Compiling Dictionaries Using Semantic Domains*

Ronald Moe, *Linguistics Consultant, SIL International*[1] *(ron_moe@sil.org)*

Abstract: The task of providing dictionaries for all the world's languages is prodigious, requiring efficient techniques. The text corpus method cannot be used for minority languages lacking texts. To meet the need, the author has constructed a list of 1 600 semantic domains, which he has successfully used to collect words. In a workshop setting, a group of speakers can collect as many as 17 000 words in ten days. This method results in a classified word list that can be efficiently expanded into a full dictionary. The method works because the mental lexicon is a giant web organized around key concepts. A semantic domain can be defined as an important concept together with the words directly related to it by lexical relations. A person can utilize the mental web to quickly jump from word to word within a domain.

The author is developing a template for each domain to aid in collecting words and in describing their semantics. Investigating semantics within the context of a domain yields many insights. The method permits the production of both alphabetically and semantically organized dictionaries. The list of domains is intended to be universal in scope and applicability. Perhaps due to universals of human experience and universals of linguistic competence, there are striking similarities in various lists of semantic domains developed for languages around the world. Using a standardized list of domains to classify multiple dictionaries opens up possibilities for cross-linguistic research into semantic and lexical universals.

Keywords: SEMANTIC DOMAINS, SEMANTIC FIELDS, SEMANTIC CATEGORIES, LEXICAL RELATIONS, SEMANTIC PRIMITIVES, DOMAIN TEMPLATES, MENTAL LEXICON, SEMANTIC UNIVERSALS, MINORITY LANGUAGES, LEXICOGRAPHY

Opsomming: Samestelling van woordeboeke deur gebruikmaking van semantiese domeine. Die taak van die voorsiening van woordeboeke aan al die tale van die wêreld is geweldig en vereis doeltreffende tegnieke. Die tekskorpusmetode kan nie gebruik word vir minderheidstale waarin tekste ontbreek nie. Om in die behoefte te voorsien, het die skrywer 'n lys van 1 600 semantiese domeine opgestel wat hy suksesvol gebruik het om woorde te versamel. In 'n werksessie-omgewing kan 'n groep sprekers tot soveel as 17 000 woorde in tien dae versamel. Hierdie metode lei tot 'n geklassifiseerde woordelys wat doeltreffend uitgebrei kan word tot 'n volledige woordeboek. Die metode werk omdat die mentale leksikon 'n groot web is wat rondom sleutelbegrippe gestruktureer is. 'n Semantiese domein kan gedefinieer word as 'n belangrike konsep saam met die woorde wat direk daarmee verband hou vanweë leksikale verwantskappe. 'n Persoon kan die mentale web gebruik om vinnig van woord tot woord binne 'n domein te spring.

* This article was presented as a paper at the Seventh International Conference of the African Association for Lexicography, organized by the Dictionary Unit of South African English, Rhodes University, Grahamstown, 8–10 July 2002.

Lexikos 13 (AFRILEX-reeks/series 13: 2003): 215-223

Die skrywer is besig om vir elke domein 'n profiel te ontwikkel om te help met die versameling van woorde en met die beskrywing van hul semantiek. 'n Ondersoek van semantiek binne die konteks van 'n domein lewer baie insigte. Die metode laat die totstandbrenging van sowel alfabeties as semanties gerangskikte woordeboeke toe. Die lys domeine is bedoel om universeel in omvang en toepassing te wees. Moontlik as gevolg van universalia van menslike ervaring en universalia van taalkundige vermoë, is daar treffende ooreenkomste tussen verskillende lyste semantiese domeine wat ontwikkel is vir tale oor die hele wêreld. Die gebruik van 'n gestandaardiseerde lys domeine om veelsoortige woordeboeke te klassifiseer, skep moontlikhede vir kruislinguistiese navorsing oor semantiese en leksikale universalia.

Sleutelwoorde: SEMANTIESE DOMEINE, SEMANTIESE VELDE, SEMANTIESE KATEGORIEË, LEKSIKALE VERWANTSKAPPE, SEMANTIESE PRIMITIEWES, DOMEINPROFIELE, MENTALE LEKSIKON, SEMANTIESE UNIVERSALIA, MINDERHEIDSTALE, LEKSIKOGRAFIE

The problem (It's going to take forever)

The mental lexicon is far larger than either the grammatical component or the phonological component in a person's linguistic competence. Investigating and describing it is the largest and most time-consuming task in descriptive linguistics. With perhaps 6 000 languages in the world and perhaps 20 000 words in each, we need to collect and describe something on the order of 120 000 000 words.[2] The major languages of the world often have several large published dictionaries available to them. The major publishing companies can afford to hire scores of professional lexicographers to compile massive text corpora and do the research necessary to produce quality dictionaries. But for minority languages the picture is far bleaker. With few or no published texts, few or no professional lexicographers available to them, and little or no funding, the minority languages face a daunting challenge.

I have been involved in the production of dictionaries for minority languages since 1985 and have taught lexicography seminars to train others in the process. I estimate that linguists working in a language development project add words to their lexical database at the average rate of only 650 words per year, or about 2.5 words per working day.[3] At this rate it frequently takes 20 years to produce even a modest dictionary. For many years I have been concerned about this abysmal rate of progress and have attempted to find ways to make the process of compiling a dictionary simpler and more efficient. If we are ever going to finish the task of documenting the world's languages, we need a mass production technique.

The journey (Searching for a solution)

For several years colleagues within SIL, together with other interested scholars, have discussed ways in which we could leverage the linguistic similarities among the Bantu languages to facilitate linguistic investigation and language

development within the Bantu family. We have called this movement the 'Bantu Initiative'. In September 2000 the Bantu Initiative asked me to begin work on a dictionary template, including the production of a list of semantic domains that could be used to classify Bantu language dictionaries. I was a bit sceptical, since I had heard from numerous sources that the semantic category systems of the world's languages were vastly different, and even varied from individual to individual. But since the Bantu languages are closely related, I thought it was worth a try.

In order to construct a list of domains for Bantu languages, I needed to know how Bantu peoples categorized the words of their languages. So in December 2000 and January 2001 I held two workshops[4] for Gikuyu and Lugwere[5] in which I asked 12 speakers of each language to sort and group a list of 1 000 words chosen from a wide variety of semantic domains. I was curious to see how non-westernized peoples would classify the words of their language. My expectation was that they would set up very different domains than an English speaker. They didn't. Their domains were strikingly similar to other lists of semantic domains that I had collected from around the world. As I compared the lists, it became apparent that the universality of human experience and some sort of universal linguistic competence resulted in similar classification systems. The differences came from minor differences of culture and the necessity to squash a multi-dimensional system of relationships into a two dimensional list. So I decided (perhaps presumptuously) to attempt to compile a universal list of semantic domains.

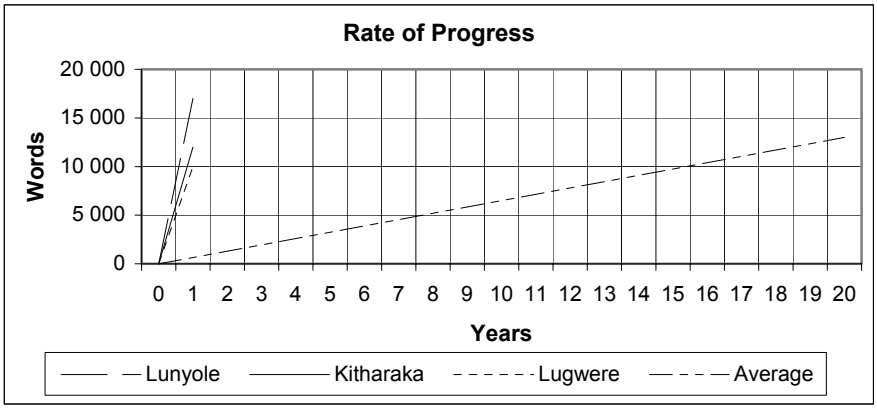
The challenge was to compile an exhaustive list of domains that could be used for any language in the world. None of the lists I had were complete. All were designed for a particular language and purpose. For instance, the *Outline of Cultural Materials* (Murdock et al. 1987) presents a list of anthropological domains, but is missing many lexical domains. *Roget's Thesaurus* (Roget 1958) has 1 000 domains, but due to its purpose it also omits many domains. Newer editions of *Roget's* (e.g. Morehead 1985) contain 600 major domains and thousands of smaller entries. Neither presentation is suitable for our purpose. Louw and Nida (1989: xix) admit that their list is uneven due to the subject matter of the New Testament. Recent semantically organized dictionaries such as the *Longman Language Activator* (Summers 1993) and the *Oxford Learner's Wordfinder Dictionary* (Trappes-Lomax 1997) are highly selective in the domains they include. So I concluded that a new list was needed. I contrasted and compared all the lists at my disposal, ensuring that every domain in every list was covered by a domain in my list. As I studied the organization of the lists, more and more similarities began to emerge. There was a logic to the domains, and a logic to how they were organized.

I knew from the beginning that a list of semantic domains could be used to collect words. Eliciting vocabulary has been a topic of interest for some time, and the literature contains a wealth of practical suggestions, such as using lexical relations (Beekman 1968: 4), concording a text corpus (Naden 1977: 14), and

using semantic domains (Newell 1986: 20).[6] I decided to try it out and see just how easy it would be. I took the semantic domain 'Bodies of water' and started listing words that belong to the domain (e.g. *ocean*, *lake*, *river*, *shore*, *wave*, etc.). In fifteen minutes I had collected and subcategorized 169 words. The rate for collecting words had just jumped from 2.5 words/day to 11 words/minute. I realized that all I needed was a list of semantic domains and I could collect the words of a language in a matter of days rather than years.

As I thought about how the list of domains could be used to collect words, I realized that a simple domain label, such as 'Bodies of water', would not be adequate. Three things were needed: (1) a simple statement of the central idea of the domain, (2) elicitation questions that would prompt a person to think of words that might belong to the domain, and (3) sample words from English.[7]

I have tested the materials and method in three workshops. The first test, held in May 2001, used a beta version of the semantic domains list with a group of fifteen speakers of the Lugwere language. In ten days, the participants collected over 10 000 words and 1 000 example sentences.[8] In January 2002, 30 speakers of Lunyole used version one to collect 17 000 words in ten days. In February 2002, 12 speakers of Kitharaka[9] collected 12 000 words in eight days. In the months since the workshops, speakers of each language have been editing and glossing the word lists. As the result of a few months work, we expect to have a classified dictionary in each language of over 10 000 words, including part of speech, noun class, the plural form of each noun, and a simple gloss. The chart below compares the historical average rate of progress with the results of the three workshops.



Why does it work? (Semantic domains, lexical relations, and semantic primitives)

The field of semantics has yet to reach a consensus on the nature and validity of semantic domains and semantic primitives. 'Semantic domain' is just another

way of saying 'area of meaning', but the notion that a meaning occupies an area is obviously figurative. Wierzbicka (1996: 170) comes close to endorsing the notion of universal semantic domains when she says: "The idea that words form more or less *natural* groupings, and that at least some of these groupings are *non-arbitrary*, is intuitively appealing, even *irresistible*" (emphasis added). She also indicates that domains vary in their nature from "self-contained fields of semantically related words" to "irregular and open-ended networks of interlacing networks". The question remains — just what is a semantic domain?

I envisioned that the list of domains would serve several purposes. It could be used to collect words, it could serve to classify a dictionary, and it could aid in semantic investigation. In order for it to be an effective tool in collecting words, I felt I should list sample words from English that belong to each domain. As I analyzed the words that I was listing under each domain, and compared them to the words others had included in the same domain, I began to see patterns. Some domains consisted of a generic term, such as 'Game', and a list of specifics: *chess, checkers, charades, monopoly*. Others were based on the Whole-Part lexical relation, such as 'Head' and *eye, nose, mouth*. Other domains included a variety of words related by different lexical relations, such as 'Wave' and *tidal wave, crest, break, roar, surfboard*.

It became apparent that a semantic domain was really some important concept and all the words directly related to it by some lexical relation. The words of a language are all linked together in the mind in a gigantic multi-dimensional web of relationships. But these mental links tend to cluster around a central nexus. A semantic domain isn't so much an area of the web as it is one of these central hubs. One of the intriguing questions about these hubs is: What is their relationship to semantic primitives? Many domains appear to be based on semantic primitives or a combination of two or three primitives (e.g. 'Bad behavior' = do + bad; 'Parts of things' = part (of) + something). Many are headed by high frequency words which constitute the core vocabulary of a language.

Several recently published dictionaries employ a "defining vocabulary". For instance, the *Longman Language Activator* (Summers 1993) lists the 2 222 words of its defining vocabulary in an appendix. When one excludes the functors (e.g. the, to, of), what is left is very similar to a list of domains. The notions of "semantic domain", "semantic primitive", "core vocabulary", and "defining vocabulary" seem to be converging.

As I developed the list, I began organizing the sample English words into lexical sets. I found that each lexical set was related to the central idea of the domain by a single lexical relation. I have already mentioned that lexicographers recommend that we employ lexical relations in collecting words. This seemed like a very useful idea in the light of what I was discovering. However, lexical relations are very hard to grasp in the abstract (e.g. $Conv_{13}$ (buy) = sell (Grimes 1987: 27)). Grimes (1994) has attempted to make lexical relations more user-friendly. But there are so many of them[10] that it is extremely inefficient to

have to think through the entire list of lexical relations for each new word encountered, in order to determine which ones might be productive. So I felt it was best if I thought through the list and identified which lexical relations were productive for each domain. I worded each productive lexical relation in the form of a simple question. For example, the domain 'Sing' has the following productive lexical relations:

What words refer to singing? *sing, serenade, warble, yodel, burst into song*
What words refer to singing without using words? *hum, whistle*
What words refer to a person who sings? *singer, vocalist, soloist*
What words refer to a group of people singing together? *choir, chorale, singing group, duet, trio, ensemble*
What words refer to something that is sung? *song, singing, tune, melody*
What types of songs are there? *lullaby, hymn, psalm, carol, national anthem, lament, ballad*
What words refer to a part of a song? *verse, chorus, theme, note, melody, harmony*
What words describe how well a person sings? *beautiful singing voice, can't carry a tune in a bucket, sing on/off key, monotone*
What words describe how high or low a person sings? *pitch, soprano, alto, baritone, bass*
What words describe whether or not people are singing the same thing together? *sing in unison, sing in harmony, sing the melody/harmony*

The questions and sample words are not meant to be exhaustive. It doesn't take much effort to think of other words. In practice, it has turned out that the combination of semantic domains and lexical relations is extremely productive. The mind quickly jumps from one word to another along the mental paths formed by lexical relations.

What do we need? (Domain templates)

Atkins (1997) has recommended that lexicographers produce a template for each lexical set they are investigating. She points out that a template enables the lexicographer to gather information faster, prompts the lexicographer to look for common features, and makes the approach to the whole lexical set much more systematic. I believe we could produce universal templates, which would be based on cross-linguistic research and would present features that the lexicographer would be likely to encounter in each domain.

Thus far I have worked to identify the lexical relations that are productive in each domain and have listed sample words from English. My purpose is to produce a tool which can be used to collect words. Here is an example:

2.4.1 See
What words refer to seeing something (in general or without conscious choice)? *see, behold, come into view*

What words refer to consciously looking at something? *look at, view, observe, scan*
What words refer to looking at something in order to learn? *watch, scrutinize*
What words are used of looking at something for a long time or in amazement? *stare, gaze, gape, gawk*
What words are used of looking at something for a short time? *glance, cursory glance, look at briefly, (eyes) flicker over*
What words refer to the sense of sight? *sight, sense of sight, vision*
What words refer to someone who sees? *observer, beholder, witness*
What words refer to a group of people who are watching something? *audience*
What words refer to what is seen? *sight, view*

Once the members of a lexical set are identified, we can identify the semantic features which distinguish them. For instance, the English words which belong to the domain 'Movement' often incorporate a component of direction, such as *advance* (front), *retreat* (back), *step aside* (side), *climb* (up), and *descend* (down). Other components include manner (*walk, run, jump*), beginning or ending point (*leave, arrive*), and medium (*fly, swim*). Once we have investigated the semantics of this domain for several languages, commonly occurring features can be noted. So the template for 'Movement' would prompt the researcher to look for these components. We could also include sample definitions, pragmatic and cultural issues to look out for, possible subcategorizations, and possible variations in the conceptualization of the domain. As each template is enriched, its usefulness will grow.

Where to from here? (Semantic universals and beyond)

Using semantic domains to produce a dictionary has numerous benefits in addition to speeding up the process of collecting words. We can sort our computer databases alphabetically or by domain. Translators and writers need lists of related words to facilitate composition. We can produce semantically organized dictionaries, such as the *Longman Language Activator* (Summers 1993), the *Oxford Learner's Wordfinder Dictionary* (Trappes-Lomax 1997), and the *Greek-English Lexicon of the New Testament, Based on Semantic Domains* (Louw and Nida 1989). Or we can publish alphabetical dictionaries and include an appendix of domains.

It is far more insightful to study the members of a lexical set together than to study them in isolation. As Wierzbicka (1996: 170) has pointed out: "Although the meaning of a word does not depend on the meanings of other words, to establish what the meaning of a word is one has to compare it with the meanings of other, intuitively related words."

Wierzbicka concludes her chapter on semantic primitives and semantic fields by saying: "I think, therefore, that the semantic primitives approach to semantic analysis also offers a necessary firm ground for the study of semantic

fields" (1996: 183). I would agree, and add that the study of semantic fields is necessary for the study of semantic primitives and universals.

The existence of the International Phonetic Alphabet permits cross-linguistic comparisons of phonological systems. The existence of (fairly) standardized grammatical categories allows us to search for universals of grammar. Anthropology has the *Outline of Cultural Materials* (Murdoch 1987). Chemistry has the periodic table. What does semantics have? I suggest that we cooperate to produce a standardized list of semantic domains. Such a list would enable us to do cross-linguistic comparisons and search for linguistic universals in the field of semantics, just as our colleagues are doing in the fields of phonology and grammar. What I have done is only a poor first attempt in this direction, but I hope it will lead to productive avenues of research.[11]

Endnotes

1. SIL International (the Summer Institute of Linguistics International) is an organization of volunteers, devoted to the promotion and development of minority languages. SIL International works in over 50 countries and over 1 000 languages.
2. In the interests of simplicity and naturalness, if not accuracy, this article employs the term 'word' to refer to lexical items of all sorts, including roots, derivatives, compounds, idioms, and phrases.
3. This estimate is based on observation of the number of years it has taken to produce published dictionaries, both within and outside of SIL, and has been confirmed by numerous SIL colleagues.
4. Thanks are due the Bantu Initiative for funding these workshops.
5. Both languages are Bantu. Gikuyu is spoken in Kenya, and Lugwere in Uganda. Dr. Mary Muchiri of Daystar University organized the Gikuyu workshop, and Dr. Ruth Mukama of Makerere University the Lugwere workshop.
6. Ideally lexicographic research should utilize both semantic domains and a concordance. However, unless a computerized text corpus running into the millions of words is available, using a list of domains is the only effective way of collecting words. If no corpus is available, it would be good to begin collecting or producing one.
7. These materials are currently being translated into Swahili, and plans are to have them translated into French, Spanish, Chinese, and other major languages of the world.
8. By comparison many bilingual dictionaries are published with only 3 000–5 000 entries.
9. All three languages are Bantu. Lugwere and Lunyole are spoken in Uganda, and Kitharaka in Kenya.
10. In fact, there are far more than the literature would suggest. It is apparent that lexical relations are not all the same sort of thing. I believe that lexical relations are based on similarities of meaning, and are as varied as the meanings of words.
11. Copies of the author's list of semantic domains and related materials are available from him via email at ron_moe@sil.org. The materials are also available in Swahili.

References

Atkins, Sue. 1997. Template Entries. Unpublished SALEX seminar handout.

Beekman, John. 1968. Eliciting Vocabulary, Meaning, and Collocations. *Notes on Translation* 29:1-11.

Grimes, Charles E. 1994. Mapping semantic relationships in the lexicon using lexical functions. *Notes on Linguistics* 66: 5-25.

Grimes, Joseph E. 1987. Relations and Linkages in the Lexicon. Prepublication Draft. Dallas: Summer Institute of Linguistics.

Louw, Johannes P. and Eugene A. Nida. 1989. *Greek-English Lexicon of the New Testament: Based on Semantic Domains, Vol. 1*. New York: United Bible Societies.

Morehead, Albert H. (Ed.). 1985. *The New American Roget's College Thesaurus in Dictionary Form*. New York: Signet.

Murdock, George P., et al. 1987[5]. *Outline of Cultural Materials*. New Haven: Human Relations Area Files.

Naden, Tony. 1977. *Words and Meanings*. Ghana: Institute of Linguistics.

Newell, Leonard E. 1986. *Lexicography Notes*. Manila: Summer Institute of Linguistics.

Roget, Peter Mark. 1958. *Roget's Thesaurus*. Harmondsworth, Middlesex: Penguin Books.

Summers, Della (Ed.). 1993. *Longman Language Activator*. Essex: Longman.

Trappes-Lomax, Hugh. 1997. *Oxford Learner's Wordfinder Dictionary*. Oxford: Oxford University Press.

Wierzbicka, Anna. 1996. *Semantics, Primes and Universals*. Oxford: Oxford University Press.