Using Concept Maps in NDLTD as a Cross-Language Summarization Tool for Computing–Related ETDs

Ryan Richardson (ryan@vt.edu), Edward A. Fox

Department of Computer Science, 2050 Torgersen Hall 0106, Virginia Tech, Blacksburg, VA 24061

Abstract heading

Concept maps, introduced by Novak, aid learners' understanding. We hypothesize that concept maps also can function as a summary of large documents (e.g., ETDs). Our system automatically generates concept maps from English-language ETDs in the computing field. The system also will provide Spanish translations of these concept maps for native Spanish speakers. Because of the results of our enhanced machine translation techniques, we believe concept maps could allow researchers to discover pertinent dissertations in languages they cannot read, helping them to decide if they want a potentially relevant dissertation translated.

We are using a state-of-the-art natural language processing system, called Relex, first to extract noun phrases and noun-verb-noun relations from ETDs, and then to produce concept maps automatically. We also have incorporated information from the table of contents of ETDs to create novel styles of concept maps. Currently we are producing concept maps for the Virginia Tech CS collection (175 ETDs), which covers a broad range of computer science topics. We intend to automatically produce concept maps for computing-related ETDs for a larger segment of the NDLTD holdings. We have recently conducted two user studies, to evaluate user perceptions about these different map styles.

We are using several methods to translate node and link text in concept maps from English to Spanish. Nodes labeled with single words from a given technical area can be translated using word lists, but phrases in specific technical fields can be difficult to translate. Thus we have amassed a collection of about 580 Spanish-language ETDs, from Scirus and two Mexican universities, and we are using this corpus to mine phrase translations that we could not find otherwise.

We also have tested the usefulness of the automatically-generated and translated concept maps in a user experiment conducted at Universidad de las Americas (UDLA) in Puebla, Mexico. This experiment provides insights regarding if concept maps can augment abstracts (translated using a standard machine translation package) in helping Spanish speaking users find ETDs of interest.

Motivation

The growth of the World Wide Web has led to increased availability of many large documents, such as electronic theses and dissertations (ETDs). In fact, NDLTD [10] has made over 300,000 ETDs available in at least 12 different languages. However, it is difficult for users to determine if such large documents, especially those written in a language they cannot read, are relevant to their information needs—so they can seek a translation. Unfortunately, automatically translating large documents, like ETDs, so they easily can be found, read, and understood, is beyond the current state of the art in machine translation (MT), especially in technical fields.

A goal of our research is to make ETDs more readily accessible as an information resource for students and researchers. Some challenges are that ETDs tend to be very long, and often only one section will pertain to a user's information need. Also, we would like to make it easier for users to determine if an ETD, or part of an ETD, is relevant to their information need, even if they cannot read (or have difficulty reading) the language in which the ETD is written. If the user determines that the ETD is relevant, they can either attempt to read the original ETD (if they have some language proficiency in the original language), and/or they can have part or all of the ETD professionally translated.

Concept Map Generation

We have conducted experiments in producing concept maps [11] involving statistical techniques such as log likelihood [6] , t-score, and association rules [2]. We determined these

to be inadequate and decided to apply more advanced natural language processing techniques. We adapted a tool named Relex [12], which was developed at VettaLabs in Brazil, under the direction of Ben Goertzel (at Virginia Tech). It is based on ideas from Minipar by Dekang Lin at the University of Alberta [8]. Relex is a system that translates syntactic dependencies into a graph of semantic primitives [13], by means of template matching algorithms. Relex is implemented within IBM's UIMA framework [7]. This framework describes a series of design patterns, interfaces, and metadata to implement, combine, and deploy analysis capabilities. The default entity tagger used with Relex is based on the Another Nearly-New Information Extraction (ANNIE) system, distributed as part of the General Architecture for Text Engineering [5].

The base version of Relex has two customized entity taggers, one specialized for the biomedical domain and another for world news and finance. Obviously neither of these was appropriate for our purposes regarding computer science ETDs. For Relex to be able to recognize computing terms as entities, we needed a comprehensive source of terminology.

We fulfilled this requirement by using term lists from the Ontology Project [4], under development at Villanova University, and sponsored by ACM. This project has divided computing into 21 topic level domains, and provides a hierarchy of terms that go from one to six levels deep for each of these main topic areas. The ANNIE extractor needs to recognize when an instance of a class in the ontology appears in the text. We accomplished this by selecting various nodes in the ontology and encoding "gazetteers" – lists of individual terms that would be the surface representations of that node in the document. Since the computing ontology has about 900 leaf nodes, it would be very time-consuming to write gazetteers for all of them. Therefore we selected a few areas of computer science for which Virginia Tech has a large number of ETDs (i.e., digital libraries, human-computer interaction, virtual environments) and wrote gazetteers for the 244 nodes involved. The gazetteer currently contains over 3,200 computing related words and phrases.

Due to the large size of ETDs, we split each ETD into chapters and use Relex to produce a concept map for each chapter. We also produce a "top-level" map, which has the title and author of the ETD as well as a node for each chapter. Each chapter node is a clickable link to the concept map for that chapter. The maps were written out to the CXL format used by IHMC CmapTools [3]. Then, for relations in concept maps produced with the help of Relex, if a user hovers over the link text, the sentence in the original document that produced that link will be shown.

Monolingual ETD Experiment

Previous tests with users revealed that including part or all of the table of contents (ToC) of ETDs improved the quality of the maps. Thus we wished to conduct an experiment to determine if the concepts found by Relex really added value to the concept maps. We conducted a user study with three different styles of automatically-generated maps.

A) Maps generated using only ToC information (called ToC maps)
B) Maps with only Relex-found terms from the computing ontology, plus chapter titles (called Relex-only maps)
C) Maps which combine types A) and B) above (called ToC+Relex maps – see Figure 1)

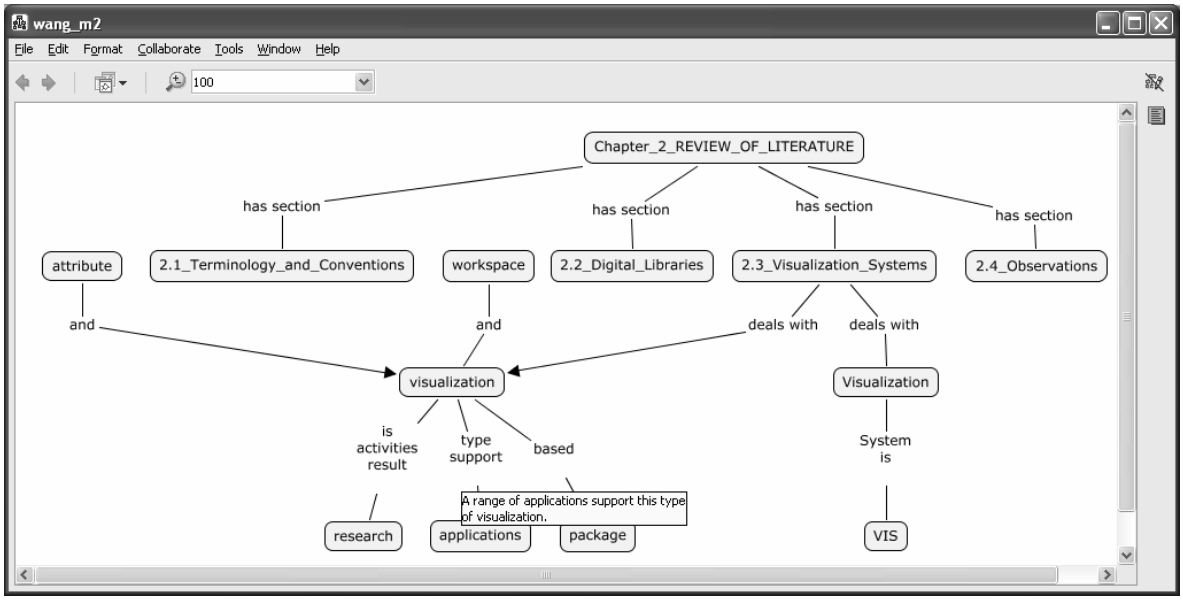

Figure 1: Concept map of Chapter 1 of the ETD by Jun Wang about VIDI, in style C. The hovertext shows the sentence from the original ETD that contains the relation.

Thirty-five subjects participated in the experiment. Subjects were presented with an ETD in electronic form and were given 20 minutes to skim the ETD. Subjects were asked to rate the concept maps on a 5-point Likert scale for key attributes: node selection, link selection, whether the relationships were important in the ETDs, helpfulness of hovertext, and overall usefulness in determining what the ETD is about. Subjects then repeated the process on a second ETD. We performed paired t-tests comparing the layouts in styles A, B, and C, based on these criteria. The presentation order was alternated, such that half of the subjects saw A first, the other half saw B first, etc.

The results were that both Relex and ToC+Relex concept maps were rated significantly higher by users than the ToC concept maps ($p=0.05$). Relex and ToC+Relex maps were rated almost the same by users across the board, with the only significant difference being that users thought that the hovertext for the Relex-only maps was more informative. For ToC+Relex maps, there are two types of hovertext. For a link between a chapter and a section, or between a section and a Relex-found term, the hovertext shows the first 200 characters of the text of that section. For relations between Relex-found terms, the hovertext shows the sentence in the original thesis where both of these terms occur. The reasons that users preferred the hovertext between Relex-found terms were not clearly explained in the user comments.

Cross-language Experiment

To test the usefulness of automatically-generated concept maps as cross-language summaries, we conducted a study involving concept maps and abstracts, with students at the Universidad de las Américas (UDLA), Puebla, Mexico. For the experiment, our three hypotheses were as follows:

1) Automatically-generated concept maps can be a useful summary of an ETD.
2) Automatically-generated concept maps can augment abstracts in helping subjects determine if a document is relevant to an information need.
3) Automatically-generated concept maps can be translated via MT "well enough" so that they can be used as a cross-language information discovery tool.

The experimental materials were based on 30 dissertations from the Virginia Tech computer science ETD collection. For 15 of these, the abstracts and concept maps were

translated by a paid professional translator familiar with IT/Computing. For the other 15, the abstracts and concept maps were translated by MT.

Automatic Translation

We developed a CS-related corpus which is comparable to the VT-CS corpus, by combining computing ETDs from UDLA, the Universidad Nacional Autonoma de Mexico (UNAM), and Spanish ETDs identified through Scirus. Our corpus contains over 500 ETDs. We used this collection as a source of phrase translations into Spanish, using an enhanced version of an algorithm developed by Lopez-Osteñero [9].

The English concept maps were translated into Spanish using several translation resources. For a given concept, our software first checks if the word or phrase occurs in a 47,000 entry word and phrase list, produced by data mining at the University of Maryland. The Maryland list contains multiple possible Spanish translations for each English word/phrase, so we reordered the Maryland wordlist based on frequency in our Spanish ETD collection. For instance, both "web" and "tela" (spider-web) are possible Spanish translations of the English word "web". Since "web" occurs more often in the our Spanish computing ETD collection, occurrences of the English word "web" are translated as "web", not "tela". We supplemented the Maryland word/phrase list with translations from the ACM CC2001 classification scheme [1], provided by Fernando Das Neves.

If an English phrase does not occur in the Maryland wordlist, our software checks in the list of phrases that were mined from the Spanish ETD collection using our implementation of the Lopez-Osteñero algorithm. If it is still not found, our software sends the word/phrase in question to Systran.

For this experiment, instead of using CmapTools, we decided to employ a simpler implementation of concept maps using AT&T's Graphviz software. This produced JPEG images, embedded in HTML pages. The main reason was that the students at UDLA had never used CMapTools before, and so would need time for training before becoming comfortable with the software. Since this was a remote experiment, we decided it would be unreliable to make the students take an online tutorial, especially since there would be no one present in Mexico who could demonstrate the software to them. Therefore we decided to use a more familiar interface, i.e., HTML with embedded JPEG images.

We also omitted the hover-text feature that we used in the previous experiment. We did this since the original document is in English, and we did not have sufficient resources to have all of these sentences translated into Spanish. For the cross-language experiment, we wanted everything shown to the subjects to be in Spanish, so that subjects who can read English fluently did not have an advantage.

The following figures (i.e., 2-4) are of concept maps that were used in the cross-language experiment. All materials presented to the subjects in the experiment were in Spanish.

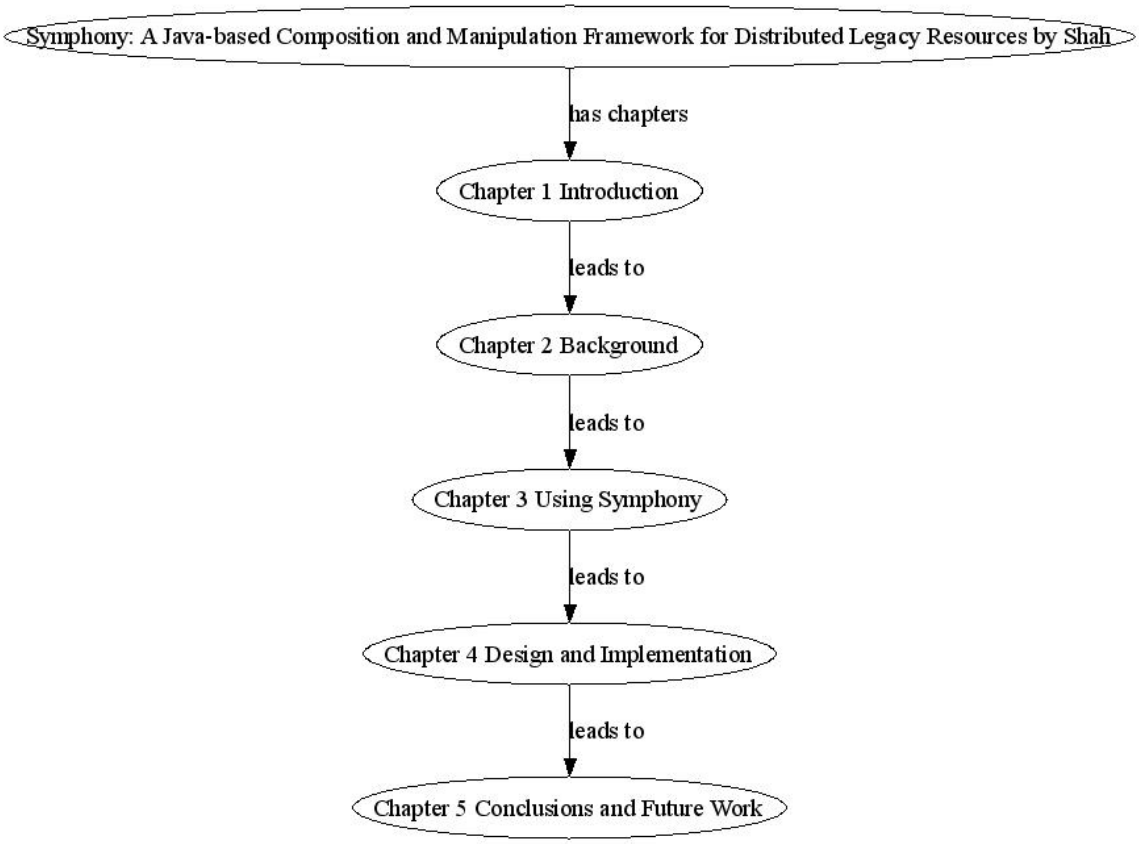

Figure 2: Overview map of ETD by A. Shah (English version)

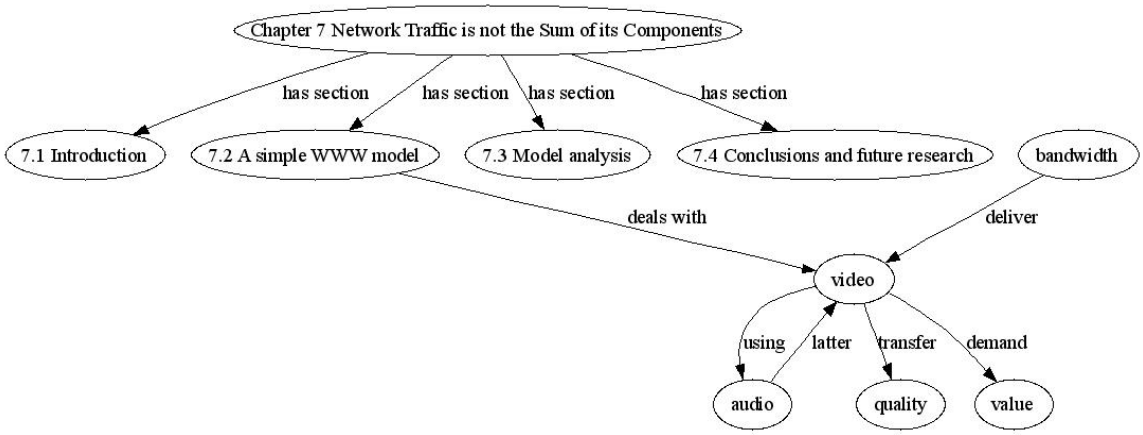

Figure 3: Concept map of ETD by G. Abdulla, chapter 7 (English version)

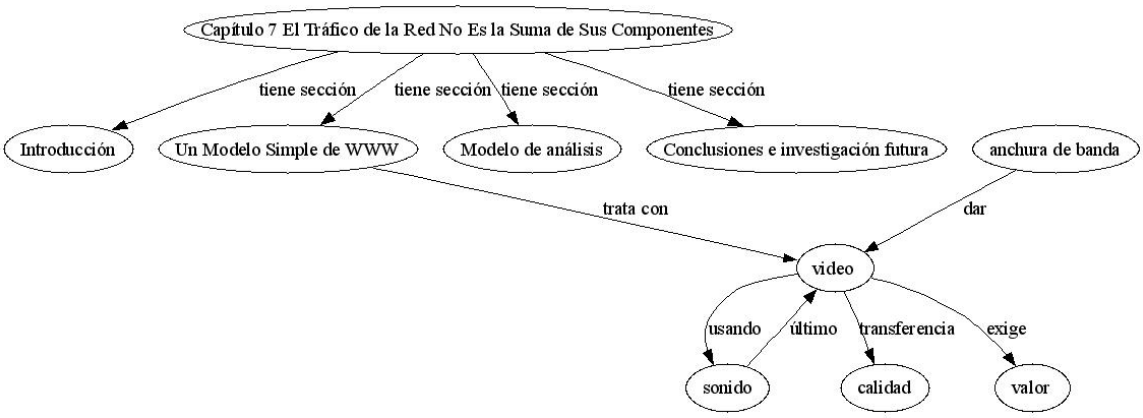

Figure 4: Concept map of ETD by G. Abdulla, chapter 7, machine translated into Spanish

Experimental Conditions

Subjects were asked 6 questions about which dissertations were relevant to a particular question. A domain expert came up with these questions, based on the 30 English ETDs, and listed which ETDs he considered relevant. Two more domain experts made their own relevance determinations about these dissertations for comparison purposes. The domain experts looked only at the English versions of the documents and concept maps.

In order for the Mexican students to answer the questions, for each question they were provided one of 6 types of summaries of the dissertations (see Table 1).

Table 1: Six Treatment Conditions for Cross-Language Experiment

1. Human translated abstract (A1)	4. Machine translated abstract (A2)
2. Human translated concept map (B1)	5. Machine translated concept map (B2)
3. Human translated abstract + Human translated concept map (C1)	6. Machine translated abstract + Machine translated concept map (C2)

Each subject was presented with 6 questions, one in each of the treatment conditions. For each question, they were allowed to pick from 5 dissertations. For instance, they were given a relevance question, and presented with 5 machine-translated concept maps (condition B2 above), based on 5 dissertations, and were asked which dissertations were relevant to that particular question based just on these concept maps.

Each question had between 1 and 3 ETDs that are relevant to it. Each of the 6 questions always had the same 5 ETDs as possible answers. These 6 sets of ETDs were disjoint (hence, 6 questions, and 5 dissertations for each, yielded 30 ETDs). Thus a subject never saw abstracts/concept maps of the same ETD for different questions. In fact, each subject was presented with each ETD exactly once.

There were 22 students in the class at UDLA who participated. Presentation order was randomized. Counting the relevance determinations of the three experts as a 'gold standard', we can compare the subjects' effectiveness at determining relevance of an ETD to a given computing-related topic.

Results

In the human-translated condition, users presented with concept maps, or with abstracts plus concept maps, had significantly greater agreement with experts than those presented only with abstracts ($p=0.05$). In the machine-translated condition, users presented with concept maps had significantly greater agreement with experts than those presented with abstracts ($p=0.05$). The results are summarized below (see Table 2).

Table 2: Agreement between Subjects and Experts for 6 treatment conditions from 0 (no agreement) to 5 (perfect agreement).

	Abstract	Concept Map	Abstract + CM
Human translated	3.23	3.95	3.68
Machine translated	3.00	3.91	3.41

Interestingly, in the machine-translated condition, users presented with abstracts and concept maps did not perform significantly better (average = 3.41) than those presented only with abstracts (average=3.00). Since the machine-translations were of lower quality than human ones, perhaps being presented with more information, because the added information was of questionable quality, was confusing to the subjects.

Future Work

We plan to further investigate the effectiveness of providing links to the context of the original document, using techniques in addition to just hovertext. This could be done with the help of tools developed to support superimposed information [9].

We are investigating to make our concept-map generation and translation scale to full collections, with the eventual goal of making the entire Virginia Tech CS-ETD collection, and perhaps collections from other universities, available in concept map form, including in other languages.

Acknowledgements

This research work was funded in part by NSF through grants DUE-0121679 and IIS-0080748, and also by a grant from IMLS. We thank Ben Goertzel, Hugo Pinto, and Vicente Cordeiro for their help with modifying Relex. Thanks go to Lillian Cassell and her team at Villanova for their work on the Ontology Project. We thank Craig Scott for making the Scirus collection available. We also thank Alberto Castro Thompson and Silvia González Marín at UNAM for making their ETD collection available to us. We thank Alfredo Sanchez for making the UDLA ETD collection available, and (along with Gabriel Torres) for making their UDLA students available for the cross-language experiment. Special thanks go to Fernando Das Neves for his proofreading and translation work for the UDLA experiment. Thanks to Seungwon Yang, Seonho Kim, and Doug Gorton for being our domain experts in digital libraries. Finally, thanks go to Greg Hill, Larry Bunch, Carlos Perez, Tom Eskridge, Roger Carff, and all the rest at the Institute for Human and Machine Cognition for their invaluable help interfacing with CMapTools.

References

[1] ACM, IEEE-CS, and AIS. "Curriculum Recommendations of the ACM, IEEE-CS, AIS": ACM, 2005. http://www.acm.org/education/curricula.html.

[2] Agrawal, R., Imielinski, T., and Swami, A. *Mining association rules between sets of items in large databases*. Presented at ACM SIGMOD Conference on Management of Data, Washington, D.C., May, 1993.

[3] Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., Eskridge, T. C., Arroyo, M., and Carvajal, R. *CMapTools: A Knowledge Modeling and Sharing Environment*. Presented at First Int. Conference on Concept Mapping, Pamplona, Spain, 2004.

[4] Cassel, L. N. Ontology Project, 2006. *what.csc.villanova.edu/twiki/bin/view/Main/OntologyProject*.

[5] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. *Framework and Graphical Development Environment for Robust NLP Tools and Applications*.

Presented at 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, 2002.

[6] Dunning, T. *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics, 19(1), pp. 61-74, 1993. *http://citeseer.ist.psu.edu/dunning93accurate.html*.

[7] IBM. Unstructured Information Management Architecture (UIMA), 2006. *http://www.alphaworks.ibm.com/tech/uima*.

[8] Lin, C. Y. and Hovy, E. *Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics*. Presented at Human Technology Conference 2003 (HLT-NAACL-2003), Edmonton, Canada, May 27, 2003.

[9] Murthy, U., Richardson, R., Fox, E. A., and Delcambre, L. *Enhancing Concept Mapping Tools Below and Above to Facilitate the Use of Superimposed Information*. Presented at Concept Mapping Conference 2006, San Jose, Costa Rica, 2006.

[10] NDLTD. "Networked Digital Library of Theses and Dissertations", 2005. *http://www.ndltd.org*.

[11] Richardson, R., Fox, E. A., and Woods, J. *Evaluating Concept Maps As A Cross-Language Knowledge Discovery Tool for NDLTD*. Presented at Proceedings of Electronic Theses and Dissertation Conference, Sydney, 2005. *http://adt.caul.edu.au/etd2005/papers/061Richardson.pdf*.

[12] Ross, M., Pinto, H., Pennachin, C., Goertzel, B., Looks, M., Senna, A., and Silva, W. *INLINK: An interactive knowledge-entry and querying tool*. Presented at Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL) 2006, New York, 2006.

[13] Wierzbicka, A. *Semantics, Primes and Universals*. Oxford, UK, Oxford University Press, 2006.