TWO PRINCIPLES AND SIX TECHNIQUES FOR RAPID MT DEVELOPMENT

Sergei Nirenburg, Stephen Beale, Stephen Helmreich,
Kavi Mahesh, Evelyne Viegas and Remi Zajac
Computing Research Laboratory
Box 3000l, Dept. 3CRL
New Mexico State University
Las Cruces, NM 88003-0001 USA
Phone: (505) 646-5466 Fax: (505) 646-6218
{sergei,sb,shelmrei,mahesh,viegas,rzajac}@crl.nmsu.edu

Abstract

In this paper we describe a range of techniques used at NMSU CRL for accelerating the development of MT systems. These techniques enable semi-automatic development of a number of components of a multilingual MT system, thereby enabling rapid deployment of MT capabilities in a new language. First, we describe the core multi-engine, multilingual architecture that enables the different techniques to be rapidly integrated to build an MT system. We show how off-the-shelf components were used in this architecture for fast development. Then we illustrate a set of techniques for semi-automatic acquisition of static resources: (a) automatic induction of grammars, (b) corpus-based acquisition of bilingual glossaries, and automatic acquisition of semantic lexicons through (c) lexical rules and (d) reversal of analysis lexicons to generation lexicons. Finally we describe an automatic testing environment that enables rapid validation of automatically acquired resources.

1 Rapid Development Techniques

Static knowledge sources — grammars, lexicons, world knowledge bases — are the most time-consuming concerns in any rule-based machine translation system. It is, therefore, imperative to find ways of speeding up the creation and updating of high-quality, useful static knowledge sources. It is equally imperative to rely on a robust and flexible core computational architecture that allows the concurrent manipulation of a large number of static and dynamic knowledge sources as well as documents and document collections. In this paper, we describe several techniques for facilitating rapid development of MT capabilities for a new language in the framework of an existing multilingual system.

Our approach is based on the following two principles:

- **Heterogeneous, Multi-Engine, Multilingual Architecture:** a multi-engine architecture where different subsets of MT techniques can be combined for different languages, accelerates development; it takes longer to perfect any one prespecified MT method for a new language to deliver comparable initial capabilities.

- **Manual Validation after Automatic Acquisition:** automatic acquisition of static resources, such as lexicons and glossaries, followed by automatic testing and machine-aided human validation is faster than manual acquisition; it is also faster than developing a fully automatic acquisition method that can deliver comparable quality.

We illustrate the above principles by describing our experiments with the following techniques:

1. Core multi-engine, multilingual architecture enables sharing of processing components, including off-the-shelf components.
2. Automatic induction of grammars accelerates the development of morphological and syntactic analyzers.
3. Corpus analysis enables automatic construction of bilingual glossaries.
4. Lexical rules (based on derivation morphology and ontological representations) enable rapid expansion of manually acquired lexicons.
5. Lexicon reversal produces generation lexicons from analysis lexicons.
6. Automatic testing methods enable rapid validation of automatically acquired resources.

2 The Core Environment

The Computing Research Laboratory (CRL) has developed a multi-engine architecture designed for rapid development of machine translation functionalities in a translator's workstation. This architecture is a descendant of the approach to multi-engine human-aided machine translation (HAMT) developed by Sergei Nirenburg and his associates at CMU CMT (Nirenburg et al., 1992; 1994; Frederking et al., 1993; Nirenburg and Frederking, 1994; Frederking and Nirenburg, 1994). At present, it serves as the environment for HAMT from Japanese, Arabic, Russian and Spanish to English. The system includes a general multilingual document handling environment supporting the translator's workstation and a basic multilingual machine translation architecture. Emphasis is made on reusing off-the-shelf linguistic components and on empirical, corpus-based acquisition of bilingual glossaries.

The translator's workstation is designed to be used concurrently with information extraction and retrieval tools. To this end, the Temple Translator's Workstation, which provides an integrated set of multilingual tools for the translator — including machine translation functions, is integrated with a Tipster Document Management system developed at CRL, which provides text management functionalities and allows a smooth integration of a translator's workstation with the rest of the user's working environment.

The Tipster Document Management system allows flexible integration of NLP tools to build an integrated application through a unifying mechanism: tools communicate through *annotations* on the document. Each annotation is associated with a span of text and contains information represented as attributes on objects. Thus, not only machine translation systems but also other NLP tools like taggers, spell-checkers, etc., can work concurrently on the same document. Alternative translations are also stored as annotations on the target text. The workstation allows the human translator to edit the target text to produce the final translation. To facilitate tool integration, the architecture defines a standard way of representing linguistic information used by all the engines for all the languages.

This architecture enables the overall system to have different sets of tools and engines for different languages. Such flexibility is essential for rapid development, since certain tools or engines may be too expensive to build for some languages but readily available for others. The basic architecture also includes Unicode-based support for concurrent display and manipulation of a very broad variety of writing systems.

Our work continues to advance the framework of the Pangloss MT project (e.g., Nirenburg (ed.), 1994), in which the word-for-word, glossary-based, transfer-based, knowledge-based and example-based translation engines have been used in a variety of configurations.

2.1 Reuse of Off-The-Shelf Linguistic Components

In order to develop working prototypes of lower-end modules for the MT engines in a short time, off-the-shelf linguistic components were used whenever possible, in the spirit of, for example, work at Cm in Montreal (e.g., Isabelle, 1992). For example, the SPOST Spanish part-of-speech tagger (Farwell et al., 1994), the Juman Japanese morphological analyzer (Matsumoto et al., 1993) and the Penman English morphological generator (Penman, 1988) were integrated. Often, off-the-shelf resources were used as input for a semi-automatic process of knowledge acquisition. Thus, in our experiments Spanish, Japanese and Arabic bilingual dictionaries were derived from corresponding machine-readable versions of paper dictionaries using the LexBase approach (Guthrie et al., 1993).

In reusing off-the-shelf components, a major problem has been the lack of compatibility between linguistic representations used by the various programs and databases. A typical example is the integration of Juman and the bilingual Japanese- English dictionary. The morphosyntactic categorization used by Juman is different from the one used by the dictionary, and perhaps more importantly, the segmentation of a word into its stem and morphs is also different. Thus, a converter had to be developed to interface the morphological analyzer and the bilingual dictionary. A converter was developed in the form of a Tcl script which takes the output of Juman and maps it to the Temple annotation structure. The converter also accesses the Japanese part of the bilingual dictionary to further ensure compatibility.

3 Rapid Development of Computational Grammars for languages with limited on-line resources

Hand-crafted computational grammars that support higher-level translation engines have largely failed to support robust parsing of unseen, realistic text. Dealing with this kind of text is a *sine qua non* for general-purpose MT. Incorporation of statistical techniques into the task of deriving large-scale grammars is widely expected to increase as well as provide tuning of grammar to a particular corpus or subject field. Statistical techniques have been developed for semi-automatic generation of taggers and parsers for any language (Brill, 1994; Brill and Marcus, 1994), though they do not yet provide a mature enough technology for general use.

CRL is working on semi-automatic development of robust grammars for languages with limited on-line resources. Limiting the resources is vital since the languages for which rapid MT development is most important are frequently those where electronic resources are scarce. Results reported to date refer to developing grammars for English using the extensive resources available for that language, in particular the hand-corrected parsed text of the Penn Tree Bank (Marcus et al., 1993). See Carroll and Charniak (1992) and Lari and Young (1990) for examples of this approach.

In our experiments, a small, core grammar for the languages under development (in our case, Russian and Serbo-Croatian) is used instead of the information provided by the parsed Penn Tree Bank corpus. This grammar will "seed" the development of a robust, large-scale grammar for these languages. This approach has been discussed, for example, in the work of Pereira and Schabes (1992) or Briscoe and Carroll (1993). We plan to combine this non-automatic top-down approach with any number of bottom-up organizing approaches, for instance, N-gram analyses of corpus which extend the methods applied by Finch and Chater (1991).

4 Rapid Lexical Acquisition

MT-related lexicons at CRL include both large bilingual terminological glossaries and deep-coverage computational-semantic lexicons to support various MT engines. For rapid development, acquisition of all these resources must be enhanced.

4.1 Corpus-Based Bilingual Glossary Acquisition

Bilingual glossaries support the glossary-based translation engine in the general multi-engine architecture. The glossaries are built using an N-gram generation program that produces a list of phrases recurring in a given corpus. These phrases are converted into the source-language parts of the glossary entries. The entries are completed either manually, using a multilingual glossary editor or, if a bilingual machine-readable dictionary (MRD) exists between the source language and English, the system suggests possible English translations by finding, in an English corpus from the same subject area, phrases that contain the most matches with the list of all possible translations (found in the MRD) of the terms in the source language phrase. Even manual acquisition of bilingual glossaries is rather fast. For example, it took about 6 person-months to acquire a bilingual Arabic-English glossary of over 10,000 entries.

4.2 The Content of the Lexicon

The CRL MT lexicon is composed of superentries, each containing entries corresponding to word senses, without reference to their part of speech (the senses of both the noun and the verb "walk," for example, will be listed inside the same superentry). Each word meaning is identified by a unique identifier, or lexeme (Mel'cuk et al., 1984; Onyshkevych and Nirenburg, 1995). No distinction is made between homonyms and polysemous words, and all homonyms and all meaning shifts of polysemous words are under one single superentry, adopting Fillmore's (1971) rather than Weinreich's approach (1964). Moreover, for "logically polysemous" words as defined by Pustejovsky (1991; 1995), we keep one entry (as opposed to two in most approaches). The meaning of a lexical entry is encoded in a (lexical) semantic representation language whose primitives are predominantly terms in an independently motivated world model, or ontology (see Carlson and Nirenburg, 1990; Mahesh and Nirenburg, 1995; Mahesh, 1996). The information contained inside a lexeme is divided into **zones** corresponding to various levels of lexical information. The different zones include, among others, *category, morphology, syntactic structure, semantic mapping, lexical relations and rules* and *stylistics.*

4.3 Lexical Rules for Accelerated Acquisition

The central idea of lexical rules (LRs)— that there are systematic paradigmatic meaning relations among lexical items, such that, given an entry for one such item, other entries can be derived automatically — is certainly not novel (Mel'cuk, 1979; Atkins, 1991; Ostler and Atkins, 1992; Briscoe and Copestake, 1991).

Most LRs in the generative lexicon approach (Pustejovsky, 1995) deal with small classes of words and explain such grammatical and semantic shifts as +*count* to - *count* or - *common* to +*common.* While shifts and modulations are important, we find that the main significance of LRs is in their promise to aid the task of *massive* lexical acquisition by allowing automatic generation of entries from a set of "core" entries acquired manually. The major problem in using LRs is overgeneration: LR generators suggest inappropriate forms which need to be weeded out by humans. This problem is usually overlooked if the goal of studying the LRs is theoretical (as is the case in many of the above-mentioned approaches). In practice, as we argue in (Viegas et al., 1996b), the costs of using LRs in knowledge acquisition must be carefully weighed against the benefits.

The lexicon for which our LRs are introduced is intended mainly to support the computational specification and use of text meaning representations used in the KBMT engine, though the syntactic, morphological, pragmatic, stylistic and other information stored there will also aid other MT engines (see, e.g., Onyshkevych and Nirenburg, 1995). The acquisition of such a lexicon, with or without the LR mechanism, involves a substantial investment of resources.

The LR processor applies to each word sense in a superentry. For example, *pronunciar* has (at least) two entries (one could be translated as "articulate" and one as "declare"); the LR generator, when applied to the superentry, would produce (among others) two entries for *pronunciacion,* derived from each of the two verbal senses/entries.

The nature of the links between the lexicon and the ontology is critical to the entire issue of LRs. Representations of lexical meaning may be defined in terms of any number of ontological primitives, called *concepts.* Any of the concepts in the ontology may be used (singly or in combination) in a lexical meaning representation.

Correlations between syntactic category and semantic or ontological class of a lexical unit are at best underspecified and, in the general case, must be considered largely arbitrary. For example, although meanings of many verbs are represented through reference to ontological EVENTS and a number of nouns are represented by concepts from the OBJECT sublattice, one cannot rely on this being the norm. Many LRs change the syntactic category of the input form; in our model the basic semantic category is usually preserved. For example, the verb *destroy* will have the semantics of an EVENT, as will the noun *destruction* (naturally, with a different linking in the syntax-semantics interface). Similarly, *destroyer* (as a person) would be represented using the same event with the addition of a HUMAN as a filler of the agent case-role. This built-in transcategoriality strongly facilitates applications such as interlingual MT, as it renders vacuous many problems connected with category mismatches and misalignments that plague those paradigms in MT that do not rely on extracting language-neutral text meaning representations (e.g., Dorr 1995; Kameyama, Ochitani, and Peters, 1991).

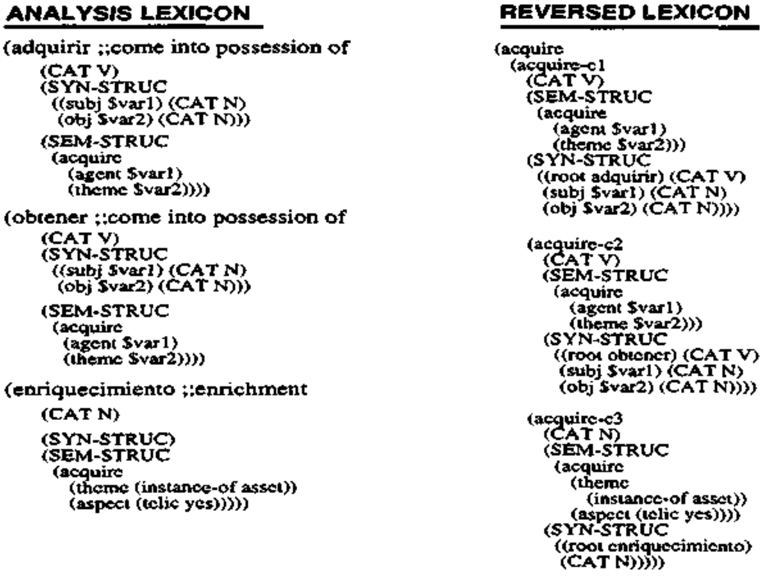
The following phenomena seem to be appropriate for treatment with LRs:

- Inflected Forms - Specifically, those inflectional phenomena which accompany changes in subcategorization frame (passivization, dative alternation, etc.).
- Word Formation - The production of derived forms by LR is illustrated in a case study below, and includes formation of deverbal nominals *(destruction, running),* agentive nouns *(catcher).* Typically involving a shift in syntactic category, these LRs are often less productive than inflection-oriented ones. Consequently, derivational LRs are even more prone to overgeneration than inflectional LRs.
- Regular Polysemy - This set of phenomena includes regular polysemies or regular non-metaphoric and non-metonymic alternations such as those described in Apresjan (1974), Pustejovsky (1991; 1995), and other literature on LRs.

In our initial experiment, a set of LRs was used for the acquisition of the Spanish computational lexicon in the Mikrokosmos project. In this experiment, about 7,000 manually acquired entries resulted, after the application of a battery of LRs, in a lexicon of over 35,000 entries, as described in (Viegas et al., 1996a). LRs were also used at run time as a failure recovery procedure when a lexical item present in a text had not yet been encoded in the lexicon.

4.4 Reversing Analysis Lexicons for Use in Multilingual Generation

Another way of speeding up lexicon compilation is converting available computational lexicons to new languages and uses. In this section, we briefly describe the issue of semi-automatic generation of multilingual computational semantic lexicons. In our experiment, we used the existing large-scale analysis lexicon as the starting point for building a generation lexicon for the same language. This process produces a conceptual lexicon which must be enhanced to produce a lexicon suitable for generation. The analysis lexicon is indexed on source language words. Each entry includes information on syntactic dependency (SYN-STRUC), along with a mapping to its meaning representation (SEM-STRUC). For example, a number of Spanish words map into an **acquire** concept:



```
ANALYSIS LEXICON
(adquirir ;;come into possession of
    (CAT V)
    (SYN-STRUC
     ((subj $var1) (CAT N)
      (obj $var2) (CAT N)))
    (SEM-STRUC
     (acquire
       (agent $var1)
       (theme $var2))))
(obtener ;;come into possession of
    (CAT V)
    (SYN-STRUC
     ((subj $var1) (CAT N)
      (obj $var2) (CAT N)))
    (SEM-STRUC
     (acquire
      (agent $var1)
      (theme $var2))))
(enriquecimiento ;;enrichment
    (CAT N)
    (SYN-STRUC)
    (SEM-STRUC
     (acquire
       (theme (instance-of asset))
       (aspect (telic yes)))))
```

```
REVERSED LEXICON
(acquire
  (acquire-c1
    (CAT V)
    (SEM-STRUC
     (acquire
       (agent $var1)
       (theme $var2)))
    (SYN-STRUC
     ((root adquirir) (CAT V)
      (subj $var1) (CAT N)
      (obj $var2) (CAT N))))

    (acquire-c2
      (CAT V)
      (SEM-STRUC
       (acquire
         (agent $var1)
         (theme $var2)))
      (SYN-STRUC
       ((root obtener) (CAT V)
        (subj $var1) (CAT N)
        (obj $var2) (CAT N))))

    (acquire-c3
      (CAT N)
      (SEM-STRUC
       (acquire
         (theme
           (instance-of asset))
         (aspect (telic yes))))
      (SYN-STRUC
       ((root enriquecimiento)
        (CAT N)))))
```

Figure 1: Sample Entries from Analysis and Reversed Lexicons.

The algorithm to "reverse" the analysis lexicon to produce the generation lexicon involves rearranging, modifying, deleting, and adding information. It centers on reindexing the material based on the meaning and not on the lexical units. The generation lexicon must encode information which is specific to the process of generation, such as word order and collocations. This information must be added by some means to the reversed lexicon. However, a reversed lexicon has advantages beyond its practical use in generation. We have begun using it in the following areas:

- Evaluation of semantic analysis by reverse generation of input text in the same language. With a conceptual lexicon that is based on the original analysis lexicon, it is possible to take the output semantic representations from the analyzer and submit them to a text generator. The output surface structures can then be compared to the input text.

- Evaluating Text Meaning Representations (TMRs) with respect to the granularity of semantic representation. Is the representation precise enough to correctly translate all meaning components, or is a specific source term mapped into a generalized one from which the original meaning cannot be recovered?

- Testing lexicon entries (see below).

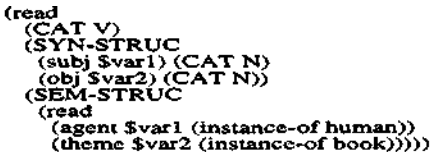
5 Automatic Testing and Validation

As the size of computational lexicons begins to increase, and as more and more automated methods are used to populate them, the need for new testing methodologies grows. Testing lexicons can be divided into the following three areas:

Format (syntax): Tools to detect improperly formed lexicon entries. For example, misplaced parentheses and quotes, although uninteresting theoretically, can cause many practical problems.

Meaning (semantics): Quality and consistency of information across the lexicon and the ontology needs to be checked for each zone of the entry. For instance, inside the syntax zone, the syntactic structures indicated should be consistent with the grammar of the language. In the semantics zone, meaning assignment and constraint checking should be consistent with meaning definitions in the ontology. Assigning an AGENT to a PLACE in some lexicon entry, for example, is suspect, as is constraining the AGENT of a KILL event to be a MONKEY.

Correctness (pragmatics): We have created a set of tools that automatically create input test sentences (or test semantic structure inputs) that can be used to evaluate specific lexicon entries. This helps in two ways. First, simply seeing a sentence that fits the syntax and semantics present in the lexicon entry can often highlight errors. This type of testing ensures that the lexicon items will be useful and will be applied in the correct situations. Second, after generating the test items, the lexicons can be used to process them, with the results also subject to review. This ensures that the lexicon item produces the intended results. For example, a simple entry for the English word *read* might look like:



```
(read
  (CAT V)
  (SYN-STRUC
    (subj $var1) (CAT N)
    (obj $var2) (CAT N))
  (SEM-STRUC
    (read
      (agent $var1 (instance-of human))
      (theme $var2 (instance-of book)))))
```

Figure 2: Lexical Entry for "read".

We can then use the conceptual lexicon to generate sentences 1) from the given SEM-STRUC and 2) by substituting into the SYN-STRUC appropriate text for any VARs present. If the lexicon entries are correct, the two sentences should be similar **and** should make sense to the tester. Sentences such as the following would signal problems:

The book read John.
John read into the book.
John read the cheese.

This is especially helpful for automatically generated lexicon entries such as nominal entries created from verbal entries using lexical rules.

6 Current and Future Work

The environment for rapid prototyping MT applications briefly described above has been initially created, under DoD sponsorship, in the framework of the Pangloss and Mikrokosmos projects and extended in the Temple project (Zajac and Vanni, submitted). The environment also incorporates results from the Tipster (DARPA, 1993) and Norm (Ogden, 1996) efforts at CRL; all of the extant NLP-related projects at CRL use it or some of its parts.

Development continues of both the core architecture and the concrete tools. Applications typically involve a subset of the tools within the basic architecture. Thus, for instance, the first system in the Corelli project (Zajac and Vanni, submitted) will demonstrate a relatively low-level capability for Serbo-Croatian translation support.

We plan to demonstrate our environment and acquisition tools at the conference.

References

Apresjan, Yu. (1974). Regular Polysemy, *Linguistics* vol. 142, pp. 5-32.

Atkins, B.T.S. (1991). Building a lexicon: The contribution of lexicography. In B. Boguraev (ed.), "Building a Lexicon", Special Issue, *International Journal of Lexicography* 4:3, pp. 167-204.

Brill, E. (1994) Advances in rule-based part of speech tagging, In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, WA.

Brill, E. and Marcus, M. (1992) Automatically acquiring phrase structure using distributional analysis. Proc. DARPA Speech and Language Workshop.

Briscoe, T. and Carroll, J. (1993) Generalized probabilistic LR parsing of natural language corpora with unification-based grammars. Computational Linguistics.

Briscoe, E.J. and A. Copestake. (1991). Sense extensions as lexical rules. In *Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language.* Sydney, Australia, pp. 12-20.

Carlson, L. and Nirenburg, S. (1990). World Modeling for NLP. Technical Report CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Carroll, G. and Charniak, E. (1992) Learning Probabilistic Dependency Grammars from Labeled Text. AAAI Fall Symposium on Probabilistic Approaches to Natural Language.

DARPA "Proceedings of the Tipster Text Program (Phase I)" Fredericksburg, Va., Morgan Kaufmann, 1993.

Dorr, B.J. (1995) A lexical-semantic solution to the divergence problem in machine translation. In St-Dizier P. and Viegas E. (eds), Computational Lexical Semantics: CUP.

Farwell, D., Helmreich, S., Jin, W., Casper, M., Hargrave, J., Molina-Salgado, H. Weng, F. 1994. Panglyzer: Spanish Language Analysis System. In Proceedings of the Conference of the Association of Machine Translation in the Americas (AMTA). Columbia, MD.

Fillmore, C. (1971). Types of lexical information. In D.D. Steinberg and L.A. Jakobovits (eds.) *Semantics,* pp. 370-392, Cambridge University Press.

Finch, S. and Chater, N. (1991) A hybrid approach to the automatic learning of linguistics categories. Artificial Intelligence and Simulated Behaviour Quarterly. 78: 16-24.

Frederking, R. and S. Nirenburg. 1994. Three Heads Are Better than One. Proceedings of ANLP-94. Stuttgart, October.

Frederking, R., D. Grannes, P. Cousseau, and S. Nirenburg. 1993. "An MAT Tool and Its Effectiveness". Proceedings of the DARPA Human Language Technology Workshop, Princeton, NJ.

Guthrie, Louise, Rauls, Venus, Luo, Tao, Bruce, Rebecca. 1993. "LEXI-CAD/CAM, A Tool for Lexicon Builders". CRL Technical Report MCCS-93-259.

Isabelle, P. (1992). Bi-textual aids for translator. *Proc. Eighth Ann. Conf. UW Centre for the New OED and Text Research,* UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Ontario, Canada, 1992.

Kameyama M., Ochitani, R. and Peters, S. (1991) Resolving Translation Mismatches With Information Flow. In Proceedings of ACL'91.

Lari, K. and Young, S. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language.

Mahesh, Kavi, and Sergei Nirenburg (1995). A situated ontology for practical NLP. In the Proceedings of *IJCA1'95* Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, August 19-21.

Mahesh, K. (1996). *Ontology Development: Ideology and Methodology.* Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University.

Marcus, M., Santorini, B. and Marcinkiewicz, M. (1993) Building a large annotated corpus of English: The Penn Treebank, Computational Linguistics, Vol. 19, No. 2.

Matsumoto, Yuji, Sadao Kurohashi, Takeji Utsuro, Yutaka Myougi, Makoto Nagao. 1993. "Japanese Morphological Analysis System, JUMAN". Kyoto University, Nara Science and Technology Graduate School University (in Japanese).

Mel'cuk, Igor. (1979). *Studies in Dependency Syntax.* Ann Arbor, MI: Karoma.

Mel'cuk, I., N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja and A. Lessard (1984) : *Dictionnaire explicatif et combinatoire du françcais contemporain: recherches lexico-semantiques I.* Montréal: Presses de l'Université de Montréal.

Nagao, M. and Mori, S. (1994) A new method of N-gram statistics for large number of n, and automatic extraction of words and phrases from large text data of Japanese (1994). in Proc. COLING 1994, Kyoto.

Nirenburg, S., P. Shell, A. Cohen, P. Cousseau, D. Grannes and C. McNeilly. 1992. The Translator's Workstation. Proceedings of the 3rd Conference on Applied Natural Language Processing. Trento, Italy, April.

Nirenburg, S. And R. Frederking. 1994. Toward Multi-Engine Machine Translation. Proceedings of the Human Language Technology Conference, Princeton,

Nirenburg, S., R. Frederking, D. Farwell and Y. Wilks. 1994. Two Types of Adaptive MT Environments. Proceedings of COLING-94. Kyoto, August.

Nirenburg, S. (editor) 1994. The Pangloss Mark III Machine Translation System. NMSU CRL, USC ISI and CMU CMT Technical Report.

Ogden, W.C. (1996). Oleada: An Intergrated Multilingual Software System for Language Analysts, Instructors and Learners. Symposium on Advanced Information Processing and Analysis, AIPA96, Tysons Corner, Virginia, March 26-28,1996.

Onyshkevych, B. and S. Nirenburg (1994) *The Lexicon in the Scheme of KBMT Things.* Technical Report MCCS-94-277, Computing Research Laboratory, New Mexico State University.

Ostler, N. and B. T. S. Atkins. (1992). Predictable meaning shift: Some linguistic properties of lexical implication rules. In J. Pustejovsky and S. Bergler (eds), *Lexical Semantics and Knowledge Representation.* Berlin: Springer, pp. 87-100.

Onyshkevych, B.A. and Nirenburg, S. (1995). A lexicon for knowledge-based MT. *Machine Translation,* 10:1-2, pp. 5-57, Special issue on Building Lexicons for Machine Translation II.

The Penman Primer, User Guide, and Reference Manual. 1988. Unpublished USC/ISI documentation.

Pereira, F. and Schabes, Y. (1992) Inside-outside re-estimation from partially bracketed corpora. Proc. 20th Meeting of the ACL, Newark, DE.

Pustejovsky, J. (1991) The Generative Lexicon. In *Computational Linguistics,* 17(4).

Pustejovsky, J. (1995) The Generative Lexicon. Cambridge, MA: MIT Press.

Viegas, E., Gonzalez, M., and Longwell, J. (1996a). Morpho-semantics and Constructive Derivational Morphology: a Transcategorial Approach to Lexical Rules. Technical Report MCCS-96-295, CRL, NMSU.

Viegas, E., Onyshkevych, B.A., Raskin, V., and Nirenburg, S. (1996b). From *Submit* to *Submitted* via *Submission:* On Lexical Rules in Large-Scale Lexicon Acquisition. In *Proc. ACL-96.*

Weinreich, U. (1964). Webster's Third: A Critique of its Semantics. In *International Journal of American Linguistics* 30:405-409.

Zajac, R. and Vanni, M. "Glossary-Based MT Engines in a Multilingual Analyst's Workstation Architecture". To appear in *Machine Translation.* Special Issue On New Tools For Human Translators.