**University of Brighton**

# *ITRI-00-15* Business Models for Dictionaries and NLP

Adam Kilgarriff

**June, 2000**

Also published in Int. Jnl. Lexicography 13 (2), pp 107–118

Supported by the EPSRC under Grant M/54971

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL:
firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

# Business models for Dictionaries and NLP

Adam Kilgarriff

January 22, 2001

**Abstract**

NLP needs dictionaries, and dictionary-makers can use NLP to make better dictionaries, so there is great potential for synergy between the two activities. To date, there has been only very limited collaboration. The two reasons for this are (a) dictionary publishers' concerns regarding intellectual property, and (b) the different languages that lexicographers and NLP researchers speak. In this paper I present a model for overcoming the first and suggest some strategies for the second.

## 1 Introduction

NLP needs dictionaries, and dictionary-makers can use NLP to make better dictionaries, so there is great potential for synergy between the two activities.[1] There is ample motivation for NLP to court dictionary publishers, and *vice versa*.

To date, NLP research has used dictionaries and dictionaries have used NLP, but the two processes have not been brought together. The NLP that has gone into making dictionaries has not been the NLP that was based on an earlier version of the same dictionary (or indeed of any published dictionaries).[2] The NLP has either been dictionary-independent if brought in from outside, or developed in house. While NLP groups have made innumerable corrections, improvements, additions and extensions to the dictionary databases they have licensed from publishers, these changes have never been used by the publisher to improve the next printing or the next edition of the dictionary.

[1] NLP (Natural Language Processing) is here taken to describe all those technologies that manipulate human language inputs by computer, examples being automatic part-of-speech tagging, parsing, concordancing, machine translation, information extraction and text generation. 'Language engineering' and 'computational linguistics' are near-synonyms.

[2] Sophisticated NLP software that has been used for English includes taggers (eg COBUILD used ENGCG from Helsinki; Longman, OUP and Chambers-Harrap used the BNC, which was POS-tagged by CLAWS, from Lancaster) and parsers (the experimental HECTOR project used the Fidditch parser (Hindle, 1990)). These programs do not use publishers' dictionaries. Nor do the statistics-based collocation-finders that have been quite widely used. An exception, among English dictionaries, is (CIDE, 1995), where the development of tagging software was closely integrated with the dictionary production process.

There have been two reasons for this failure of synergy. Firstly, dictionary publishers are concerned about a number of questions relating to intellectual property. Secondly, lexicographers and NLP researchers speak very different languages, so it is not straightforward to develop the atmosphere of trust in which the lexicographers fully understand what the NLP research has done or believe that the NLP 'enhancements' to their dictionary are useful to them.

In the remainder of the paper I first describe what the two sides have to gain from each other, then the history of the issue, and then present a mutually beneficial business model. I then comment on the language problem.

The paper may seem to replicate the 1991 Special Issue of this journal edited by Bran Boguraev (Boguraev, 1991). Indeed, the overall motivation is the same: to promote mutual understanding and collaboration between the two fields. But whereas the Special Issue set out to describe what benefits there could be, in this paper, this is all but assumed (with the following two sections summarising the main themes.) Rather, we look at the institutional and business reasons why more collaboration has not happened and consider how they can be overcome.

## 2 Dictionaries for NLP

NLP has cause to court the publishers because it needs lexical information for almost everything it does. Lexical information tends to be expensive and difficult to produce, so licensing it from those who have already invested in it — the dictionary publishers — makes good sense. This applies to the full range of lexical information: orthography, phonology, morphology, syntax, semantics, pragmatics, translations, domain, thesauri, collocations. All of these have, over the last twenty-five years, been extracted from electronic versions of dictionaries (on a typesetter's tape, or CD-ROM, or other electronic medium, or scanned or keyboarded) and used in NLP, for research in all cases, for products in most. NLP has found dictionaries valuable for all these purposes in spite of the overheads of finding the relevant information in what is usually a poorly-structured input, and the inevitable errors, inconsistencies and omissions. They would be much more enthusiastic if it were not for these failings.

## 3 NLP for better dictionaries

Dictionary publishers have cause to court NLP for a number of reasons. The most obvious is that there is money to be made from licensing arrangements. Large sums frequently change hands. This has been the dominant motivation for dictionary publishers approaching NLP groups. (A distinction must be made at this point between dictionary publishers and lexicographers. This motivation has been vivid in the minds of the publishers, but of only indirect interest to the lexicographers.)

The more intriguing reasons relate to the potential NLP has to improve dictionary quality. Of these, two varieties can be distinguished: those that act on a dictionary database, and those that act on text corpora.

## 3.1 Benefits of NLP applied to dictionary databases

The benefits are closely related to the benefits to the dictionary production process of, simply, having the dictionary in a database at all. Many errors can be detected automatically, and consistency of all manner of features checked. Superficially, these may seem to be simple database functions. However, many features can only be checked for consistency if the data model underlying the database is based on an understanding of lexicography. The skills required to model the data lie at the intersection of linguistics, lexicography and computation, viz, in NLP.

Moreover, there are always further checks and analyses which go beyond anything the database will provide directly. For example, one might want to check to what extent the words in the same thesaural category get corresponding definitions. Answering this will require NLP expertise, as well as both a well-structured database (so the relevant data is accessible) and lexicographic expertise (to work out where the non-correspondences are justified).

A common situation is that a good database model is available but the dictionary database does not fit it. What is then required is a process sometimes called "up-translation", of carrying the data over from the existing format into the new one. Standardly, much of this will be doable by simply translating mark-up, but in many cases, the mapping will be one-to-many, or will be context-dependent, or distinctions will be required for the new data model which are implicit in the text or font-changes of the existing dictionary. In most such cases, a very high percentage of the up-translation can be done automatically, but it is an NLP task to do it.[3]

## 3.2 NLP applied to text corpora

The arrival of electronic text corpora is causing a revolution in lexicography. Previously, the primary source of evidence for a word's behaviour was the lexicographer's intuition. Now, wherever publishers have been able to assemble a large text corpus (and have concordancing technology) it is the corpus. There is widespread agreement that this is a huge boon for lexicography.

Yet it clearly introduces many new challenges. Organising the lexicographer's working environment so that s/he has instant access to a concordance for the word s/he is working on is not easy (particularly where lexicographers work at home). Even the simplest concordancing program requires some minimal NLP input. Minimally, it needs to know which characters are punctuation characters,

[3]Two EU projects which have investigated up-translation in various ways are CONCEDE (`http://www.itri.bton.ac.uk/projects/concede`) and DicoPro (`http://issco-www.unige.ch/projects/dicopro_public/`).

so should be ignored when defining what a word is: if a space is treated as the only word-delimiter, the search term, "`butter`", will not match the corpus object, "`butter,`". (In English, the characters ' and - present problems at this level.) Morphological analysis will make the corpus more useful, as then all forms of a word can be found with a simple search. (For English, with its simple morphology, it is easy enough to search for all four forms separately, or with a regular expression. For Finnish, it is not.) The corpus will be more useful again if it is part-of-speech tagged, word-sense disambiguated or parsed (see eg (Tapanainen and Järvinen, 1998)). Each level of annotation allows the lexicographer more control in searching the corpus for the linguistically interesting phenomena, so that, when looking through the concordance, noise is reduced and duplicating patterns do not need to be scanned exhaustively.

Beyond the concordance, there are corpus statistics and machine learning. Statistically-organised collocation lists have proved their worth for all the dictionary projects that have had access to very large corpora. ((Church and Hanks, 1989), the paper that opened the current debate in NLP about collocation statistics, is a noteworthy example of lexicography/NLP collaboration.) Acquiring lexical information from corpora is currently one of the most dynamic areas in NLP. My purpose in saying this is not to put fear of redundancy in the hearts of lexicographers but to indicate how much more satisfactory their work is to become, when the tools at their disposal are so much more powerful. The techniques tend to find many plausible hypotheses for how a word behaves in a corpus, but are unable to sort the wheat from the chaff, or, evidently, to assign meanings to the patterns they find. The lexicographer's task is as before but with less drudge.

# 4 History

The recent history of dictionary/NLP interaction begins with (Amsler, 1980) and (Michiels, 1982). Each took a typesetter's tape for a dictionary. Amsler's agenda was to see whether the dictionary could be used as a source of general knowledge of the everyday world — knowledge that, for example, Alsations are dogs — for use with artificial intelligence programs. Michiels explored how the more specifically linguistic information might be used for NLP. Both of these agendas were pursued at length in the course of the 1980s, notably in the EU ACQUILEX project, at the Computer Research Laboratory at New Mexico State University, and at IBM in Yorktown Heights. (Boguraev and Briscoe, 1989) represents the activity at its high water mark, with (Byrd et al., 1987) demonstrating the ascent, (Wilks, Slator, and Guthrie, 1996) reviewing the whole, and (Ide and Veronis, 1993) offering a post mortem (their subtitle is: Have we wasted our time?).

As a research topic, the use of dictionary databases is less active than it was. This has a number of interpretations. The most positive is that the NLP world has now developed a fair understanding of what information dictionary databases contain and what is involved in extracting it for NLP use, so that the topic has shifted from 'research' to 'development'. (This would offer a perspective on why

almost all the research was on English dictionaries: they were used for testing out methods, and the methods could then be used for dictionaries for other languages.)

The next interpretation is simply that the research fashion has changed, particularly to language corpora (and methods for extracting lexical information from them).

Then there is the interpretation that motivates this paper. In the academic sector, work done to enhance a dictionary is usually work wasted if others are not permitted to use the enhanced resource. Publishers, anxious about intellectual property, have frequently not permitted it. A number of NLP workers have explicitly chosen not to do any further work on dictionaries (except those in the public domain) for this reason: they suspect that any such work, however good, will be destined for oblivion.

In the event, the academic world's own product arrived. As of 1990, WordNet has been available free, over the web and without constraint. WordNet has been the lexical resource of the 1990s, and various people have argued that WordNet senses are the *de facto* standard for NLP. WordNet was produced by linguists and psychologists, according to a psycholinguistic agenda, and its suitability for NLP research remains a lively topic of debate. But, as against the practicalities of getting hold of it, these purist concerns have carried little weight and everyone uses WordNet now.

WordNet addressed semantics: for syntax, the NLP community invested in COMLEX (and, more recently, NOMLEX) (Grishman, MacLeod, and Meyers, 1994; MacLeod et al., 1998).[4] COMLEX and NOMLEX are designed from the outset to meet the needs of the NLP community and there is currently work in progress on linking the COMLEX syntax patterns to the WordNet word senses.

The 1998 Euralex conference in Liège may have inaugurated a new phase in the debate, with the impetus this time coming from lexicography. Whereas earlier EURALEXes have been resistant to computation or viewed it warily, at Liège, this author's perception was that it was universally accepted that NLP had a role to play in dictionary production.[5] The question was no longer whether to use it, but how to use it well. Except in the 'dictionary use' sessions, it was hard to find papers which did **not** assume the availability of corpora and concordancing software or more. One paper of particular salience to the argument was jointly presented by Ulrich Heid, an NLP academic and Vincent Docherty, a dictionary publisher (Docherty and Heid, 1998): here, at last, the state of the art in NLP was being used to provide inputs to lexicographers for compiling a better dictionary for people.

[4] There are alternatives to COMLEX. Other lexicons such as XTAG (Group, 1998) and ANLT (Carroll and Grover, 1989) have also been developed by the NLP community. (The ANLT lexicon was developed in the mid-eighties, it included material from a dictionary and copyright issues prevented it being used widely for NLP until the mid-nineties.)

[5] The longstanding engagement with NLP of the hosts in Liège may well have played a role.

# 5 Publishers' anxieties

The overt reason why publishers' dictionaries have not been more widely used is copyright. The publishers' argument is simple and direct: the publisher's trade is in intellectual property, so it is not reasonable to expect them to give the dictionary away or risk it falling into the public domain.

There are several threads to the case, which I first present, and then respond to.

The first is, simply, piracy. If a dictionary starts being copied freely, or, worse still, starts being copied for a fee but where the fee does not go to the legal copyright owner, the publisher is the loser. The publisher wants to avoid this happening above all else. A licence agreement with an NLP research group provides an avenue by which a dictionary may find its way to being illegally copied and re-copied.

The other threads relate to the possible re-use or re-sale by the publisher of versions of their dictionary which have been upgraded in some way by an NLP research group.

The issues here concern contamination. Where a dictionary publisher is the sole owner of the copyright for a dictionary, it would like to keep it that way. There will be no other parties to consider in future negotiations, and all the profits will come to the publisher. If a dictionary database has been enhanced by an NLP group, then, *prima facie*, a share of the intellectual property belongs to the NLP group. It is not straightforward to arrive at a model for how shares of the intellectual property should be allocated. The starting point would usually be the quantity of labour that each party had put in, or the fraction of the text that each produced. The latter cannot be applied at all: whatever the NLP enhancements may be, it certainly will not make sense to quantify them as a fraction of total text-length. The former is also of little use: the NLP laboratory will probably have been using various pieces of software and expertise to aid with the enhancement, which will have been undertaken for internal use, in other NLP applications or research projects, so the re-sale or re-licensing of the upgraded dictionary would be a side-effect.

The publisher may well have doubts about the accuracy and consistency of the enhanced dictionary, but may not be well-placed to evaluate it, as this in itself may require NLP techniques and an understanding of the issues which are likely to be critical for NLP applications.

The enhanced dictionary may have potential uses for print products, for electronic products for the consumer market, or for licensing on for NLP use. A question in relation to the first two is addressed in section 7. Regarding the third, the potential complexity of the arrangements is forbidding. Where a dictionary is licensed for NLP use, the licence may be, broadly, for research use or for product development. Research use is straightforward, provided the demarcation in the NLP group between research and development is clearly drawn. Where it is for product development, the publisher may receive a licence fee or a royalty for each product 'containing' the dictionary, or some of each. 'Containing' in inverted commas, because the dictionary will not generally exist in the product in any recognisable form, but as one of a number of inputs to the system's lexicon, which may well

be compiled and unreadable. That product may itself be a consumer product, or may be a component of some larger application, so there may be many steps, each potentially involving licences and royalties, between the dictionary publisher and the end-product. One has some sympathy with the publisher not wishing to have its negotiating hand constrained by undertakings to NLP groups.

## Responses

Piracy is not a strong argument. There are many ways in which a dictionary may be pirated or may enter the public domain. It may be re-keyed, it may be OCR'd from the printed version, the contents of CD-ROMs may be unscrambled, hackers may hack into the publisher's system, tapes may be stolen.[6] It will of course be the duty of licence-holders to ensure that people who should not have access to their version of the dictionary, do not, and that it does not get copied outside the confines of the laboratory, and if they fail in this duty they will be culpable. But they are not dramatically more likely to fail in the duty than assorted other of the publisher's employees, agents and licensees who have access to the data. [7]

Regarding copyright on an enhanced dictionary, firstly, many research groups in Universities will not want to claim a share of copyright. The original licence allowing them to use the dictionary could then contain a clause stipulating that lexicons for which the dictionary has been an input can only be passed on to third parties by the publisher, and all the copyright in such lexicons shall be vested in the publisher.

Secondly, while some research groups may be unwilling to relinquish copyright for free, they may well be likely to do so for a fee.

Where NLP groups are not happy to hand over copyright in this way, there will be potentially complex negotiations required, as there will when a lexicon is a component of some other application at one or more further removes from the end product. These sorts of considerations are, however, becoming commonplace in the software and multimedia industries, so the publisher's concerns reduce to the following: are the gains to be had from synergy with NLP sufficient to merit the effort? Is the publisher willing to take on the challenge of the complex negotiations?

[6] It may also be photocopied. In Bali, getting a dictionary photocopied is an inexpensive and commonplace procedure. The copying was of good quality, with the cover copied in colour, and the most notable difference between original and copy being one of bulk.

[7] As one reviewer pointed out, research organisations may be places where some individuals have an ideological commitment to free information, and others simply find copyright considerations irritating and low-priority. Also, in general, the risk of piracy is minimised by the publisher limiting the circulation of the electronic form of the product as far as possible. Computer science students and researchers are also particularly likely to be able to decipher encrypted texts. But to say that, is to say little more than that no business ventures are risk-free; each different scenario brings with it its own particular risks. It is the task of the publisher to weight the risks against the potential benefits.

# 6 The model

The primary concern of the publisher is to retain control of the resource and to maximise the flow of income it gives rise to. The primary concern of many NLP researchers is that their work is available for others to use and extend. (Reputation and status is dependent on being cited, in the academic world. If people use your resource they will cite you. Licence income is a lesser consideration, firstly, because it it is not expected, secondly, because it would probably go to the institution rather than the individual, and thirdly, simply because it is not the currency of academic status.)

The appropriate model is for the publisher to encourage NLP researchers to use the dictionary, on the basis that any enhancements to the dictionary will be returned to the publisher, for it to use itself if it so wishes (for dictionary revision and other improved consumer products) and to market and generally make available to other NLP groups. The agreement would put the publisher under an obligation to make the enhanced dictionary available (under similar terms again, and for a fee that would not be prohibitive) and this would be the benefit to the NLP group to counterbalance the fact that they would, in the simple case, relinquish claims to a share in the copyright.

Limitations are required on this obligation of the publisher to publish. Firstly, the enhanced dictionary would have to be of adequate quality, and to this end, the publisher must acquire the expertise to assess the quality.[8] This will demand some investment in NLP expertise on the part of the publisher, but then, the publisher should not expect to reap the fruits of the new market without any investment.

Secondly, the publisher will want to strike different deals for the dictionary with different customers: one would not expect terms for Microsoft to be as for a University group or startup company. The terms of the agreement would have to allow flexibility, with the obligation-to-make-available probably only covering agreements with non-profit organisations for research use only.

Thirdly, there might be a number of enhanced versions of the same dictionary, from the same, or different, NLP groups. The publisher might reasonably require the different versions to be consolidated, so that it did not have slightly different and mutually incompatible products on its list.

The NLP group will not always be willing to relinquish copyright, and in that case the negotiations regarding possible future licence fees and royalty income will inevitably be complex but the principle — that the publisher publishes and licences the enhanced version — remains the same.

The model does not impose any unfamiliar demands on publishers. It merely supposes that they identify the NLP publishing function as one that is worth investing in, and that, in pursuing it, they bear in mind the particular interests of academics, who are their likely collaborators, and the particular characteristics of

[8] Some guidelines for assessing dictionary quality, from an NLP perspective, are available from the European Language Resources Association: `http://www.icp.grenet.fr/ELRA/validat.html`

dictionaries. Amongst dictionaries, they should pay particular heed to the character of dictionaries for NLP use, notably that they are the kinds of objects that potentially accumulate greater and greater value, with each accretion not only adding value in itself, but contributing, like compound interest, to future additions.

# 7 The out-of-date dictionary problem

In the best of all possible worlds, computational enhancement and lexicographical upgrading would build upon each other in a virtuous circle that knew no bounds. There are, however, snags. One is timing. If the enhanced dictionary is based on, say, a dictionary first published in 1990, which the NLP group started using in 1994 and worked on until 1998, then the publisher may well consider the underlying analysis too old to be worth considering as a basis for further work starting in 2000. This was a poignant issue for Longman, whose 1978 first edition of LDOCE has been very widely used for NLP, with NLP groups still licensing it into the mid 1990s. By that time the (pre-corpus) lexicographic analysis was of purely historical interest from an EFL dictionary publishing perspective and two further editions had appeared in the meantime.

There are various reasons for starting a new dictionary from scratch. The lexicographic team is not encumbered by the prevailing philosophy, and can work out a new perspective from the outset. It encourages the lexicographers to look afresh at the evidence. It means the words "new" and "completely revised" can be blandished across the cover without fear of contradiction.

If a new dictionary is written from scratch then it cannot readily benefit from enhancements made to its predecessor. It may be possible to automatically import some enhancements, but probably only for data that applies at the 'headword' level, not at the word sense level: a major rewrite will involve re-analysis of word meaning; new sense distinctions will frequently not coincide with old ones; and allocating sense-specific information from the senses in the old dictionary to senses in the new will be a difficult manual job.

Concerns about enhanced dictionaries being out of date by the time they make their way back to the publisher are valid, and may mean that, in the future, dictionaries for NLP and dictionaries for people do part company.

# 8 Different worlds

The gulf of understanding between lexicographers and NLP researchers is not to be underestimated. It often relates to level of detail. The NLPer is looking for generalities whereas the lexicographer has a profound awareness of the level of idiosyncrasy in the lexicon. Asked to "take a verb", the NLPer is likely to offer *make* or *break* — these are after all the common items that serve as prototypes for the class. The lexicographer, painfully aware of the often atypical behaviour of very common items, and the sheer number of verbs there are to choose between,

might offer *accost* or *simmer* or *handcuff* or *vary*, depending on which stretch of the alphabet they were last working on.

Tasks which seem to the NLPer obviously in need of automation continue to be done by people in dictionary publishing. Sometimes this is the outcome of undue conservatism, but more often, the dictionary publishers find it easier, cheaper, and more dependable to employ people than to acquire and install software which is not extensively tested and may not be robust, which needs to be integrated into their setup, and which is unlikely to cover all the cases (so there may well be an extensive post-editing function as well).

Lexicographers write dictionaries for a living, whereas researchers write research papers. For the researcher, the natural end point of an activity is to write it up, and that is central to what they are paid for and gain status for. For a lexicographer, any writing up is likely to be an activity for evenings and weekends. Many lexicographers work freelance, making the cost of writing up all the more apparent: it amounts to paid work foregone. The economic difference results in a difference of perspective at various points. Anecdotal evidence of NLP researchers spending time visiting lexicography departments (in, eg, the ILD project[9]) is that they found what was going on bewildering.

Lexicographers' core task is analysing and describing meaning, a task on the arts side of the arts/science divide. NLP is firmly on the science side. NLP sometimes approaches problems of lexical description with more formal or quantitative approaches than the lexicographer can happily apply: where lexicographers try to apply the formal systems, they are likely to find it necessary to stretch the meanings of the categories so far that the NLPer no longer recognises the system.

Lexicographers and NLP researchers come from very different cultures. If there is to be collaboration between the two worlds, it is necessary to allow for the difficulties that communication between cultures will always present.

# 9 Conclusion

Collaboration between NLP and dictionary publishers offers great benefits to both sides. However, there are hurdles to be overcome. In addition to cultural differences, there are some specific issues regarding copyright. In this paper we disentangle the interests of publishers and NLP research groups, establish that they are compatible, and present a business model which allows both parties to get what they want.

## Acknowledgements

I would like to thank Nicolas Dufour, Thierry Fontenelle, Andrew Harley and Mary O'Neill for comments on earlier drafts of this paper. The work was supported by EPSRC grants K/18931 (SEAL) and M/54971 (WASPS).

[9] `http://www.ltg.ed.ac.uk/projects/ild`

# References


Amsler, Robert A. 1980. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph.D. thesis, University of Texas at Austin.

Boguraev, Branimir K., editor. 1991. *Special Issue: Building a Lexicon*, volume 4(3).

Boguraev, Branimir K. and Edward J. Briscoe, editors. 1989. *Computational Lexicography for Natural Language Processing*. Longman, Harlow.

Byrd, Roy J., Nicoletta Calzolari, Martin S. Chodorow, Judith L. Klavans, Mary S. Neff, and Omneya A. Rizk. 1987. Tools and methods for computational lexicology. *Computational Linguistics*, 13:219–240.

Carroll, John and Claire Grover. 1989. The derivation of a large computational lexicon for english from LDOCE. In Branimir K. Boguraev and Edward J. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman, Harlow.

Church, Kenneth and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, pages 76–83, Vancouver.

CIDE, 1995. *Cambridge International Dictionary of English*. CUP, Cambridge, England.

Docherty, Vincent and Ulrich Heid. 1998. Computational metalexicography in practice – corpus-based support for the revision of a commercial dictionary. In *Proc. EURALEX*, pages 333–346, Liège, Belgium, August.

Grishman, Ralph, Catherine MacLeod, and Adam Meyers. 1994. Comlex syntax: Building a computational lexicon. In *COLING 94*, Tokyo.

Group, XTAG Research. 1998. A lexicalized tree adjoining grammar for english. Technical report, IRCS, University of Pennsylvania. http://www.cis.upenn.edu/ ircs/reports/trs/abstracts98.html.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *ACL Proceedings, 28th Annual Meeting*, pages 268–275, Pittsburgh.

Ide, Nancy M. and Jean Veronis. 1993. Extracting knowledge bases from machine-readable dictionaries : Have we wasted our time? In *KB&KS Workshop*, pages 257–266, Tokyo.

MacLeod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: a lexicon of nominalisations. In *Proc. EURALEX*, pages 187–194, Liège, Belgium, August.

Michiels, Archibald. 1982. *Exploiting a Large Dictionary Database*. Ph.D. thesis, University of Liège, Belgium.

Tapanainen, Pasi and Timo Järvinen. 1998. Dependency concordances. *Int. Journal of Lexicography*, 11(3):187–204.

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