[From: *Current issues in computational linguistics: in honour of Don Walker*, ed. Antonio Zampolli, Nicoletta Calzolari, Martha Palmer (Linguistica Computazionale, vol. 9-10); Pisa, Dordrecht, [1994]

Acquisition and Exploitation of Textual Resources for NLP

Susan Armstrong-Warwick
ISSCO
University of Geneva
e-mail: susan @ divsun. unige. ch

Abstract

Electronic access to large collections of texts and their translations provides a new resource for language analysis and translation studies. Empirical and statistical methods offer the means to organize the data and develop alternative models in view of a better understanding of our use of language. From a practical point of view they provide a basis for progress in the performance of NLP systems. A prerequisite for this work is the availability of machine-readable texts in an appropriate format. This paper will present current initiatives to acquire and prepare the necessary textual resource for corpus-based work and review current methods under development to exploit the data.

1 Background

For a growing number of researchers, electronic access to large collections of texts and their translations has become an essential resource for language analysis and translation studies. Empirical and statistical methods are being developed to organize the data in order to elaborate more adequate models of the structure and use of natural languages. Reliable methods for English are now available to tag texts for part-of-speech, predict word sequences, recognize collocations and automatically align sentences with their translations. These methods offer a starting point for deeper studies and practical applications in varying fields such as lexicography, speech recognition and machine(-assisted) translation.

This quite recent and growing interest in corpus-based studies is somewhat reminiscent of the empirical and statistical methods popular in the 50s. Initial work on machine translation (MT) – one of the first computational linguistic applications – was then related to problems of code-breaking (Weaver 1949). However, the computing resources were far from adequate and the textual resources, necessary as a basis for the statistical models, did not exist. Technological advances in computing power have certainly favored the reintroduction of this approach, as has the growing availability of electronic texts.

Another important factor which has contributed to the interest in data-oriented methods is the realization that rule-based systems have not produced the desired results nor the hoped-for basis for future progress. However, this new direction has also brought forth

a number of critics, a debate which can be characterized in terms of 'statistics-based vs rule-based', 'empiricist vs. rationalist' or 'Shannon-inspired vs. Chomsky-inspired'. Ironically, this controversy was also very much discussed in the 50s when Bar-Hillel (1951), arguing for the importance of semantics, wrote:

> Let me warn in general against overestimating the impact of statistical information on the problem of MT and related questions. I believe that this overestimation is a remnant of the time, some ten years ago, when many people thought that the statistical theory of communication would solve many, if not all, of the problems of communication, (p. 172)

This very same communication model has, in fact, been revived in current MT work (Brown et al. 1988) and with it, the same debate. However, most researchers engaged in corpus-based studies do not regard statistics as a total solution to the problem of describing language but rather as a means of modeling what occurs in texts as part of different computational applications. One important direction in the field is to integrate probabilistic models into rule-based systems and conversely, to augment statistical models of language with more traditional linguistic information.

Based on recent work, it has become clear that the new data-oriented methods offer potential solutions to key problems in computational linguistics:

- **acquisition:** identifying and coding all of the necessary information
- **coverage:** accounting for all of the phenomena in a specific domain, a given collection of texts, an application, etc.
- **robustness:** accommodating 'real data' that may be corrupt, ungrammatical or simply not accounted for in the model
- **extensibility:** applying the model and data to a new domain, a new set of texts, a new problem, etc.

While these problems are not new, access to large text resources does offer the means to investigate new directions that promise some important progress in the field theoretically and also to help solve very practical problems such as building dictionaries, classifying proper names and unknown words, and identifying noun phrases and other collocations.

In what follows we will present some general issues in acquiring and preparing corpora and report on a number of data collection activities currently in progress in Europe and North America. We will then review a number of studies which have shown how this data can be exploited.

2 Availability of Textual Data

In the latter half of the 1980s, when interest in statistical methods and corpus-based work was emerging in the computational linguistic community, there was very little material widely available for research purposes. This situation is in contrast to the

speech community where probabilistic models and statistical methods had become the standard and where data gathering was thus considered an integral part of any project (cf. Church and Mercer (1993) for a discussion of the development of statistical methods in speech research and its effect on work in computational linguistics and Liberman (1992) for some case studies in how publicly available corpora have benefited the speech community).

Older corpora for English, such as the Brown Corpus (see Francis and Kučera 1982) were quite small by current standards and others such as the Birmingham Corpus (see Sinclair 1987) were not publicly accessible. In continental Europe, where the new interest in corpus-based studies has only recently emerged, the situation is similar: the texts held in the national language centers are either too expensive for the individual researcher, not accessible in a manner conducive for current methods[1] or simply not available to the public.

Though there is a vast potential amount of data in electronic form, little of this material is currently available to the research community. The texts are privately held in centers all over the world and the holders of the data are often printing houses that do not have ownership and distribution rights. This separation of holders and owners is also apparent in large organizations where the technical services managing the archives are quite separate from other departments. Simply identifying where the data is located is often a problem itself once the texts have been printed (and are thus no longer in use).

The lack of appropriate textual materials (in quantity and range of data) has restricted research work in various ways. For languages other than English very little material is available; thus work has concentrated on the English language and methods have been tailored to take advantage of some language specific phenomena, e.g., fixed word order and limited morphology. It remains to be seen how far these can be extended to other languages. Translation studies up to the present have concentrated on essentially one language pair and one text type due to the public availability of only one corpus.[2] The proprietary nature of much of the data currently in use has meant that work was often duplicated since sharing results was discouraged.

Fortunately, this situation is slowly changing and it is this progress we wish to document here. A number of initiatives (cf. below) have served to increase the awareness of the desire and need for public access to data and to demonstrate the interest in cooperating in the acquisition and preparation of these resources. The data that has been made available through initiatives and [some individual efforts] has tended to be a rather ad-hoc collection. The community has been working under the motto that almost any data are better than no data and certainly, the more, the better. However, once large amounts of texts do become available, the issue of how to construct a 'balanced' or 'representative' corpus will have to be addressed – what Walker (1991) has termed the "ecology of language".[3] In order to provide adequate coverage of language at a given time or for a given domain, we will need to consider matters such as style, register, text type, frequency, etc., Biber (1993).

[1] Many centers offer remote access through in-house query programs whereas current practices require that the entire text must be available for manipulation in one's own laboratory.

[2] The Canadian Parliamentary Debates referred to as the Hansard Corpus, available from the ACL/DCI.

[3] See Walker's paper in this volume. This topic is of central concern to all corpora developed for lexicographical work.

One important issue that any data collection enterprise must address is how to protect the interests of the originators of the texts – a matter of critical concern in this new electronic era. Whereas texts are normally acquired and consulted for their information and amusement value, the use of texts in corpus-based studies is quite different. The interest is in the use of language rather than the content of a given document. This view of texts measured in *kilobytes* rather than *content* is often difficult to explain to the data holder who views texts in terms of copyright issues, scientific, artistic or popular value or simply as a potential source of revenue. And unlike the past, when a research environment simply meant access to a well-endowed library (or inter-library loan system) and adequate computing resources, for corpus-based studies each research group must have a personal copy of all of the material.

In light of this situation, all data collection enterprises make formal agreements with the data providers and those who wish to use the data. In the case of the organizations described below, each applicant for data must sign an agreement not to redistribute the data and to respect all restrictions as stipulated by the data providers. Though many issues of access, copyright and data protection in general (e.g. sensitive or private material, text collections and derived data as a potential source of revenue, etc.) are still in need of clarification, these agreements provide the legal basis to guard against misuse.

3 Data Collection Initiatives

We now turn to a brief description of the new text collection and distribution activities that have emerged over the past few years. We begin with the largely volunteer efforts and then look at the later official projects that will assure a sounder structural basis.

3.1 ACL/Data Collection Initiative

The first such initiative, *the ACL Data Collection Initiative (ACL/DCI)* was established in 1989 by the Association for Computational Linguistics. The ACL provided the aegis of a not-for-profit scientific society to oversee the acquisition and preparation of a large text corpus to be made available for scientific research and without royalties. The acquisition work was carried out on a volunteer basis in a somewhat opportunistic manner, relying on availability rather than concerns of balance or representativeness. The clean-up and preparation (minimal SGML mark-up) of the material was done by a few individuals (Liberman 1989).

In 1991 the ACL/DCI produced and distributed its first CD-ROM and hundreds of sites are now working with this data. The disk contains over 600 Kb of mostly American English data and includes a large collection of newspaper articles from the Wall Street Journal, a dictionary of English donated by Collins Publishers and some grammatically annotated data from the Penn Treebank Project (Marcus et al. 1993), among others; a second CD-ROM is currently in preparation.

3.2 European Corpus Initiative

A similar initiative was established in 1991, the *European Corpus Initiative* (see Thompson 1992), to acquire a large multilingual corpus for research work in Europe. In partic-

ular, emphasis was put on gathering texts in languages other than English to provide the basis for researchers in all European countries to work on their own national language. An additional goal was to acquire a set of parallel texts (texts and their translations) in light of the importance of multilingual document production in Europe and the interest in translation studies.[4]

A large amount of data has now been collected for most European languages, with at least 5 million words of text for each of the major languages. A variety of parallel corpora have also been acquired from international organizations and swiss banks (English, French, Spanish and English, French, German, respectively). The texts are currently being prepared and will be available on a CD-ROM by the end of the year.[5]

3.3 Establishing Text Repositories

The two aforementioned initiatives have been singled out as exemplary for a new direction to meet the needs of researchers in the computational linguistic community. These volunteer efforts are now slowly being followed up by official projects which should establish a funding basis and the proper infrastructure for a longer term development of these resources.

3.3.1 Linguistic Data Consortium

In the United States the *Linguistic Data Consortium (LDC)* was established by the federal government in recognition of the necessity to follow up the largely informal efforts with a sounder structural basis. Another concern was to provide the resources to all researchers, not just those in large and private laboratories (who already had access to in-house data and/or a budget to acquire and prepare private collections). The LDC was founded in 1992 with an initial start-up grant from the Advanced Research Projects Agency (AREA) to acquire, prepare and distribute material for the research community (Liberman 1992). In less than one year the LDC has produced nearly 100 CD-ROMs and is actively working on acquiring a great deal more data. One of the major goals for the next year will be the acquisition of multilingual text to support machine translation and other activities.[6]

3.4 Multinational Efforts

In Europe, where the multilingual environment poses special problems for centralized action in this field, work has begun on defining a framework for further actions regarding building up textual resources in Europe. An initial feasibility study was carried out under a project called the *Network for European Corpora* (NERC). A follow-up project intended to establish the appropriate infrastructure for the collection of texts and the

[4]The work has been sponsored by the *European Chapter of the ACL (EACL)*, the *European Network in Language and Speech (ELSNET)*, the *Network for European Reference Corpora*, and the *Linguistic Data Consortium (LDC)*. Most of the work has been carried out at HCRC, Edinburgh and ISSCO, Geneva.

[5]The CD-ROM will be pressed by the LDC; distribution in Europe will be assured by ELSNET (email contact: elsnet@cogsci.edinburgh.ac.uk) and in the US by the LDC.

[6]Contact: The Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104-6305; email: ldc@unagi.cis.upenn.edu

distribution of the data in Europe will begin next year under the CEC funded project *RELATOR.*[7]

Another project to begin this year (in follow-up to the *ECF)* is the collection and preparation of a large multilingual corpus (Thompson 1993). The corpus will consist of a set of comparable polylingual documents in at least six European languages (newspaper articles in the field of finance) and a multilingual parallel corpus in all nine languages (most likely drawn from the Official Publications of the European Community).[8]

4 Data Preparation

Aside from the basic problems of acquisition and negotiating rights for distribution making the data useful often requires a good deal of effort to 'clean-up' and reformat it. Simply having data in electronic form is not necessarily sufficient, though in the future, as electronic document publishing evolves and mark-up and coding standards are established, this problem may disappear. Given the amount of time this work currently implies in any corpus collection activity - a situation that is likely to continue for perhaps a decade or more - it is not a task to be underestimated.

Older texts which were prepared uniquely for printing are usually stored on tapes in an undocumented and complex format and the correspondence between the logical structure of the text and the typographical structure is often not easy to establish. The large collection of texts from the United Nations, recently acquired by the LDC is but one example of this (Graff 1993). The documents for English, French and Spanish were archived on tapes made by the Wang computer system, an efficient means for storage but not for automatic extraction of all the files. Extracting the actual text data from the tapes required considerable effort (with help from Wang itself) to decipher the system specific character coding, format control codes and file structure. The recovery of the parallel texts could only be done semi-automatically due to the somewhat ad-hoc filename conventions coupled with numerous human-introduced errors. These problems arise from the fact that designers of older systems did not foresee such an application: the mark-up language was developed for physical display purposes only, rather than for logical representation of the information.

Beyond concerns of text mark-up schemes for formatting of texts is the issue of standards for annotation, i.e., additional, interpretive mark-up added to the data. Given the relatively little amount of data widely available and the current explorations of what information can reliably be identified in texts, it is not surprising that each corpus project has adopted in-house conventions for, e.g., sentence and word marking, part-of-speech-tagging and phrasal bracketing. As long as the mark-up is clear, well-documented, unambiguous and easy to convert for local machine processing, different conventions may suffice for these tasks.

However, as the information associated with the data becomes more complex, the standards, or conventions, adopted do become an issue. One major international project,

[7]The two projects NERC and RELATOR are carried out under contract to the CEC, DG-XIII, Luxembourg; contact: Roberto Cencioni, Jean Monnet Bldg., 2920 Luxembourg, or Nino Varile, email M444@eurokom.ie.

[8]The project is part of the CEC International Scientific Cooperation Program.

the *Text Encoding Initiative,* has been working on a set of guidelines for coding material for all types of text mark-up with special attention to the complex needs of humanities researchers. [9]

Working with texts in a multilingual environment also raises a number of issues not necessarily apparent when only working with one language. The interpretation of a given symbol may be different for a given language (e.g., alphabet code conventions). The problem is more serious when some information associated with textual data in one language does not have an equivalent in another language. Studies are currently under way to determine to what extent the essentially English-based tagging systems in use can be adopted to European languages which display a wider range of morpho-syntactic phenomena (Monachini and Östling 1992).

As more hand-corrected data are prepared with sophisticated linguistic mark-up, an expensive and time-consuming task, annotation standards become an issue. To promote the sharing of resources and comparison of results, common coding schemes become a desirable goal. One major European project, *MULTEXT,* which plans to make a large multilingual, (partially) hand-validated corpus available with annotations for logical text structure, sentence marking, tagging and alignment of parallel texts, will address this issue in a systematic way.

These few remarks on text preparation and mark-up point to a large range of issues that will have to be confronted as more data becomes available in a wide variety of languages. Low-level issues of character sets and text formatting codes are in need of standardization to enhance international exchange of data. For higher-level mark-up it is perhaps premature to look for any standardization in the field. As new methods evolve and are applied to the data and as these results are shared, new conventions and standards will certainly emerge.

In the remainder of this paper we will review the various corpus-based studies currently under development.

5 Exploiting the Data

The increasing range of new methods being developed to exploit the data can be followed in the rise in publications and the number of tutorials and workshops dedicated to this topic. Whereas in the 80s, a large proportion of research work in computational linguistics concentrated on improving (unification-based) grammar formalisms and extracting data from machine-readable dictionaries, the publications of the 90s are witness to the new interest in data-oriented approaches. The journal *Computational Linguistics,* for example, recently devoted a large two volume special issue to "Using Large Corpora" and the theme of the conference on Theoretical and Methodological Issues in Machine Translation '92 was 'empirical vs. rationalist methods'. Workshops and tutorials addressing these topics are now held regularly in conjunction with the major conferences on NLP.[10] This approach has also become the main focus of all work under the AREA

[9] An initial set of guidelines was published in 1991, a more comprehensive version will be available this year. Contact: tei-l@uicvm.bitnet

[10] In the following sections we can only mention a few of the numerous studies currently underway. The interested reader is referred to the collections given in the references to the cited papers. Cf. the tutorial by Liberman and Schabes (1993) for a topical bibliography.

program.[11]

5.1 Part-of-Speech Tagging

In contrast to the in-depth studies of grammatical phenomena in limited domains, data-oriented methods focus on (statistically) easily observable phenomena that can be determined with certain reliability over large quantities of data. The new methods aim at total (though perhaps superficial) coverage. The most-well established of these methods is that of part-of-speech tagging (e.g., Church 1988 and Cutting et al. 1992[12]). Given a sequence of words as input, these programs assign a sequence of part-of-speech tags with a very high success rate. The programs consist of a lexical component to assign a set of potential tags to each word and a component to disambiguate over sequences of tags (based on n-gram models that compute the probability of a tag given a previous sequence of tags). These taggers serve as the basis for a wide range of subsequent tasks, e.g., as a pre-processor for a parser (cf. Hindle and Rooth 1993, Marcus et al. 1993), as a basis for identifying phrasal expressions (Church 1988, Cutting et al. 1992, Smadja 1993), and in applications such as speech recognition, information retrieval and computational lexicography.

The widespread use of taggers is due to their ability to work on large amounts of quite variable data (given an appropriate training phase). They also represent the first step in solving at least one aspect of the ambiguity problem, one of the major problems of natural language analysis.

5.2 Grammar Development

In recognition of the meager results that traditional grammar development has generally produced, efforts have turned to incorporating data-oriented methods to improve performance and coverage along two different lines. One approach is concerned with augmenting existing grammars and traditional methods with probabilities, the other with inducing new grammars from large corpora. What both have in common is the need for annotated material in the training process.

The inclusion of probabilities in a parser, by ranking the rules according to their frequency of use for a given corpus, is reported on in Briscoe and Carroll (1993). They argue for the need to accommodate linguistically motivated constraints in contrast with some of the grammar learning programs that assign regular but arbitrary structures to the texts. Similarly, Black et al. (1993) discuss the development of history based grammars meant to accommodate a large variety of information, proposing a division of the parsing problem "into two sub-problems: one of grammar coverage for the grammarian to address and the other of statistical modeling to increase the probability of picking the correct parse of a sentence" (p. 36). The work by Hindle and Rooth (1993) to determine correct attachment of prepositional phrases by lexical probabilities is an example of this view.

[11]Cf. the *Proceedings of the Speech and Natural Language Workshop*, published by Morgan-Kaufmann. which provide a rich source of information about current work.

[12]The latter program is available via anonymous ftp from parcftp.xerox.com. See the bibliography in Liberman and Schabes (1993) for references to the numerous taggers.

The second direction in grammar development from corpora has concentrated on methods that are reliable and efficient to induce grammars automatically from texts. Pereira and Schabes (1992) demonstrate how the inside-outside algorithm can be successfully used to infer the parameters of a stochastic context-free grammar from a partially bracketed corpus. Bod (1993) derives a grammar from a corpus of labeled bracketings using statistical techniques. A less computationally intensive approach is presented in Brill (1993), who relies on only a very small training corpus to induce a grammar by simple transformations, i.e., by adding and deleting parentheses.

It is perhaps worth noting that all of the work reported on above was only possible due to the availability of annotated corpora. In fact, most of the current projects used material prepared by the Penn Treebank (Marcus et al. 1993). There is also work underway on training grammars from unlabeled texts (see Kupiec and Maxwell 1992). The underlying idea is to probabilistically identify word equivalence classes for subsequent use in part-of-speech tagging and parsing programs.

5.3 Lexical Acquisition

One of the major bottlenecks in NLP development has been the human labor-intensive task of acquiring the necessary lexical resources. The efforts to re-use existing machine-readable dictionaries have only partially alleviated this problem, and for languages other than English, there are no dictionaries that contain the explicit and detailed subcategorization information as found in the popular learners' dictionaries. Methods are being explored to automatically derive subcategorization frames, identify syntactic and semantic classes, discover phrasal expressions and build bilingual dictionaries (cf. papers in the *Proceedings of the SIGLEX Workshop*, Boguraev and Pustejovsky, 1993).

In partial answer to the need for detailed syntactic information, Brent (1993) developed a program to identify subcategorization frames of verbs based on the occurrence of pronouns. Manning (1993) and Ushioda et al. (1993) also report on work to acquire subcategorization frames. In the program developed by Manning (1993), the tagged data are first run through a finite state parser to identify potential complements and then filtered on the basis of statistical regularities over the candidate words. Methods for proper name identification and classification, an important phenomena in texts of all kinds, have been developed by McDonald (1993) among others. Weischedel et al. (1993) discuss a range of probabilistic methods for identifying unknown words and for dealing with ambiguity in more robust NLP applications. Work on the identification of noun phrases and collocations, another major problem for current NLP applications, is reported on in Smadja (1993) and Kupiec (1993).

Access to textual data provides the resource for learning about the different uses of a word, in particular uses not previously attested to in dictionaries or simply overlooked by human introspection (Church and Hanks 1990). Class-based approaches to lexical discovery have been investigated by Futrelle and Gauch (1993) who automatically identify classes on the basis of mutual information and position, and Resnik (1992), whose work is also based on mutual information measures, the initial word classes being constrained by a thesaurus (WordNet) (Miller et al. 1990). Word associations as they occur in text are compared to psycholinguistic studies in Wettler and Rapp (1993). Pustejovsky et al. (1993) and Waterman (1993) demonstrate how lexical semantic information can be

identified in texts.

5.4 Work with Multilingual Corpora

With the growing availability of large amounts of parallel texts,[13] corpus-based studies on translation have begun to emerge. Reliable alignment techniques for different text types have been developed (Brown et al. 1991; Kay and Röscheisen 1993; Church 1993) with a high level of accuracy. These alignment methods work with very simple notions of similarity of patterns of sentence lengths and regularity of (approximate) word pairs across the texts. Extensions to refine problematic alignment cases have been proposed using cognates (Simard et al. 1992) and predefined word lists such as bilingual dictionaries and terminology banks (Catizone et al. 1989). Two more recent studies by Church (1993) and Chen (1993) allow for more robust alignment in case of corrupted data (e.g., misplaced footnotes or missing segments of texts).

Partially annotated multilingual data is being used in studies to automatically identify word pair correspondences Dagan et al. (1993), in word-sense disambiguation Church (1991), in example-based machine translation Sato and Nagao 1990; Matsumoto et al. 1993; and Sumita and Iida 1992) and even in fully automatic MT (Brown et al. 1988)[14]. Example-based machine translation, first advocated by Nagao (1984), relies on a database of structured bilingual texts which are automatically matched according to lexical and structural regularities and various distance measures based on, e.g. thesauri.

Intelligent access to multilingual texts also provides the basis for a new generation of tools for translators (des Tombe and Armstrong 1993; Shemtov 1993; Simard et al. 1992). These new systems provide access to previously translated texts as a resource for identifying possible translations by searching on aligned text segments. The tools can also provide facilities for checking for potential translation errors such as missing segments and inconsistent use of terminology. Lexicography is another application domain where useful tools are being developed (Church and Hanks 1990) both for monolingual and bilingual work. In Smadja (1992), his initial work on extracting collocations is extended to include phrasal expressions and their translations. These multilingual investigations will certainly become more widespread as more parallel data becomes available.

5.5 Evaluation of Methods

An important issue which has been systematically addressed in US government funded NLP projects under the ARPA programs is the evaluation of methods and measurement of overall progress in the field. In Europe work is under way to elaborate policies and programs to better evaluate current work.[15] The issue of comparing results has been hampered by the limited textual resources available. The lack of public corpora for languages other than English has meant that much of the current work carried out

[13] Though currently only the Hansard corpus is publicly available a number of new corpora are currently in preparation by the LDC and the ECI.

[14] Statistical machine translation is an important focus of the ARPA program.

[15] E.g. under the recently created evaluation sub-group within the Expert Advisory Group for Linguistic Engineering Standards.

in different countries, working with different languages, has remained a local matter. And lack of comparable corpora in different languages has meant that no comparison is possible on how successful the current methods might be for languages other than English.

The issue on the adequacy of methods in use is yet another topic that deserves more attention in light of the potential misuse of statistical data. Church and Mercer (1993) address this issue in general and Dunning (1993) provides a case study of the potential weakness of using the wrong measures for a given problem. A comparison of different methods in view of a more systematic elaboration of evaluation techniques is presented in Grefenstette (1993) - a topic that will certainly become more important as methods proliferate and the 'claimed' results are brought under more rigorous scrutiny.

6 Conclusion

In this paper we have summarized a new and exciting direction in work in NLP. The growing availability of on-line corpora provides the basis for development of new methods to account for natural language phenomena, to further our insights in language use and to develop practical NLP programs. The necessary textual resources are still lacking, but some progress has been made to overcome this problem and current programs promise to deliver even more in the future. A representative sample of the wide range of new studies currently underway have been presented as a demonstration of the potential of the new data-oriented approaches to language study.

Acknowledgments

The data collection work undertaken at ISSCO and reported on here has been supported by SWISSTRA and in part by a grant from the Linguistic Data Consortium. Thanks are also due to my colleagues Afzal Ballim, Graham Russell and Louis des Tombe for comments on previous drafts of this paper.

References

[1] Bar-Hillel, Y., "The state of machine translation in 1951", *American Documentation,* 2:229-237, 1951.

[2] Biber, D., "Using register-diversified corpora for general language studies", *Computational Linguistics,* 19(2):219-242,1993.

[3] Black, E., F. Jelinek, J. Lafferty, M. Magerman, R. Mercer, and S. Roukos, "Towards history-based grammars: Using richer models for probabilistic parsing", In *Proceedings of the ACL,* pages 31-37, Columbus, Ohio, 1993.

[4] Bod, R., "Using an annotated corpus as a stochastic grammar", In *Proceedings of the Conference of the European Chapter of ACL,* pages 37--44, Utrecht, Holland, 1993.

[5] Boguraev, B. and J. Pustejovsky, (eds.) *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text.* Association for Computational Linguistics, Columbus, Ohio, 1993.

[6] Brent, M., "From grammar to lexicon: Unsupervised learning of lexical syntax", *Computational Linguistics,* 19(2):243-262,1993.

[7] Brill, E., "Automatic grammar induction and parsing free text: A transformation-based approach", In *Proceedings of the ACL,* pages 259-265, Columbus, Ohio, 1993.

[8] Briscoe, T. and J. Carroll, "Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars", *Computational Linguistics,* 19(1): 25-60,1993.

[9] Brown, P., J. Cocke, S. Delia Pietra, V. Delia Pietra, F. Jelinek, R. Mercer, and P. Roossin, "A statistical approach to language translation", In *Proceedings COLING-88,* pages 71-76, Budapest, 1988.

[10] Brown, P., J. Lai, and R. Mercer. "Aligning sentences in parallel corpora", In *Proceedings of the ACL,* pages 169-176, Berkeley, California, 1991.

[11] Catizone, R., G. Russell, and S. Warwick-Armstrong, "Deriving translation data from bilingual texts", in Zernik, (ed.), *Proceedings of the Lexical Acquisition Workshop,* Detroit, Michigan, 1989.

[12] Chen, S., "Aligning sentences in bilingual corpora using lexical information", in *Proceedings of the ACL,* pages 9-16, Columbus, Ohio, 1993.

[13] Church, K. and P. Hanks. "Word association norms, mutual information, and lexicography", *Computational Linguistics,* 16(l):22-29, 1990.

[14] Church, K. and R. Mercer. "Introduction to the special issue on computational linguistics using large corpora", *Computational Linguistics,* 19(1): 1-24,1993.

[15] Church, K., "A stochastic parts program and noun phrase parser for unrestricted text2, in *Proceedings of the Second Conference on Applied Natural Language Processing,* pages 136-143, Austin, Texas, 1988.

[16] Church, K., "Concordances for parallel text", *in Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research,* pages 40-62, Oxford, England, 1991.

[17] Church, K., "Char_align: A program for aligning parallel texts at the character level", in *Proceedings of the ACL,* pages 1-8, Columbus, Ohio, 1993.

[18] Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger", in *Proceedings of the Conference on Applied Natural Language Processing Processing,* Trento, Italy, 1992.

[19] Dagan, I., W. Gale, and K. Church. "Robust bilingual word alignment for machine aided translation", in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives,* pages 1-8, Columbus Ohio, 1993.

[20] des Tombe, L. and S. Armstrong, "Using function words to measure translation quality", In *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research,* pages 1-18, Oxford, England, 1993.

[21] Dunning, T., "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics,* (19)1:61-74,1993.

[22] Francis, W. and H. Kučera, *Frequency Analysis of English Usage.* Houghton Mifflin, Boston, Massachusetts, 1982.

[23] Futrelle, R. and S. Gauch, "Experiments in syntactic and semantic classification and disambiguation using bootstrapping", In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text,* pages 117-127, Association for Computational Linguistics, Columbus, Ohio, 1993.

[24] Graff, D., "The UN multilingual text corpus", in *LDC Newsletter,* Vol. 1, No. 3. Linguistic Data Consortium, 1993.

[25] Grefenstette, G., "Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches", In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text,* pages 128-142, Association for Computational Linguistics, Columbus, Ohio, 1993.

[26] Hindle, D. and M. Rooth. "Structural ambiguity and lexical relations", *Computational Linguistics,* 19(1): 103-120,1993.

[27] Kay, M. and M. Röscheisen, "Text-translation alignment", *Computational Linguistics,* 19(1): 121-142,1993.

[28] Kupiec, J., "An algorithm for finding noun phrase correspondences in bilingual corpora", in *Proceedings of ACL,* pages 17-22, Columbus, Ohio, 1993.

[29] Kupiec, J. and J. Maxwell, "Training stochastic grammars from unlabelled text corpora", in *Workshop Notes from the AAAI Workshop on Statistically-Based Natural Language Processing Techniques,* pages 14-19, San Jose, California, 1992.

[30] Liberman, M. and Y. Schabes, "Tutorial on statistical methods in natural language processing", held in conjunction with the Conference of the European Chapter of ACL, 1993.

[31] Liberman, M., "Text on tap: The ACL/DCI", in *Proceedings of the 1989 DARPA Speech and Natural Language Workshop,* Cape Cod, Massachussetts, 1989.

[32] Liberman, M., "Introduction to the Linguistic Data Consortium", distributed at COLING-92, Nantes, 1992.

[33] Manning, C., "Automatic acquisition of a large subcategorization dictionary from corpora", in *Proceedings of the ACL,* pages 235-242, Columbus, Ohio, 1993

[34] Marcus, M., B. Santorini, and M. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics,* 19(2):5 -331,1993.

[35] Matsumoto, Y, H. Ishimoto, and T. Utsuro, "Structural matching of parallel texts", in *Proceedings of ACL,* pages 23-30, Columbus, Ohio, 1993.

[36] McDonald, D., "Internal and external evidence in the identification and semantic categorization of proper names", In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text,* pages 32-43, Association for Computational Linguistics, Columbus, Ohio, 1993.

[37] Miller, G. et al., Five papers on WordNet, Technical report, Cognitive Science Laboratory, Princeton University, 1990.

[38] Monachini, M. and A. Östling, "Morphosyntactic corpus annotation - a comparison of different schemes", Technical report, Istituto di Linguistica Computazionale CNR, Pisa, 1992. Report for NERC project.

[39] Nagao, M., "A framework of a mechanical translation between Japanese and English by analogy principle", in A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence,* pages 173-180. North-Holland, 1984.

[40] Pereira, F. and Y. Schabes, "Inside-outside reestimation from partially bracketed corpora", in *Proceedings of ACL,* pages 128-135, Newark, Delaware, 1992.

[41] Pustejovsky, J., S. Bergler, and P. Anick, "Lexical semantic techniques for corpus analysis", *Computational Linguistics,* 19(2):331-358,1993.

[42] Resnik, P., "WordNet and distributional analysis: a class-based approach to lexical discovery", in *Workshop Notes from the AAAI Workshop on Statistically-Based Natural Language Processing Techniques,* pages 54-64, San Jose, California, July, 1992.

[43] Sato, S. and M. Nagao, "Towards memory-based machine translation", in *Proceedings of COLING-90,* pages 247-252, Helsinki, 1990.

[44] Shemtov, H., "Text alignment in a tool for translating revised documents", in *Proceedings of the European Chapter of the ACL,* pages 449-453, Utrecht, Holland. 1993.

[45] Simard, M., G. Foster, and P. Isabelle, "Using cognates to align sentences in bilingual corpora", in *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation,* pages 67-82, Montreal, 1992.

[46] Sinclair, J. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing,* Collins, London, 1987.

[47] Smadja, F., "How to compile a bilingual collocational lexicon automatically", in *Workshop notes from the AAAI Statistically-Based NLP Techniques Workshop,* pages 65-71, San Jose, California, July, 1992.

[48] Smadja, E, "Retrieving collocations from text: Xtract", *Computational Linguistics,* 19(1): 143-178, 1993.

[49] Sumita, E. and H. Iida, "Example-based natural language processing techniques - a case study of machine translation", in *Workshop notes from the AAAI Statistically-Based NLP Techniques Workshop,* pages 90-97, San Jose, California, July, 1992.

[50] Thompson, H., "European Corpus Initiative", *ELSNEWS,* 1(1), 1992.

[51] Thompson, H., "Multilingual corpora for cooperation (MLCC)", Proposal submitted under the LRE program for International Scientific Cognitive science-operation, 1993.

[52] Ushioda, A., D. Evans, T. Gibson, and A. Waibel, "The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora", in *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text,* pages 95-106, Association for Computational Linguistics, Columbus, Ohio, 1993.

[53] Walker, D., "The ecology of language", in *Proceedings of the International Workshop on Electronic Dictionaries,* pages 1-22, Tokyo, Japan, 1991.

[54] Waterman, S., "Structural methods for lexical/semantic patterns", *in Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text,* pages 128-142, Association for Computational Linguistics, Columbus, Ohio, 1993.

[55] Weaver, W., Translation, (memorandum), 1949.

[56] Weischedel, R., M. Meteer, R. Schwartz, L. Ramshaw, and J. Palmucci, "Coping with ambiguity and unknown words through probabilistic models", *Computational Linguistics,* 19(2):359-382,1993.

[57] Wettler, M. and R. Rapp, "Computation of word associations based on co-occurrences of words in large corpora", in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives,* pages 84-93, Columbus Ohio, 1993.