Lexicons for Human Language Technology

Mark Liberman

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104-6305

myl@unagi.cis.upenn.edu

ABSTRACT

Information about words—their pronunciation, syntax and meaning—is a crucial and costly part of human language technology. Many questions remain about the best way to express and use such lexical information. Nevertheless, much of this information is common to all current approaches, and therefore the effort to collect it can usefully be shared. The Linguistic Data Consortium (LDC) has undertaken to provide such common lexical information for the community of HLT researchers. The purpose of this paper is to sketch the various LDC lexical projects now underway or planned, and to solicit feedback from the community of HLT researchers.

1. Introduction

This paper will give an overview of current LDC efforts to develop lexical resources and describe some efforts now in the planning stage. Readers are invited to join an on-going discussion of priorities, methods and even formats for our present and future efforts in this area.

1.1. Intellectual Property Rights

Since lexicons, unlike text and speech databases, are likely to be incorporated (perhaps in derived form) in commercial HLT products, intellectual property rights come to center stage. The LDC's charter as a consortium requires us to leverage the U.S. government's investment by sharing the cost of resource development among our members. This forces us to limit usage of such resources to consortium members, or others who have paid an appropriate fee. However, we also want to encourage rapid development and broad exploitation of commercial HLT technology. Therefore, we need to protect our members' investment in research based on LDC resources by ensuring their rights to future commercial exploitation without additional license negotiations, royalty payments, or other intellectual property issues.

This contrasts with the general practice for research use of machine-readable dictionaries, in which all rights to derived works are typically reserved to the publisher. For this reason, our lexicons will not be derived from existing lexicons, except as permitted by normal provisions of copyright law, or in case we are able to purchase appropriate rights from the owner of the existing resource. This also contrast with our practice with respect to text databases, where we have negotiated agreements to distribute for research purposes many bodies of text whose copyright remains with the original owner. The difference here is that the text corpora themselves will not typically be incorporated in future products, and our understanding of the applicable law (which we openly explain to information providers) is that language models trained on such text are free of any IPR taint.

We have worked hard, in consultation with our members, to develop an appropriate license for LDC lexicons. A copy of the draft license agreement for COMLEX syntax will be furnished on request to the author, or ldc@unagi.cis.upenn.edu.

2. Current English Lexicon Efforts

Our primary effort is to provide lexicons for English. We are funding a large, high-quality English pronunciation lexicon; an English syntactic lexicon, including detailed information about syntactic properties of verbs; and a set of improvements in an existing lexicon of English word-sense differences. All three lexicons will eventually be tied to appropriately sampled occurrences in text and speech corpora.

Based on an original proposal from Ralph Grishman and James Pustejovsky, we have been called our English lexicon "COMLEX," for COMmon LEXicon.

2.1. Pronunciation: PRONLEX

For the COMLEX English pronouncing dictionary ("PRONLEX"), the LDC has obtained (by purchase or donation) rights to combine four existing large and high-quality lexicons. Bill Fisher at NIST has been carrying out a pilot project to design a consensus segment set, and to map the representations in the multiple sources into it automatically. Then words where the various sources agree will be accepted, while disagreements will be adjudicated by human judges, and new words will be added as needed.

The sources we are starting from will provide coverage of more than 250K word forms. Appropriate coverage of the words found in the various ARPA speech recognition databases will also be guaranteed. We solicit suggestions for lists of other words to cover, such as proper names, including surnames, place names, and company names. The pronunciation representations used in the first release of PRONLEX, being based on those in the lexicons we are starting from, will be similar to those provided in typical dictionary pronunciation fields and used in most of today's speech recognition systems. This level is best described as "surface phonemic" rather than phonetic—it abstracts away from most dialect variation, context-conditioned variation, and casual-speech reduction.

Pat Keating at UCLA has been carrying out a pilot project to examine systematically the relationship between such normative pronunciations and the actual phonetic segments found when the corresponding words are used in conversational speech. We provided a sample of occurrences of words with high, medium and low frequencies of occurrence, drawn from the Switchboard data base. We will use the results of this study to plan how to improve the pronunciations in the initial release of PRONLEX. Readers are invited to join an on-going email discussion of this topic.

2.2. COMLEX Syntax

A lexicon of syntactic information, known as "COMLEX Syntax," is under development by Ralph Grishman and others at NYU. After designing the feature set and representational conventions, Grishman created a zeroth-order mock-up from existing resources. This has been circulated for comments and is available to interested parties from the LDC, along with the specifications for the syntactic features and lexical representations to be used. The project at NYU is now doing the lexicon over again by hand, guided by corpus-derived examples. The first release will occur later this year.

2.3. COMLEX Semantics

The existing WordNet lexical database, available from George Miller's group at Princeton, provides a number of kinds of semantic information, including hypo-/hypernym relations and word sense specification. In order to improve the quality of its coverage of real word usage, and to provide material for training and testing "semantic taggers," the LDC has funded an effort by Miller's group to tag the Brown corpus using WordNet categories, modifying WordNet as needed.

2.4. COMLEX Corpus

Because of the Zipfian 1/f distribution of word frequencies, a corpus would have to be unreasonably large in order to offer reasonable sample of an adequate number of words. Although it is no longer difficult to amass a corpus of hundreds of millions or even billions of words, complete human annotation of such a corpus is impractical. Therefore the LDC proposes to create a new kind of sampled corpus, offering a reasonable sample of the words in a lexicon the size of COMLEX Syntax, so that human annotation or verification of (for instance) four million tokens would provide 100 instances of each of 40K word types. This sampled corpus (in reality to be sampled according to a more complex scheme) can then be "tagged" with both syntactic and semantic categories. The entire corpus from which the sample is drawn will also be available, so that arbitrary amounts of context can be provided for each citation. The design of this sampled corpus is still under discussion, and reader participation is again invited.

3. Other Languages

This past year, we cooperated with the CELEX group in the Netherlands to publish their excellent lexical databases for English, German and Dutch. In this case, our willingness to pay for CD-ROM production, to handle the technical arrangements for publication, and to shoulder some of the burden of distribution was enough to help bring this resource into general availability.

As a first step towards providing new lexical resources in languages other than English, the LDC has begun an effort to provide medium-sized pronouncing dictionaries in a variety of languages. This effort, which will be coordinated with efforts to provide transcribed speech and text resources in the same languages, is beginning this year with Japanese, Mandarin, and Spanish. It aims at coverage comparable to an English dictionary with about 20K words.

As the U.S. speech research community begins to work on languages other than English, it is confronted with new issues that have reflexes in the design and implementation of even such a simple-seeming object as a pronouncing dictionary. Again, we solicit the community's participation in helping us choose a useful approach. In the next section, we would like to highlight one of the questions that will need to be answered, language by language, in the early stages of such a project.

Morphology? The question is, how should orthographically-defined units be broken up or combined in a lexicon? One answer, which is the easiest one to give for an English pronouncing dictionary

for speech recognition applications, is "not at all: list all and only the orthographic units paired with their pronunciations." For other languages, this answer may no longer apply.

Table 1 shows (for 5 databases of journalistic text) how many word types are needed to account for various percentages of word tokens. In all languages except Chinese, the principles for defining "words" in the text were the same: a contiguous string of alphabetic characters flanked by white space preceded and followed by any number of punctuation characters, with case distinctions ignored. All "words" containing digits or other non-alphabetic characters were left out of the counts, except that a single internal hyphen was permitted. In the case of Chinese, the notion of "word" was replaced by "character" for purposes of calculating this table.

As Table 1 shows, languages with a larger number of inflected forms per word, or with more productive derivational processes not split up in the orthography (such as German compounding), tend to require a larger number of word types to match a given number of word tokens. The counts for Chinese represent the other extreme, in which every morpheme (= Chinese character) is written separately, and the orthography does not even indicate how these morphemes are grouped into words (either in the phonological sense, or in the sense that any Chinese dictionary lists tens of thousands of 2-, 3-, or 4-character combinations whose meaning is not predictable from the meaning of the parts).

In English, the orthographic word is a fairly convenient unit both for pronunciation determination and for language modeling. Depending on the mix of word types in the sample, there are only about 2 to 2.5 inflected forms per "lemma" (base form before inflection), and the rules of regular inflection are fairly easy to write. Productive derivation of new words from old (e.g. "sentencize") is not all that common. Most compounds are written with white space between the members, even if their meaning and stress are not entirely predictable (e.g. "red herring," "chair lift"). For these reasons, a moderate-sized list of English orthographic forms can be found that will achieve good coverage in new text or speech.

Smoothing is required for good-quality n-gram modeling of English word sequences in text, but morphological relations among words have not been an important dimension in most approaches. Language models, like pronunciation models, can thus treat English orthographic words as atoms. As a result, from the point of view of speech recognition technology, there has not been a strong need for an English pronouncing dictionary that encodes morphological structure and features.

However, the situation in German may be different. As Table 1 suggests, simple reliance on word lists derived from a given amount of German text will produce a significantly lower coverage than for a corresponding English case, and even very large lexicons will leave a surprisingly large number of words uncovered. Thus the Celex German lexicon, which contains 359,611 word forms corresponding to 50,708 lemmas, failed to cover about 10% of a sample of German text and transcribed speech. Of the missing words, about half were regular compounds whose pieces were in the lexicon (e.g. *Lebensqualität*), while by comparison less than 1/6 were proper names.

The same sort of relative difficulty in unigram coverage can be seen in Table 2, where we look at the count of word types for a lexicon derived from one sample in order to cover a given percentage of word tokens in another sample. German requires a two- or three-times larger lexicon than English does to achieve a given level of coverage, and the factor increases with the coverage level. This is not because of differences in the type of text—all samples are drawn from the same or similar newswires, covering the same or similar distributions of topics. Spanish is in between German and English in this matter.

One simple approach is to make the lexicon into a network that generates a large set of words and their pronunciations. Thus German *Lebensqualität* will be derived as a compound made up of *Leben* and *Qualität*. The point of such an exercise is not to shrink the size of the lexicon, or to express its redundancies (although both are consequences), but rather to predict how the forms we have seen will generalize to the much larger number of forms we have not seen yet.

A similar issue arises for inflectional morphology. An Italian verb has at least 53 inflected forms (3 persons by 2 numbers by 7 combinations of tense, aspect and mood, plus 4 past participle forms, 5 imperative forms, the infinitive and the gerund). Several hundred additional "cliticized" forms (joining the infinitive, the gerund and three of the imperative forms with various combinations of the 10 direct object and 10 indirect object pronouns) are also written without internal white space. In a database of 3.2M words of Italian, forms of the common verb "cercare" *to look for* occur 1818 times, but 8 of the 53 regular forms are missing, and a larger number of the possible combinations with object pronouns. Forms of the (also fairly common) verb "congiungere" occur 89 times, and 41 of its 53 forms are missing. This indicates both the difficulty of finding all inflected forms as unigrams by simple observation, and also the greater problem for language modeling caused by the distribu-

tion of a lemma's probability mass among its various forms.

It is not obvious what the right approach is to these cases, so researchers should have convenient access to lexicons that can easily be reconfigured to provide various types and degrees of subword analysis.

Chinese presents exactly the opposite problem. The Taiwanese newspaper text used in the counts (done by Richard Sproat of AT&T Bell Labs) employs a total of about 7,300 character types in a corpus of more than 17M character tokens. Each character (with a few exceptions) is pronounced in just one way, as a single syllable. However, a given syllable might be written as quite a few different possible characters, each one (roughly speaking) a separate morpheme. There is no inflection in Chinese, but there is a lot of compounding of morphemes into words with unpredictable meanings. A typical Chinese dictionary will list tens of thousands of such combinations, and new forms are seen all the time, just as in German. However, this compounding is not indicated in the orthography.

A language model based on (at least some) compound words will of course be effectively of higher order than one based only on characters. Again, there are several approaches to this question, ranging from explicit listing of the largest possible number of multiple-character words on standard lexicographical criteria, to a simple smoothed N-gram model based on individual characters as the only unigrams. This issue has a phonetic side as well, since multiple-character words in Mandarin often have a fixed or strongly preferred stress pattern, and at least for some dialects, unstressed syllables may be strongly reduced.

Both issues—explicit representation of the internal structure of certain orthographic words, and grouping of several contiguous orthographic words as a single lexical entry—have scattered echoes in speech recognition technology as applied to English. However, other languages put these (and other) question on the agenda in a much stronger form.

4. New Kinds of Lexicons

New ARPA tasks are likely to require new kinds of resources. For instance, the outcome of the on-going discussion about semantic evaluation will probably motivate new sorts of lexicons as well as new kinds of annotated corpora.

Table 1

Corpus	Size	80%	85%	90%	94%	98%	99%	100%
AP English	3.0M	2,421	3,784	6,406	11,095	26,844	38,990	66,557
Reuters Spanish	3.0M	2,510	4,178	7,514	13,496	33,581	49,416	79,843
AP German	3.0M	4,742	8,258	16,091	31,440	76,704	107,141	137,578
Italian	3.2M	5,136	8,768	16,209	29,668	70,023	99,880	132,171
Mandarin Chinese	17M	659	843	1,124	1,509	2,384	2,937	7,337

Table 1: Number of word types required to cover various percentages of word tokens within a given sample.

Table 2

Corpus	80%	85%	90%	94%	98%
AP English	2,643	4,319	7,997	17,974	*
Reuters Spanish	3,091	5,526	11,161	28,320	*
AP German	5,558	10,715	27,404	*	*

Table 2: Number of word types, in frequency order, from a 500K-word sample, needed to cover various percentages of word tokens in a non-contiguous sample (about two months away). Asterisk means coverage at that level was not possible from the given sample.