[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

The Lexicon and MT: a position paper

Jeremy Clear
Oxford University Press

June 1990

Abstract

The recent trend towards developing the lexical component of NLP systems has focussed attention on two potentially valuable sources of lexical data: printed dictionaries for humans and large text corpora. This presentation considers the types of information that might be required by MT researchers and the extent to which this information can be derived from these two sources. This raises a number of questions, among which are the following. What type of information should be recorded in the lexicon? Dictionaries are quite comprehensive in their coverage of lexical items but how reliable are they? How can the information from a dictionary be represented in a form which is appropriate for NLP systems? Text corpora can provide a statistical basis for probabilistic models of language: what are the requirements with respect to size and composition of text corpora for deriving lexical data? Can the manual effort which is currently directed towards the compilation of printed dictionaries provide spin-off benefits for those who need lexical databases for MT?

1. What type of information should be recorded in the lexicon?

I detect a tendency recently to see the lexicon as the primary source for linguistic information at all levels: morphological, syntactic, semantic and pragmatic. There is clearly a need to enrich a grammar (which is a statement of generalities) with a lexicon which will record the peculiar and idiosyncratic features of individual words and word forms. The base form of a lexical item (let me call this the *lemma) is* a very convenient and manageable unit of organisation for linguistic information which cannot be adequately captured in the grammar—at least, not if the grammar is to be formal and a parser is to be written for it. It is interesting that the *Comprehensive Grammar of English* of Quirk *et al.,* which is broad in its scope and probably fullest treatment of English grammar in print, is more like a dictionary than a set of rules which could be mapped into any formalisms currently in fashion in MT.

For each lemma one may require from the lexicon:

- information on inflections

- part-of-speech labels linked to inflections
- alternative spellings
- word sense categories
- synonyms, antonyms, superordinates, hyponyms, etc.
- "word meaning" encoded in some formalism
- selection restrictions
- verb complementation patterns
- delicate syntactic classifications
- statistics (relative frequencies of POS and senses)
- semantic field labelling
- region and register labelling
- cross references to related compounds and idioms

2. How reliable are existing dictionaries?

This question can be answered in both positive and negative ways. On the one hand, a one-volume printed dictionary intended, say, for EFL use and containing as a consequence some detailed grammatical and collocational information may have entries for around 50,000 lemmata. Each of these lemmata may have several discrete sense categories, such that the dictionary may cover perhaps 100,000 meaning units. These entries will have been hand-crafted by skilled specialists in lexical analysis. In these terms, and compared with the size and scope of lexica which are part of existing MT systems, the printed dictionary appears to be quite a reliable source of lexical information.

But the current methods of lexicography unfortunately allow massive scope for inconsistencies, vagueness, omission, over- or under-emphasis, half truths, and so on. To a large extent, the features which may be seen by the MT researcher as diminishing the reliability of the dictionary are those which help to make the printed dictionary a marketable and user-friendly reference book for the non-specialist human. The Explanatory Combinatorial Dictionary being compiled under the direction of Igor Mel'čuk can be contrasted with dictionaries such as OALD and LDOCE. The former attempts an exhaustive and formal treatment of the morphology, syntax and semantics of each entry, but is unlikely to appeal to a publisher or to the general public as a book on the shelf to be consulted while playing Scrabble, doing the crossword, reading the newspaper or arguing over the acceptability of ‘plus’ as a disjunct adverbial. The conventional printed dictionary is designed for purposes other than to drive parsing programs.

The availability of large natural-language text corpora and software to search them has introduced an important new resource to lexicographers, and one should expect the descriptive adequacy of human dictionaries to be enhanced in the future. I believe that a detailed analysis of a large general language corpus of English will make a significant contribution to the achievement of accuracy in the compilation of new reference works. Moreover the evidence furnished by a corpus should be used to bring more consistency to the recording of linguistic information in the dictionary or lexical database.

3. How can dictionary data be represented in a form suitable for NLP systems?

The most problematic issue relating to representation is the semantic information which is recorded in dictionaries. It is acceptable for the printed dictionary to encode syntactic and morphological information in some fairly rigorous formalism, particularly since the notion of grammatical classes is traditional and almost universally accepted, and the use of symbols and codes results in an economy of expression which is desirable for a printed book. But there is no widely recognised classification of semantics and the discursive, natural-language definitions of dictionaries will remain for the foreseeable future as the preferred means of "encoding" semantic features for the human user. It is interesting to note that although the Collins Cobuild dictionary *(CCED)* made a strong commitment to record the evidence derived from a large text corpus of English, and might therefore make significant steps towards increased reliability and consistency, the policy of that dictionary was to reject the lexicographic tradition of formulaic definition style. Ironically, *CCED* found the Birmingham corpus to be a rich source of lexical facts which could be recorded in more detail and more reliably than previous dictionaries, but the resulting dictionary is probably less tractable as a basis for NLP lexicons because of its informal natural-language definitions.

Work carried out by the ILC in Pisa (Calzolari & Picchi 19XX) and the Lexical Systems group at IBM Yorktown (Byrd 1989) illustrates the approach taken by many NLP specialists who are extracting formalised lexical and semantic information from machine-readable dictionaries (MRDs). A certain degree of success has been achieved in encoding lexical and semantic relations (sets, hierarchies, networks) between lexical items by, for example, parsing the definitions and identifying the genus term. Conveniently, lexicographers tend to adopt a formal style for definitions which usually places a superordinate "genus" term at the head of the definition with a sequence of criterial qualifiers following. The words *hammer, drill* and *screwdriver* all share the genus term *tool* in OALD, for example. The restriction of the defining vocabulary in LDOCE seems to offer the MT researcher an even more formal basis for parsing and restructuring the semantic information contained in definitions—though the loose interpretation of the restricted vocabulary means that in practise LDOCE definitions are unlikely to be any more lucid.

I am not optimistic that the early successes in extracting semantic information in this way will be continued. Allowing the OALD to fall open at almost any page will reveal entries in which the use of a strict "genus term plus discriminators" definition is varied, stretched or rejected altogether in the interests of clarity and accuracy for the human reader. The definition of *structure* (sense 1) begins "way in which sth is put together,...". Parsing this definition will obviously identify *way* as the superordinate item, but this word doesn't have the character of a semantic primitive like TOOL. Moreover, *way* has two closely related senses: one referring to a state of existence ("that's the way I am") and the other referring to a process or method ("that's the way to do it"). This nice distinction is irrelevant to the user of the printed book, but possibly very important for the semantic description of the word *structure—structure* describes a state, not a process. Despite the features such as restricted vocabulary and formulaic style, definitions often pose very trick parsing problems. Scoping of co-ordinated constructions, for example, as in "frontage: extent of a piece of land or a building along its front, esp bordering a road or river".

The participants at this workshop will be more familiar than I am with the development of Lisp-style encoding, relational databases, Prolog-style databases and so on as representations of dictionary entries amenable to NLP systems, and I need not discuss them further.

The problem of representing semantic information is as important to lexicographers and dictionary publishers as it is to NLP researchers. Each individual dictionary records different sense categories for a significant subset of the headword list. Since no definitive encoding for a lexical database is on the horizon, it is not clear how consistency of representation can be achieved, nor whether it is desirable in reference tools for human users.

4. What are the size and composition requirements for a text corpus for deriving lexical data?

In the last year I have been devoting special attention to the design aspect of corpus building and it is clear that we are seeing a surge of interest in corpus-based linguistic study, in speech and NLP circles as well as in lexicography. A number of notable achievements in the use of text corpora have, I believe, acted to spur efforts in this area. In speech processing, the work of Jelinek and colleagues at IBM seems to be influential (REF); Church at ATT (REF) and Leech at Lancaster (Garside et al. 1987) have shown impressive results in stochastic methods for word-class tagging; Choueka at Bar-Ilan reports progress in identifying compounds (Choueka 19XX) and Sinclair at Birmingham introduced a corpus-based EFL dictionary (Sinclair et al. 1987).

The empirical approach to language analysis, whether the statistical methods employed are sophisticated or elementary, requires the collection of very large amounts of naturally-occurring data if reliable data is to be recorded. It is important to consider the relationship between the size and constitution of the corpus upon which linguistic analysis is to be based. It appears from the current work reported at this workshop that MT applications are showing encouraging rates of acceptability within restricted subject or discourse domains. This is good news for MT researchers who are intending to derive lexical data through corpus processing, since a corpus of, say, 10M words made up entirely of road traffic accident reports will show certain domain-specific linguistic features more frequently than a general language corpus of the same size. The sparse data problem is eased if the range of vocabulary and grammatical structures to be handled is significantly reduced.

How significant is the composition of the text corpus when assessing the value of lexical and grammatical rules or probabilities which are derived from it? There has been very little published research into the effects of varying the composition of a corpus, and I believe that more effort must be directed in this area. However, the evidence so far suggests that despite the strong feeling among lexicographers that a corpus of weather reports will furnish little useful data for the purpose of compiling a dictionary of general English, the statisticians and computational linguists are not concerned about the effects of skewed[1] distribution of domain-specific linguistic features. At this workshop we have heard about

[1] I use this word loosely, since we have no model of the true distribution against which to measure skew

IBM's use of the vast Canadian Hansard corpus, and both IBM and ATT have been using large amounts of text from the AP newswire. It may be that the size of the corpus is more significant than its composition, though of course the two parameters are not completely independent. The evidence against using corpora made up only from restricted domains as a basis for the induction of linguistic rules or estimation of probabilities rests merely on anecdotal examples. Richard Sharman noted that the mutual information approach to identifying English-French translation units resulted in an oddity whereby *hear* had two potential French equivalents *bravo* and the null string: such absurdities are clearly attributable to the peculiar nature of the corpus, but it is not clear whether noise of this type has a serious overall effect on the linguistic information which is extracted. Since the testing of probabilistic language models is often carried out over text from the same domain as the training corpus, the domain-specific effects are unlikely to show up.

5. Can manual effort in compiling dictionaries provide benefits for MT?

Dictionary compilation is still a very labour intensive operation. The widespread introduction of computer technology into lexicography is improving productivity and quality, but there is some way to go before any significant amount of linguistic analysis can be carried out automatically. Since publishers are investing time and money in the compilation of dictionaries, to what extent can the resulting resource be "re-used"?

The question should perhaps be posed the other way around: can the compilers of reference works for humans benefit from the advances in lexicon building for MT and other NLP applications? If the answer to this question is negative, then the potential for bringing together lexicographers and MT researchers in the common pursuit of lexicon building is seriously limited. The present situation is at a sort of stalemate. NLP systems developers want the results of the manual effort and skills which lexicographers have invested in their reference works, but the products are not adapted to NLP needs. Lexicographers want the results of NLP research (lexical databases, parsers, taggers, semantic formalisms) but similarly the resources that exist at present are ill-suited to needs of reference publishing. There are major initiatives in progress to bridge this gap. An ESPRIT project, GENELEX, involves IT companies and dictionary publishers in work to create a generic lexicon for NLP. A EUROTRA project definition study is starting with the aim to define priorities for the reusability of lexical resources. Published reports and surveys of the state of the art tend to see the task as being to bring a uniformity, explicitness and formal rigour to dictionaries such that they will be tractable for NLP systems.