expected to be able to easily access such works, without having to make a trip to the nearest library.

However, if a lexicon is online and served through the Internet, then other resources can be accessed as easily as the lexicon is accessed. The online lexicon does not then need to be a stand−alone resource; it can freely reference other works, through hyperlinks.

This point has not been lost in producing FOLDOC, the *Free Online Dictionary of Computing* (Howe 1994). A good number of the entries in FOLDOC refer to important external resources. For example, the FOLDOC entry for "Structured Query Language (SQL)" first defines SQL, then gives a historical analysis of it with notes on current directions in development, and then ends by linking to three resources unaffiliated with FOLDOC: a standards document on SQL, parser program for SQL, and the procedings of a recent conference on SQL.

This general approach of providing links to more detailed information about the concept the headword denotes is a useful one. I forsee it being particularly useful in bilingual online lexicons, for providing extended information about cultural−specific terms. For example, an online lexicon of Navajo could, in the entries for terms having to do with religious ceremonies, link outside the lexicon to an article comparing the various Navajo ceremonies and detailing their significance.

I believe that a rich interstructure is a useful way for online lexicons to refer the reader to relevant encyclopedic resources, which may or may not be at the same site as, or coordinated with, the lexicon itself. Interstructure can *also* be used to provide users links to entries in other kinds of lexicons. For example, the Hypertext Webster's Interface at `http://work.ucsd.edu:5141/cgi−bin/http_webster` is a gateway to the Internet `webster` (Unknown ?1983). However, for every entry retrieved through this interface, a link is provided to search for that same word in an online version of a Roget's *Thesaurus* which is unrelated to the Internet `webster`.

Interstructure is used extensively in FOLDOC and is present in the abovementioned interface to the Internet `webster`, but in other online lexicons it is hardly to be found. Ford (1996:210) evaluates *Webster's New World Dictionary, Third Edition, with the American Concise Encyclopedia on Power CD* which claims to integrate the two resources named in the title. (Note that this is an electronic lexicon but not an online one.) However, in Ford's evaluation, the integration seems to consist of little more than the two resources coming on the same CD:

> When the dictionary is opened by itself, no access to the encyclopedia is provided, and none of the multimedia resources exploited by the encyclopedia are employed in the dictionary. I find it difficult to discover any benefits that result from the integration of these two resources, however loose; and the encyclopedia, in my view, is an embarrassing companion to WNWDCD [this lexicon]. It is flashy, noisy, and distinctly unscholarly. [1996:210]

Future attempts at interstructure should aim for more thoughtfully constructed links between lexicon entries and external resources.

Editorial Issues

An online lexicon and a paper lexicon which have the same content are not merely different expressions of the same thing. The difference in media has implications beyond the issues of macrostructure, microstructure, and interstructure which I have already discussed; the change in media alters basic details of how the content of the lexicon is produced and managed. This section discusses how this can affect the process of editing a lexicon.

Ease and Speed of Maintenance

Online lexicons differ from electronic lexicons generally (e.g., CDROM−based lexicons), in that online lexicons can be changed instantly and easily, at the will of the lexicographer, and that these changes take effect instantly for all users. This follows from our definition of *online* lexicon as one where the lexicon resource is being served from the lexicographer's computer. Dodd sees the benefits of this from a point of view of being able to include neologisms:

> In the first place, the paper dictionary is inevitably static, reflecting a state of linguistic affairs that is at best a snapshot of the period immediately preceding its publication. [...] A dictionary held in dynamic form in a computer database can far more readily be kept abreast of current language, because alterations can be made very simply to any database, with the result that there is no need to wait for accumulated corrections, additions and other changes to be sufficient to justify a new edition raher than a simple reprinting. The cost of changing an entry in a database is also lower than that of printed adjustments. [Dodd 1989:87]

Dodd sees online lexicons as the way to easily keep the lexicon up to date with neologisms, and this is very important to, say, the *Free Online Dictionary of Computing* (Howe 1994), which, by being online and being instantly updatable at zero cost, avoids the basic problem faced by books about computers, namely that they are obsolete before they can even come back from the printer.

Besides easily accomodating change in the language, online lexicons allow for easy correction of errors. This is significant in light of Landau's instructions on dealing with errors in print dictionaries:

> [...] the first edition of any dictionary contains numerous errors. Some of these errors will have been detected in the page proof stage [an agonizing process which Landau describes in pages 263−264 −−SB], but unless they are very serious[,] corrections must not be made in pages. The expense is prohibitive, and provoking delays in production makes no sense. [...] [N]otations [...] should be made in a card file, called a correction file, to be implemented at the earliest practicable time in a subsequent revision. Every dictionary should have its ongoing correction file, where no error is too trivial to be noted. [Landau 1984:267]

This was all that could be done in a production model which was based on discrete printings which had to happen by certain deadlines. However, with online lexicons, as with any online resource, changes can be made instantly. The rule for dealing with errors then becomes "if you see an error, fix it *now*, before more anyone else sees it", where "fixing it" means changing the master copy of the lexicon which all the users are accessing.

The Expectation of Maintenance

With the *ability* to easily and quickly add or revise entries in an online lexicon comes an *expectation* to do so. Users of online resources are not naïve as to the broader details of their production; users quickly learn that, as a rule with very few exceptions, online resources are easily updatable. With this in mind, users will come to expect that online lexicons of living languages will incorporate new words and senses, as they come into usage.

FOLDOC, The *Free Online Dictionary of Computing* (Howe 1994), contains a clever mechanism which is useful for ascertaining where there are gaps in the lexicon: every time a user searches for a term which is not found in the lexicon, a log entry is made noting this failed search. A log analysis program periodically generates a report listing the most common unsuccessful searches. By routinely checking this report, the FOLDOC editors can discover what new words users want to find definitions for. In the case of FOLDOC, these terms are new items of jargon −− for example, the names of new file formats or new kinds of computer chips.

This method of logging unsuccessful user queries is secondary to the main way that new entries are written for FOLDOC: a user emails the editors complaining that a term could not be found, and either requesting a meaning, or making a guess at the meaning they were hoping to see an elaboration of.

Beyond knowing that online media can be easily and quickly changed, users know that the maintainers of online content can be contacted thru electronic mail. They expect online lexicons to be kept up to date and they expect that errors reported to the lexicographers thru electronic mail will be fixed in a timely manner.

Editing for Small Speech Communities

In the case of general−use dictionaries of national languages, the basic day−to−day process of editing online lexicons need not vary greatly from the process of editing paper lexicons. If anything, because of the the ease of maintenance discussed above, editing online general−use dictionaries of national languages is a mere simplification of the process of editing for print. In contrast, the editing process may change greatly in the case of online lexicons whose scope is technical jargons shared by small groups of professionals, or of lexicons which are of "small" languages, specifically Native languages where lexicography (and probably even functional literacy, considering McLaughlin 1992) is relatively new.

While the circumstances of, say, a nematologist and of a speaker of Western Apache, may vary greatly in the conditions of use of their respective language forms and the sociolinguistic and socioeconomic implications, they do share one factor of extreme relevance to lexicographic production: they are members of small speech communities.

The market for general English dictionaries is large enough to support several large dictionary houses in the US alone, each constantly producing varied revisions and editions of their print lexicons, with the per−unit cost kept low by virtue of the economy of scale resulting from the printing and distribution of millions of dictionaries a year. Small speech communities, however, cannot financially support a comparable production cycle of constantly recompiling and re−editing print lexicons, and this leads to a serious financial quandary: if the editors opt for a large press run (in an effort to bring down per−unit cost), they end up with a quantity of volumes it may well take them a decade to sell off. And if, four years into that decade, the editors decide that the corrections and additions that have since accumulated are sufficient to warrant a new edition, they are faced with having to take a loss on (whether by remaindering or destroying) the six years' worth of unsold copies of the current edition. Conversely, if the press run is small, the run will sell out much quicker, allowing for frequent revisions, but the per−unit cost will necessarily be much higher.

The result of these issues of finance is that small speech communities are generally ill−served by print lexicons. One either pays moderately for a lexicon which may not have been re−edited in ten years −− and may not be re−edited for another ten still; or one pays dearly for a lexicon that is revised every two or three years; or one may simply have to do without.

The ease (and in financial terms, the low cost) of revision in online lexicons vastly increases their usefulness to small speech communities.

Editing for Lexicons of Specialist Jargons

In the case of specialist jargons, the editing of online lexicons may simply develop as a straightforward elaboration of the tasks currently associated with the editing of print lexicons of those specialist jargons; but the editing process *may also* come to involve input from bibliographers and other maintainers of existing online resources which are relevant to the specialty in question, but which are not lexicons.

The transition from being a bibliographer of a specialist field of knowledge to editing an online lexicon of that jargon of the field may not seem an obvious development. However, since the late 1980s, frequently updated electronic bibliographies of specialist fields of knowledge (whether accessible on CDROM or through online databases) have largely replaced print bibliographies. This means that specialist bibliographers are now already familiar with the tasks of maintaining keyword−searchable electronic databases whose value to their respective fields depends on them being up−to−date; they are also familiar with the evolving jargon of the field.

While I anticipate that bibliographic references in online lexicons (as discussed above, in the section "Interstructure") will *generally* come after the definitional content and will be secondary in importance to it, I believe that that new specialist lexicons could, in some cases, be developed as mere adjuncts to existing bibliographic databases. In fact, there have already been cases of the preliminary development of such online lexicons as subsystems of larger online information systems called "community systems":

> Electronic community systems [...] encode a research community's information and knowledge and provide an online environment to support the manipulation of that knowledge [...] An electronic community system helps researchers in the community function more efficiently and effectively by allowing them to browse the available knowledge easily, record their own knowledge for others to use, and form interrelationships between concepts. [Chen, Yim, Fye, and Schatz 1995:175]

Chen, Yim, Fye, and Schatz, speaking from experience gained in developing community systems for groups of biologists, assert the need to develop and (crucially) maintain online lexicons of the jargon of a community system's specialist audience, as part of community system engineering:

> According to Frenkel [1991, describing an experimental community system for the Human Genome Project −−SB], the meanings of concepts "become better understood as more knowledge is accumulated and integrated." This novel characteristic of changing definitions over time must be implemented into the community system to make the system more flexible. Research that deals with an "old" concept must still be accessible by the users even though the terminology is no longer in common use. [Chen, Yim, Fye, and Schatz 1995:177]

As community systems are still an emerging technolgy, it remains to be seen whether the content of existing paper lexicons will be adapted to serve as the lexicon for community systems; or whether the content for community system lexicons may be developed anew. It is also an open question whether long−term maintenance of online specialist jargon lexicons (whether part of community systems or not) will be done by people who are primarily lexicographers or primarily electronic bibliographers, or by mixtures of both groups.

Editing for Lexicons of Native Languages

Current editorial practice in the production of lexicons of Native languages involves regrettably little re−editing; the financial contraints discussed above are all the more severe for the production of Native lexicons. And so there is typically only one edition of a given Native lexicon, leaving no opportunity for corrections in later editions. This is especially unfortunate, as the first edition of a Native lexicon is bound to contain many more errors and shortcomings than the first edition of a lexicon of a specialist jargon, because languages are, of course, more complex than jargons; because of a relative lack of lexicographic tradition for Native languages; because the main lexicographer of a typical Native dictionary is likely to not be a native speaker of the language; and sometimes because of uneven coverage of the various dialects and registers of the Native language in question. With online lexicons of Native languages even more than with online lexicons of specialist jargons, there is the opportunity to greatly amend such shortcomings.

The best way to find and fix errors in Native lexicons is to have input from the Natives themselves who are speakers, teachers, or learners of the language, hopefully while they are routinely using the dictionary. But such input is very difficult to get if the lexicographers do not permanently live among the Native community where the language is spoken, as is the case more often than not.

However, with online media, it becomes easy to arrange for a group of Native speakers to have access to electronic mail. An email list (also called "a listserv", or "an email reflector") can be formed so that the Native speakers and the non−Native lexicographers can, as group, discuss adding and improving entries, can consider possible improvements to the design of the online lexicon, and can also provide perspectives on the role that the Native community expects the online lexicon to play in the larger task of teaching the Native language and culture.

There are, of course, difficulties: access to the Internet, while easy to get in larger cities and towns, is typically harder to get on reservations and in other rural Native population centers. Moreover, it is my experience that ownership of computers, and skill at using them, is not as common among Natives as among the general populace. However, I have seen both these situations improve immensely just within the past few years, and I believe that in the next few years, these will be only minor problems; Internet access will be as easy to get as phone service, and computers for basic Internet access will get cheaper and simpler to use. Moreover, Native consultants to online lexicon projects are likely to be involved with tribal efforts at language preservation, probably as teachers, and will thereby have access to computers and Internet access at local schools.

It is my experience that interested Native speakers with Internet access may not have the time to invest in being a full editor of a dictionary of their language; or they may lack the metalinguistic or technical skill this would require. But many of them will still be interested in answering queries about word meaning and participating in discussions about design of the lexicon, in the setting of an email list. Many Native speakers understand the crucial role that development of lexicons play for language preservation, and will often gladly welcome the opportunity to influence the design and content of lexicons of their language.

Minimally, an email list of such interested Native speakers can serve as pool of informants from whom the lexicographer can seek information on unfamiliar words. But in its highest realization, such a list can function as the official board of editors of the online lexicon. In spite of technical obstacles, it is well worth the bother to assemble such a group; it is my experience that involving informed Natives as much as possible in the production of online lexicons of their languages makes the resulting works much more linguistically accurate, more sociolinguistically relevant to the Native community, and better suited to language pedagogy. Moreover, having active Native consultants or editors means that an online lexicon is more likely to be seen not as a mere academician's computerized toy, but instead as an evolving interactive tool for representing the community's language, crucially guided by members of the community itself.

Bibliography

All URLs were valid and accessible as of November 1997.

[unknown]. ?1983. Commonly called "the Internet `webster`". [This is an electronic version of Merriam−Webster 1963 which was keyed in sometime in the early Seventies, purportedly by Security Development Corporation, but has been modified several times since by parties unknown (Curry 1990, 1996; Mayer 1996). It has been available on the Internet/ARPANet since around 1983, and is commonly called simply `webster`. A Web interface to one of the servers carrying it is available at `http://work.ucsd.edu:5141/cgi-bin/http_webster`]

Albert, Roy. 1985. *A Concise Hopi and English Lexicon.* Philadelphia: J. Benjamins.

[American Heritage]. 1994. *The American Heritage Talking Dictionary, Third Edition.* El Dorado Hills, CA: Softkey International; and Boston: Houghton−Mifflin Compnay.

Bwenge, Charles. 1989. "Lexicographical treatment of affixational morphology: a case study of four Swahili dictionaries". Pages 5−17 in: *Lexicographers and Their Words*, Exeter Linguistic Studies 14 (ed. Gregory James). University of Exeter.

Cerf, Vinton, and Robert Kahn. 1974. "A Protocol for Packet Network Intercommunication", *IEEE Transactions on Communications.* Vol. COM−22, No. 5, pp 637−648, May 1974.

Chen, Hsinchun, Tak Yim, David Fye, and Bruce Schatz. 1995. "Automatic Thesaurus Generation for an Electronic Community System", *Journal of the American Society for Information Science.* 46(3):175−193.

Cognitive Science Laboratory at Princeton U. 1995. *WordNet 1.5.* `http://www.cogsci.princeton.edu/~wn/`

Cunliffe, Richard John. 1924. *A Lexicon of the Homeric Dialect.* London: Blackie and Son Limited.

Curry, David A. (`davy@vnet.ibm.com`). 1990. Post to Usenet's `comp.windows.x`, 9 August 1990, message ID `32543@sparkyfs.istc.sri.com`

−−. 1996. Personal correspondence.

Dodd, W. Steven. 1989. "Lexicomputing and the Dictionary of the Future." Pages 89−93 in: *Lexicographers and Their Words*, Exeter Linguistic Studies 14 (ed. Gregory James). University of Exeter.

Faith, Rickard E., and Bret Martin. 1997. *Request for Comments 2229: A Dictionary Server Protocol.* The Internet Society. `http://ds0.internic.net/rfc/rfc2229.txt`

Franciscan Fathers. 1910. *An Ethnologic Dictionary of the Navaho Language.* St. Michaels, Arizona: Franciscan Fathers.

Frenkel, K. A. 1991. "The Human Genome Project and Informatics." *Communications of the ACM, 34*, 41−51.

Friedl, Jeffrey E. F. 1997. *Mastering Regular Expressions.* Sebastopol, CA: O'Reilly & Associates.

Haywood, John A. 1965. *Arabic Lexicography: Its History, and Its Place In the General History of Lexicography.* Leiden, the Netherlands: E.J. Brill.

Howe, Denis, ed. 1994−. *Free Online Dictionary of Computing.* `http://wombat.doc.ic.ac.uk/`

Hunn, Eugene S. 1976. *Tzeltal Folk Zoology: The Classification of Discontinuities in Nature.* New York: Academic Press.

Jingrong, Wu, editor. 1979. *The Pinyin Chinese−English Dictionary.* Hong Kong: the Commercial Press.

Johnson, Samuel. 1755. *A Dictionary of the English Language: In Which The Words are deduced from their Originals and Illustrated in their Different Significations By Examples from the best Writers.* [Reprinted as a facsimile edition by Scott, Foresman, and Company in 1941, although some pages are missing.]

Kari, James M. 1976. *Navajo Verb Prefix Phonology.* New York: Garland Pub.

Kick, Shirley, and Reginald Henry. 1988. *Cayuga Thematic Dictionary: a List of Commonly Used Words in the Cayuga Language, Using the Henry Orthography.* Brantford, Ontario: Woodland Pub.

Lachler, Jordan F. 1997. "Navajo Momentaneous Verb Stem Inflection", *1996 Mid−America Linguistics Conference Papers.* Lawrence, Kansas: University of Kansas.

Lachler, Jordan F., Thomas McElwain, and Sean M. Burke. 1995. *Mingo−EGADS: A Mingo−language Extensible Grammar and Dictionary System.* `http://www.ling.nwu.edu/egads/mingo/`

Landau, Sidney I. 1984. *Dictionaries: The Art and Craft of Lexicography.* New York: Scribner.

Larousse. 1971. *Grand Larousse de la langue française.* Paris: Larousse.

Marchand, Hans. 1960. *The Categories and Types of Present−day English Word−Formation; a Synchronic−Diachronic Approach.* Wiesbaden: O. Harrassowitz.

Mayer, Niels P. (`mayer@netcom.com`). 1996. Personal correspondence.

McLaughlin, Daniel. 1992. *When Literacy Empowers: Navajo Language in Print.* Albuquerque, NM: U of New Mexico

Press.

Merriam−Webster. 1963. *Webster's Seventh New Collegiate Dictionary.* Springfield, MA: G. & C. Merriam−Webster [An electronic version is available as Unknown (?1983)]

Merriam−Webster. 1994. *Merriam−Webster's Collegiate Dictionary, Deluxe Electronic Edition.* Springfield, MA: Merriam−Webster Inc.

Priberam Informática. 1996−. *Dicionário da Língua Portuguesa Online.* `http://www.priberam.pt/pages/dlpo/dlpo.htm`

Rey−Debove, Josette. 1971. *Étude linguistique et sémiotique des dictionnaires français contemporains.* The Hague: Mouton. Number 13 in the series "Approaches to Semiotics".

Robinson, Richard. 1954. *Definition.* Oxford: Clarendon Press.

Slate, Clay, Jr. 1989. *Navajo Verb Theme Categories and a Navajo Lexicon Database.* PhD dissertation: U New Mexico.

−−. 1997. Personal communication.

Sproat, Richard William. 1992. *Morphology and Computation.* Cambridge, MA: MIT Press.

Taylor, Frank M. 1988. *A Lexicon of New Red Sandstone Stratigraphy.* Nottingham, England: East Midlands Geological Society.

Weekley, Ernest. 1952. *A Concise Etymological Dictionary of Modern English.* London: Secker & Warburg.

Wierzbicka, Anna. 1985. *Lexicography and Conceptual Analysis.* Ann Arbor, MA: Karoma Publishers, Inc.

Wilcox, Sherman, Scheibman, J., Wood, D., Cokely, D., & W.C. Stokoe. 1994. "Multimedia Dictionary of American Sign Language". In *ASSETS'94, Proceedings of the First Annual ACM Conference on Assistive Technologies*, Oct 31−Nov 1, 1994.

Wilks, Yorick, Brian M. Slator, and Louise M. Guthrie. 1996. *Electric words : dictionaries, computers, and meanings.* Cambridge, MA: MIT Press.

Young, Robert W. and William Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary, Revised Edition* Albuquerque: U of New Mexico Press.

Young, Robert W. and William Morgan. 1992. *Analytical Lexicon of Navajo* Albuquerque: U of New Mexico Press.

ZCI Publishing. 1995. *Webster's New World Dictionary, Third Edition, with the American Concise Encyclopedia on Power CD.* Dallas, TX: ZCI Publishing.

Zgusta, Ladislav. 1971. *Manual of Lexicography.* Prague: Academia, Publishing House of the Czechoslovak Academy of Sciences.