

Lexical Analysis, Semantic Sets and Computer Automated Semantic Classification

Les Bruce and Peter Wang

*Summer Institute of Linguistics
International Linguistics Center
Dallas, Texas*

1. Introduction

This paper introduces a demonstration of a simulated computer program tool for lexical analysis. The program leads the lexicographer through appropriate steps in his analysis and data entry. The basic theoretical principle underlying the tool is that an adequate semantic analysis and description requires that a given lexeme be contrasted with semantically similar lexemes in its own language system. The second principle applies to bilingual work; it states that a gloss must be distinguished from a definition. Due to differences between languages, a gloss in a second language usually cannot adequately express the meaning of a lexeme, even though such a gloss may serve as a translation equivalent in the right contexts.

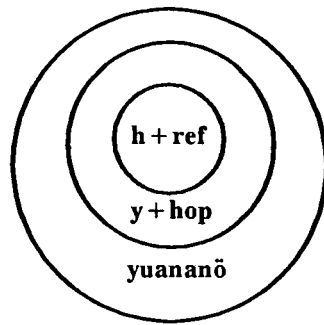
We are assuming the situation in which the semantics field worker is attempting to describe the semantics of lexemes by means of a second language. A lexicographer working on a bilingual dictionary may well find himself engaged in this type of description. Zgusta (1971:213) comments that descriptive work often characterizes a bilingual dictionary when the compilation of a monolingual dictionary is not anticipated soon by the lexicographer.

A field worker doing original analytical and descriptive work on a language which is not his mother tongue, typically goes through stages of refinement of his own understanding and expression of his analysis. Wierzbicka (1985:5) emphasizes the need for careful, thorough analysis and semantic descriptions in this type of situation because the "intuitive link between a word and a concept is missing" when concepts are defined in a foreign language. The language of description can interfere in the analyst's goal if he uncritically uses one-word glosses in his mother tongue to define lexemes in another language.

2. Contrastive analysis

It is essential that a lexeme be contrasted with all of the members of its set in a closed set of terms, or contrasted with many related terms in an open set. For example, to understand the number reference of pronouns in a language, one must know how many there are in the set. **They** refers to 'two or more' persons in English, but **nom** refers to 'three or more persons' in Alambalak of Papua New Guinea because its pronoun system includes singular, dual and plural values of its parameter of number. Similarly, **tomorrow** means something like 'the day following the present time' in English. **Y + hop** (Alambalak) must be contrasted with **yuananö** and other terms to determine its meaning, as described below.

A translation is based on maintaining the same reference in a particular context, but due to language diversity it usually will not serve to express the meaning of the term being translated. In different contexts **y + hop** can be translated 'tomorrow' and 'yesterday', but the meaning of the Alambalak term must be expressed as something like 'one day removed from the present'. The analyst is able to determine its meaning by contrasting **y + hop** with **yuananö** and other related expressions which helps define the term's range of reference (cf. Bruce 1984:86). A diagrammatic relationship among three Alambalak time words are presented in concentric circles:



The computer program described here aids semantic analysis by providing more than 660 semantic domains with which the lexicographer can classify each lexeme. The semantic classification forms links with each entry classed the same way. The program then provides the lexicographer with these sets of related lexemes with which the analyst can contrast new lexical entries and revise his definitions of existing entries.

3. Semantic Domains

The set of domains is taken from a Greek-English Lexicon produced by Dr. Eugene Nida, Johannes Louw and their editorial staff. The set is comprehensive, covering five thousand words, including twenty-five thousand senses for the ancient, Koine Greek dialect. Since it is a comprehensive set it provides a good starting point for a field worker to begin classifying the lexemes of a dictionary. It should provide a category for most of the concepts he will work on. The program enhances consistent classification by prompting the lexicographer with one or more semantic classes for each lexical entry based on the definition the lexicographer uses for the entry. Cross classification is encouraged.

The domains are included in the following general areas:

- I. UNIQUE REFERENTS (names of persons and places)
- II. CLASS REFERENTS
 - A. OBJECTS or ENTITIES
 - B. EVENTS
 - C. ABSTRACTS (including relationals)
- III. MARKERS
- IV. DISCOURSE REFERENTIALS (including pronominal and deictic expressions which refer to objects, events and abstracts,).

Louw and Nida have organized their domains in a two-level hierarchy. Only those domains in which Louw and Nida listed Greek terms have been used in this project. A sample of the 93 major domains and 663 subdomains is shown below. The single quote marks beside some of the capital letters in the outline are used as prime marks to distinguish repeated letters.

32. UNDERSTAND

- | | |
|---------------------------------------|--|
| A. Understand | C. Ease or difficulty in understanding |
| D. Capacity for understanding | |
| B. Come to understand | |
| E. Lack of capacity for understanding | |

33. COMMUNICATION

- | | |
|--------------------------------|---------------------------------|
| A. Language | O. Inform, announce |
| D'. Invite | R'. Mock, ridicule |
| B. Word, passage | P. Assert, declare |
| E'. Insist | S'. Criticize |
| C. Discourse types | Q. Teach |
| F'. Command, order | T'. Rebuke |
| D. Language levels | R. Speak truth, speak falsehood |
| G'. Law, regulation, ordinance | U'. Warn |
| E. Written language | S. Preach, proclaim |
| H'. Recommend, propose | V'. Accuse, blame |
| F. Speak, talk | T. Witness, testify |
| I. Intercede | W'. Defend, excuse |
| G. Sing, lament | U. Profess allegiance |
| J'. Thanks | X'. Dispute, debate |
| H. Keep silent | V. Admit, confess, deny |
| K'. Praise | Y'. Argue, quarrel |
| I. Name | W. Agree |
| L'. Flatter | Z'. Oppose, contradict |
| J. Interpret, mean, explain | X. Foretell, tell fortunes |
| M'. Boast | A". Prophecy |
| K. Converse, discuss | Y. Promise |
| N'. Foolish talk | B". Swear, put under oath, vow |
| L. Ask for, request | Z. Threaten |
| O'. Complain | C". Bless, curse |
| M. Pray | A'. Advise |
| P'. Insult, slander | D". Non-verbal communication |
| N. Question, answer | B'. Urge, persuade |
| Q'. Gossip | |
| | C'. Call |

34. ASSOCIATION

- | | |
|--|------------------------------------|
| A. Associate | D. Limit or avoid association |
| F. Visit | I. Kiss, embrace |
| B. Join, begin to associate | E. Establish or confirm a relation |
| G. Welcome, receive | J. Marriage, divorce |
| C. Belong to, be included in the membership of, be excluded from | |
| H. Show hospitality | |

This project does not assume or claim that the semantic domains are theoretical constructs of the semantic structure of a particular language or of the cognitive structure of the human mind. They are certainly not a universal set that can be expected to be applied equally well to all languages. In some places the set seems too finely grained, such as where Louw and Nida differentiate

- A. Learn
- B. Learn the location of something
- C. Learn something against someone
- D. Try to learn
- E. Be willing to learn
- F. Be ready to learn, pay attention
- G. Recognize within the generic category LEARN. In other cases some categories seem to be too broad. Concepts of metaphysical locations, for example, are included in the two domains: Regions above the earth; Regions below the surface of the earth.

These domains have been used by one of the authors for lexical classification in preliminary tests. Comprehensive testing with non-Indo-European languages is awaiting the completion of an English Thesaurus based on those domains which will enable the automation of the classification process. Despite these disclaimers, however, as stated earlier, it should provide a good starting point for a field worker to classify most of the lexemes of a dictionary.

4. The Computer Program

This lexicon processor tool will provide a semi-automated classification of each lexical entry. This development will include the following steps: the production of an English thesaurus based on the semantic domains; production of a computer program to match the definitions of lexical entries with one or more of the semantic domains; comprehensive testing on non-Indo-European bilingual dictionaries at varying stages of completion; translation of the thesaurus into other languages of wider communication; completion of the development of an object-oriented program environment within which the tool will run.

The work on just the thesaurus based on semantic domains is currently seventy-five percent completed.

Results of two preliminary tests conducted on 200 words from the Gurung language of Nepal (Glover, W.W., J.R. Glover and D.B. Gurung. 1977) is promising. The first test with the thesaurus

half completed is compared in the table below with a second test following further enrichment of the thesaurus and stripping of entries to reduce overlap of content between the domains. The tests comprised computerized searches throughout the thesaurus for the words in the definitions of 200 Gurung entries. Percentages of test one occur first followed by percentages of test two:

	Test One	Test Two
% of entries without any matching domain:	22%	9%
% of matched entries which matched 7 or fewer domains:	88%	92%
% of entries which were matched with an incorrect domain:	14%	9%
Average number of domains per entry for these entries which matched 7 or less domains (not including entries with no match)	2.9	2.28

A comprehensive test will be done upon the completion of the thesaurus.

Work has not yet begun on the program to be used for linking lexical entries with semantic domains. It is planned to develop a routine for matching key words of a definition with words in each domain of the thesaurus. The program will allow for multiple matching to enable the classification of entries in more than one domain. One of the goals for the program is to put no restrictions on the lexicographer as he formulates analytical definitions for the dictionary entries. This type of matching can be done by hand using a straight forward search of words in a database organized by records corresponding to the semantic domains. Such a routine must allow for a random order of words. This type of searching was used for the preliminary testing that has been done to date.

This program is being developed in an object oriented program environment. This type of environment was chosen to make the tool as interactive as possible. That is, the linguist will be able to tag data once with the relevant features he selects; the program environment then makes it possible to use that data and sections of data in a variety of ways, organizing the data in a variety of views for the linguist according to the tasks he wants to perform. The whole package should be ready for initial field testing within eighteen months. Suggestions for making it into a practical lexicographer's tool will be happily received by the authors in the meantime.

REFERENCES

- Bruce, L.P. 1984. *The Alambalak Language of Papua New Guinea*. Pacific Linguistics C-81, Canberra: The Australian National University.
- Glover, W.W., J.R. Glover and D.B. Gurung. 1977. Gurung-Nepali-English dictionary. *Pacific Linguistics*, Series C - No. 51, Canberra: The Australian National University.
- Louw, Johannes P. and Eugene A. Nida. 1988. *Greek-English Lexicon of the New Testament based on Semantic Domains*, Volume 1: Introduction and Domains. N.Y.: United Bible Societies.
- Wierzbicka, Anna. 1985. *Lexicography and Conceptual Analysis*, Ann Arbor: Karoma Publishers.
- Zgusta, L. (1971) *Manual of Lexicography*, The Hague: Mouton.