MCDONNELL DOUGLAS ELECTRONIC SYSTEMS COMPANY: DESCRIPTION OF THE TEXUS SYSTEM AS USED FOR MUC-4

Amnon Meyers and David de Hilster

McDonnell Douglas Electronic Systems Company
Advanced Computing Technologies Lab
1801 E. St. Andrew Place
Santa Ana, CA 92705-6520
{vox, hilster}@young.mdc.com
(714) 566-5956

INTRODUCTION AND APPROACH

Unlike most natural language processing (NLP) systems, TexUS (**Text Understanding System**) is being developed as a domain-independent *shell*, to facilitate the application of language analysis to a variety of tasks and domains of discourse. Our work not only develops robust and generic language analysis capabilities, but also elaborates knowledge representations, knowledge engineering methods, and convenient interfaces for knowledge acquisition.

TexUS builds on INLET (Interactive Natural Language Engineering Tool) [1][2], which was used for MUC3. Both descend from VOX (Vocabulary Extension System) [3][4], which was developed from 1983-87 at UCI and 1988-90 at MDESC. Many analysis, knowledge representation, and knowledge acquisition ideas from VOX have evolved in constructing TexUS. In particular, TexUS (1) implements completely new, robust, and tailorable analysis algorithms; (2) embeds analyzer data structures within the knowledge representation; (3) supports interactive graphical knowledge engineering; (4) runs in C on Sun workstations, to improve portability and speed; and (5) employs a pragmatic and domain-independent knowledge representation framework. TexUS and INLET differ primarily in the strength of the analysis capability. Figure 1 exemplifies the graphical utilities available in TexUS.

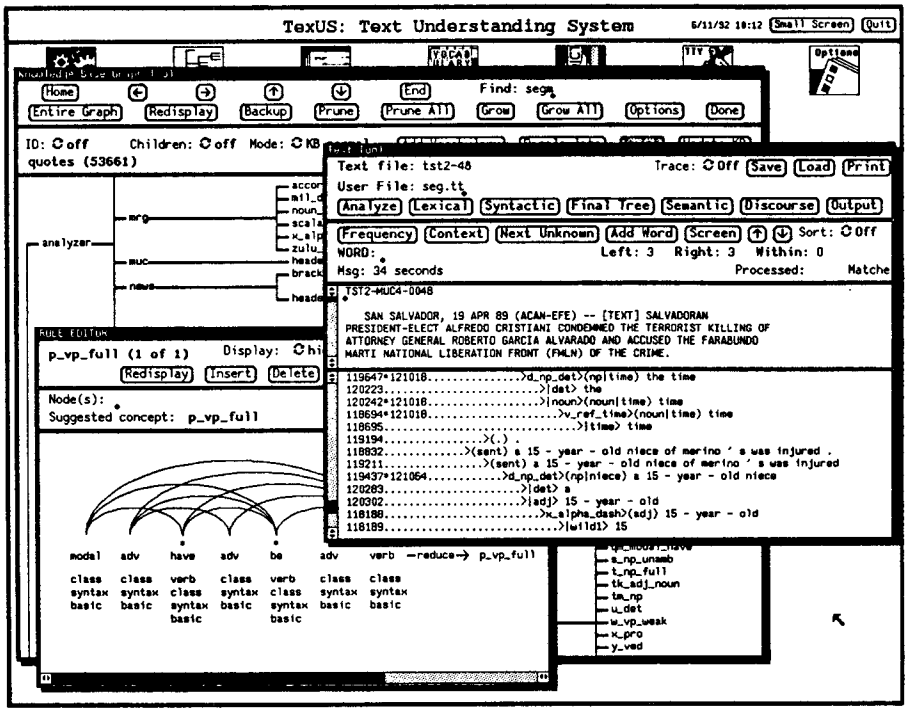

Figure 1. TexUS provides a host of interactive graphic tools for NLP development

The system we used at MUC3 was new (under development for 9 months) and incomplete. In essence, a sophisticated skimming system carried the analysis burden. Since that time, we have implemented a comprehensive analyzer that includes the following phases: (1) key pattern search, which locates potentially interesting texts and text fragments; (2) pre-processing, where idiomatic constructs are collapsed in bottom-up fashion; (3) syntactic analysis, which uses primarily top-down mechanisms to successively segment text into finer-grained units, i.e., paragraphs, sentences, clauses, and components; (4) semantic analysis, which extracts and standardizes information found in the parse tree; (5) discourse analysis, which traverses the semantic representation to establish relationships between events, actors, objects, and other cases; and (6) back-end processing, which converts the internal representation produced by the analyzer to task-specific output. Table 1 highlights differences between the TexUS and INLET systems.

	MUC3(INLET)	MUC4(TexUS)
DICTIONARY	~5k (including inflected forms)	~90k (with p.o.s., root forms only)
LEXICAL	• minimal	• 2 spelling correctors • n-grams for Spanish & English • morphological analysis • unknown word processor
SYNTAX	• embedded in semantic rules • no parse tree	•separate syntactic segmentation •syntactic parse tree
SEMANTICS	• actor/object extraction	• parse tree traversal • extraction of semantic cases
DISCOURSE	• minimal	• anaphora • composite events, actors, & objects

Table 1. TexUS builds on INLET analysis capabilities

The algorithms provided with TexUS support truly robust analysis. Construction of useful parse trees does not depend on complete syntactic characterization of the text, and succeeds even in the presence of ungrammatical, terse, and garbled text abounding in unknown words and phrases. This is accomplished by using grammar rules that successively relax constraints. For example, one hierarchy of grammar rules is dedicated to segmenting clauses into components (e.g., noun and verb phrases). Rules that represent the strongest confidence (such as "det quan adj noun" with no optional parts) are applied first. If these fail, rules with optionality are applied next (e.g., with "det" missing). Next, rules containing wildcard elements (e.g., a wildcard "noun" element) but allowing no optionality are applied, followed by rules with wildcards and optionality. In this way, we attempt to match the text to rules with highest confidence first (the ordering interleaves, e.g., rules for noun phrases and verb phrases, to maximize confidence). Using rules with wildcards and little optionality allows the analyzer to characterize fragments of text containing unknown words with reasonable confidence in many cases. Such rules also support automated knowledge acquisition (See section 2.2).

The parsing mechanisms are driven by hierarchies of grammar rules. Much of our analyzer development thus consists of refining these rule sets, rather than building code. Further, we have developed generic analysis mechanisms that serve multiple tasks, depending on the rule sets they are given as parameters. Building analyzers for new tasks is substantially reduced to selecting and using existing analysis mechanisms and creating rule sets when needed for new domains. We are currently developing an interactive graphical analyzer tool to simplify enhancement of the analyzer algorithms.

In order to exercise the NLP shell approach taken in TexUS, we are developing analyzers for Army SALUTE messages and texts provided by Federal Express. The most important validation of our approach was provided at MUC3. With only 2 man-months of customization, we achieved performance comparable to sites that devoted about one man-year of effort. Similarly, 3.5 man-months of customization for MUC4 have brought us to the same level as before, but with much greater potential for enhanced performance in the near term.

SYSTEM COMPONENTS

TexUS comprises the following major components: knowledge acquisition system, analysis system, knowledge base, knowledge management system, and primitive support system (see Figure 2).

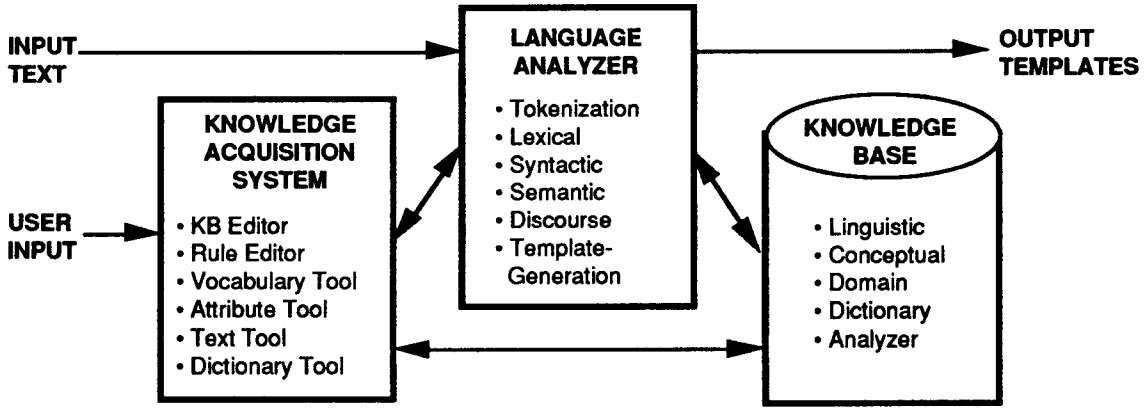

Figure 2. Overview of TexUS system

Analysis System

TexUS provides tailorable analysis capabilities. Rather than providing a monolithic analyzer that must be applied to all tasks, we have provided a set of functions that can be mixed and matched to easily construct an analyzer for a new task. Analysis functions perform tasks such as: (1) tokenize the input text, (2) perform morphological analysis and lexical lookup, (3) locate keywords and key phrases within a text, (4) apply rules to segment text, (5) apply rules to match segments of a text, and (6) perform semantic analysis on parse trees. Functions can apply top down or bottom up, a single or multiple times per node of the parse tree, recursively or not, and so on. Functions typically apply a hierarchy of grammar rules to the parse tree, so that augmenting the analyzer often consists of modifying or adding grammar rules, rather than code. Modifying the analyzer code most often consists of adding calls to existing functions, rather than implementing new functions.

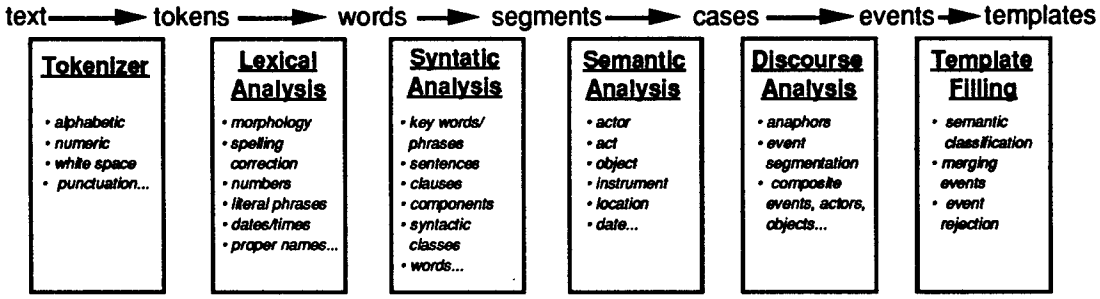

Figure 3. TexUS provides complete end-to-end analysis capabilities

The analysis algorithms are robust because they don't depend on complete characterization of the text in order to produce parse trees. When syntactic grammar rules apply, they are used in building the parse tree, but even when the text is ill-formed or the syntax knowledge of the system is incomplete, the analyzer produces parse trees from which useful information can be extracted.

Semantic analysis extracts and standardizes information by traversing the parse tree produced by the grammar-based analyzer. We have implemented semantic analysis capabilities to locate events, actors, objects, instruments, time, and location within a text. We are currently developing deep analysis capabilities that intelligently resolve discourse phenomena (e.g., whether two sentences describe the same or different events). The semantic analyzer employs

domain-specific rules whenever possible, and uses more generic knowledge when necessary, in order to extract relevant information.

The semantic analyzer constructs an internal representation of the objects and relationships between them. The discourse analyzer traverses the internal semantic representation to establish links between events, actors, objects, locations, dates, instruments, and other information extracted from text. The internal representation then serves as input to a task-specific conversion process that produces the desired output. We are investigating interactive tools to specify the output format. Figure 3 depicts the analysis passes implemented in TexUS.

Knowledge Acquisition Capability

The knowledge acquisition system provides a set of interactive graphic tools, including a hierarchy-based knowledge editor, a grammar rule editor, a vocabulary addition tool, and a dictionary tool. These tools allow a user to add lexical, syntactic, semantic, and domain knowledge to the system. The user can also build hierarchies that drive the analysis functions described earlier.

We have implemented automated knowledge acquisition capabilities that apply during analysis. For example, the analyzer applies patterns with one-word wildcards to categorize unknown words. A pattern such as "det quan WILD noun", when it matches a piece of text such as "the two rebel positions", leads the analyzer to hypothesize that "rebel" is an adjective or noun. Other mechanisms use morphological evidence and multi-word wildcards to characterize words and phrases. Expanding the automated knowledge acquisition capability is part of our ongoing research effort.

Batch knowledge acquisition tools have also been implemented, for example, to incorporate personal and geographic names into the knowledge base. We are investigating the use of on-line dictionaries to enlarge the knowledge base. We are already employing the Collin's English Dictionary (CED) provided by the ACL Data Collection Initiative for syntax class information, and are investigating the extraction of semantic information as well.

Knowledge Base

We have implemented a domain-independent knowledge management framework that improves the Conceptual Grammar framework of the predecessor VOX system. The knowledge base represents a variety of linguistic and conceptual knowledge, as well as housing the analyzer data structures and internal meaning representation data structures. Figure 4 exemplifies the knowledge representation for kidnapping concepts and their associated words.

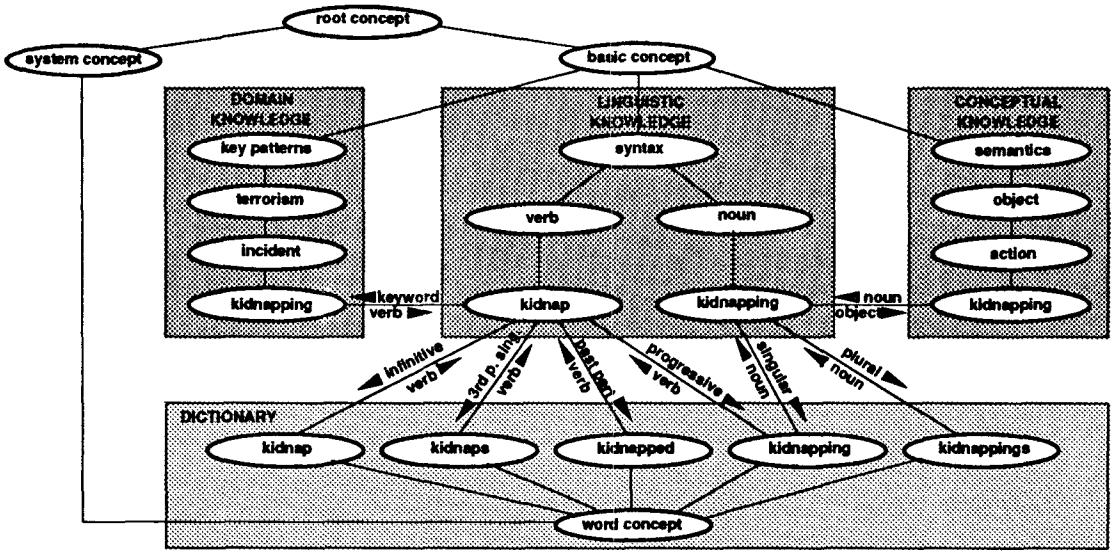

Figure 4. Knowledge representation example -- kidnapping concepts and words

The knowledge representation elements are concepts and grammar rules that are analogous to Lisp symbols and lists. Grammar rules represent any kind of sequential information; we use them for syntax rules, idiomatic phrases, attributes of concepts, patterns with wildcards, logical expressions, and so on. Attributes specify relationships between concepts, such as the parent-child relationship in a hierarchy.

The system's knowledge is stored in several forms. The raw database consists of knowledge in a form directly accessed and updated by TexUS. A second form of the knowledge consists of a set of files containing primitive knowledge addition commands (e.g., a command to add a node to a phrase). Executing the commands in this file system rebuilds the entire knowledge base from scratch. A third form of the knowledge consists of a file system of high level knowledge addition commands (e.g., a command to add a noun to the system). Each form of the knowledge provides a greater degree of independence from the system internals, and each is successively more human readable. The multiple layers of knowledge storage also provide extra knowledge protection, in case one layer is corrupted.

CURRENT WORK

Having completed an end-to-end analysis framework, our main tasks at present are to improve each pass of the analyzer and smooth the interactions between the passes. We are using the MUC development corpus to monitor our progress in fleshing out the analyzer and enhancing its performance. By early 1993, we will have completed a "beta release" of TexUS.

SYSTEM INFORMATION

TexUS has 89,000 lines of code and runs on Sun SPARCstations in C and Sunview. The system customized for MUC4 has about 1800 vocabulary words (not counting conjugations), with an additional 4,000 Hispanic names and geographic locations. The analyzer uses about 260 rules and processes text at about 2 words per second.

MUC4 DISCUSSION

That our current system is in transition is clearly evidenced by comparing the MUC3 and MUC4 scores. In fact, the rescored MUC3 results are better than those of the current system.

We have also implemented extensive automated testing facilities to augment the MUC scoring apparatus. We have used the testing system in preparing for MUC4 and will make extensive use of it during the remainder of 1992 to improve performance on the MUC task.

Processing of Message 0048 in TST2

Relevance filter: Keyword and key pattern search helps identify relevant portions of the text.

Lexical: The lexical pass is primarily concerned with identifying locations and names, and implements n-gram methods to decide if unknown words are English or not. To augment the system's vocabulary, the lexical pass made extensive use of the Collin's English Dictionary (CED). The set list of locations is used by the lexical pass, as is a set of personal names extracted from the development corpus. Spelling correction and morphological analysis algorithms also apply to unknown words.

For message 48, the lexical analyzer failed to find 'yet' in the CED, so n-gram analysis guessed that it is an English word. The word 'there' was absent from our core vocabulary, indicating the incompleteness of our coverage. The complete list of unknown words found in the CED for message 48 follows:

> abroad, accused, appointed, approve, armored, christian, closely, confirmed, considered, cordoned, credit, declared, democrat, drastic, elect, escaped, halt, including, intended, intersection, job, laws, legislative, linked, moments, napoleon, niece, noted, occasions, old, operation, possibility, prompt, reaction, replace, represent, responsible, roof, ruled, same, sources, stopped, street, termed, there, threatened, time, traveling, unscathed, warned

The spelling corrector converted "asssembly" to "assembly". Finally, all the names in the message, such as "Roberto Garcia Alvarado", were correctly determined.

Before and after lexical analysis, bottom-up passes through the message text located several types of idioms. Before lexical analysis, the following were found

Specific locations:	"San Salvador"
Date phrases:	"5 days ago"
Transition advs:	"also"
Brackets:	"[TEXT]"
Literals:	"Farabundo Marti National Liberation Front"
Verb idioms:	"took place", "carrying out"
Complex preps:	"according to"

After lexical analysis, the following were found:

Locations:	"downtown San Salvador"
Noun list	"government and military"

Erroneous findings were made by the noun-listing rule, which was recently added to the system:
"Roberto Garcia Alvarado and accused"
"police and Garcia"

Syntactic analysis: The syntactic segmentation algorithms have worked very well, considering the preliminary state of our knowledge base and the large degree of syntactic ambiguity supplied by the CED. For example, the first sentence in message 48 was segmented to the following components:

(np)	salvadoran president
(np)	- elect
(np)	alfredo cristiani
(vp)	condemned
(np)	the terrorist
(np)	killing
(pp)	of attorney general
(np)	roberto garcia alvarado and accused
(np)	the farabundo marti national liberation front (fmln)
(pp)	of the crime
(.)	.

In general, the assignment of np, vp, and pp was correct, even in this sentence. Lack of patterns such as <alphabetic> <hyphen> <alphabetic> led to the mishandling of "president-elect". "The terrorist killing" is difficult to assign correctly in general, and TexUS did well to assign the noun sense of "killing". As described for the lexical pass, "roberto garcia alvarado and accused" was misparsed because of a noun list pattern <noun> <and> <noun> that was overly unrestricted.

Semantic analysis: The semantic structures produced for the first sentence in message 48 derive directly from the syntactic segmentation shown above. We have edited the internal semantic representation to be human-readable:

event = condemned
actors = (1) Salvadoran president, (2) elect, (3) alfredo cristiani.
actions = (1) killing, (2) crime.
objects = (1) terrorist, (2) attorney general, (3) roberto garcia alvarado and accused, (4) fmln

The assignments are generally reasonable, except that merging of appositives and split noun phrases is not yet implemented. In the two weeks following the formal MUC4 test, we have improved the semantic analyzer to output separate event structures for nominal actions such as "the terrorist killing" and "the crime", so that adverbial information can be properly attached to these events.

After fixing some of the segmentation bugs noted earlier, the semantic output is greatly improved:

event = condemned
actor = (1) salvadoran, (2) president-elect alfredo cristiani.
object = the terrorist
event-object = killing

event = killing
object = attorney general

event = accused
object = fmln
event-object = crime

event = crime
actor = fmln

Discourse analysis: Discourse analysis links or separates semantic information based on syntactic, semantic, and discourse knowledge. In general, semantic information is separated or merged by comparing date/time, location, actors and objects. Actors and objects are classified as proper nouns, pronouns, or abstract nouns (e.g., "the home") and are compared by successively relaxing constraints on agreement, as in the syntactic and semantic passes. If the object being compared is the name "Garcia", then the first precedence for comparison will be other names such as "Roberto Garcia Alvarado" or "Garcia Alvarado" which contain the name "Garcia". If none are found, a proper name is then matched with pronouns such as "he" or "him". If that also fails, then "Garcia" is matched with abstract nouns such as "the attorney general". Time, location, and other concepts are compared similarly.

One construction not currently handled by the discourse analyzer is the phrase "Merino's home". The discourse analyzer does not yet link possessive nouns with other nouns in the corpus, which would help classify "Merino's home" as a GOVERNMENT OFFICE OR RESIDENCE instead of CIVILIAN RESIDENCE.

For meaningful work on message 48, discourse analysis depended on modifications to the earlier passes. In addition, we added a pragmatic rule that merged events based on location, allowing the attack on Merino's home to be merged with the fact that children were in the home at the time (sentences 11-13 of message 48). In general, the discourse process works well on MUC messages when the prior passes produce a correct internal semantic representation.

Template Output for Message 0048

SLOT	OFFICIAL OUTPUT	MODIFICATIONS
0. MESSAGE: ID	TST2-MUC4-0048	
1. MESSAGE: TMP	2	
2. INC: DATE	19 APR 89	14 APR 89
3. INC: LOC	EL SALVADOR: SAN SALVADOR (CITY)	
4. INC: TYPE	ATTACK	BOMBING
5. INC: STAGE	ACCOMPLISHED	
6. INC: INSTR ID	"EXPLOSIVES"	
7. INC: INSTR TYPE	EXPLOSIVE: "EXPLOSIVES"	
8. PERP: CATEGORY	TERRORIST ACT	
9. PERP: INDIV ID	"GUERRILLAS"	
10. PERP: ORG ID	-	
11. PERP: ORG CONF	-	
12. PHYS: ID	"MERINO ' S HOME"	
13. PHYS: TYPE	CIVILIAN RESIDENCE: "MERINO ' S HOME"	
14. PHYS: NUM	1: "MERINO ' S HOME"	
15. PHYS: FOREIGN	-	
16. PHYS: EFFECT	-	
17. PHYS: TOTAL	-	
18. HUM: NAME	-	
19. HUM: DESCR	-	"CHILDREN: "VICE PRESIDENT'S CHILDREN" "15-YEAR-OLD NIECE"

20. HUM: TYPE	-	CIVILIAN: "CHILDREN: CIVILIAN: "VICE PRESIDENT'S CHILDREN" CIVILIAN: "15-YEAR-OLD NIECE"
21. HUM: NUM	-	7: "CHILDREN: 4: "VICE PRESIDENT'S CHILDREN" 1: "15-YEAR-OLD NIECE"
22. HUM: FOREIGN	-	
23. HUM: EFFECT	-	INJURY: "15-YEAR-OLD NIECE"
24. HUM: TOTAL	-	

CONCLUSION

TexUS, substantially expands the capabilities of the system fielded at MUC3. The MUC4 scores of TexUS do not yet reflect the system's potential, a situation which we will rectify by the end of 1992. The 3.5 man-months of customization for MUC4 did not improve the overall performance over last year, because more passes of the analyzer must be upgraded for each new capability introduced into the system. Furthermore, of the total effort, at least half was devoted to corpus study rather than augmenting the analyzer and knowledge base.

Unlike MUC3, where our system had reached its full potential, TexUS is now a complete end-to-end framework which will serve to develop much stronger performance than we have shown at both MUC3 and MUC4. We look forward to exhibiting the system to good effect at the next MUC conference.

REFERENCES

[1] de Hilster, D. and Meyers, A. "McDonnell Douglas Electronic Systems Company: Description of the INLET System Used for MUC3". *Proceedings of the Third Message Understanding Conference*. DARPA. NOSC, San Diego, California. May 1991.

[2] de Hilster, D. and Meyers, A. "Heuristic Skimming of Voluminous Text". *Natural Language Text Retrieval Workshop*. AAAI, Anaheim, California. July 1991.

[3] Meyers, A., Knowles, A. and Ruoff, K. "Lexical Acquisition Tools in VOX". *Proceedings of the First International Language Acquisition Workshop*. IJCAI, Detroit, Michigan. August 1989.

[4] Meyers, A. "VOX -- An Extensible Natural Language Processor". *International Joint Conference on Artificial Intelligence*. UCLA, California. August 1985.