Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the 2nd Workshop on
Cognitive Aspects of the Lexicon

Workshop chairs:
Michael Zock and Reinhard Rapp

22 August 2010
Beijing International Convention Center
Beijing, China

Produced by
Chinese Information Processing Society of China
No.4 Zhong Guan Cun Nan Si Jie, Hai Dian District
Beijing, 100084
China

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Introduction

Whenever we read a book, write a letter or launch a query on a search engine, we always use words, the shorthand labels and concrete forms of abstract notions (concepts, ideas and more or less well specified thoughts). Yet, words are not only vehicles to express thoughts, they are also means to conceive them. They are mediators between language and thought, allowing us to move quickly from one idea to another, refining, expanding or illustrating our possibly underspecified thoughts. Only words have these unique capabilities, which is why they are so important.

Obviously, a good dictionary should contain many entries and a lot of information associated with each one of them. Yet, the quality of a dictionary depends not only on coverage, but also on accessibility of information. Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the moment of consultation (words, concepts, speech sounds). Unlike readers who look for meanings, writers start from them, searching for the corresponding words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related words) and via diverse access routes. Navigation takes place in a huge conceptual lexical space, and the results are displayable in a multitude of forms (e.g. as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

Many lexicographers work nowadays with huge digital corpora, using language technology to build and to maintain the lexicon. But access to the potential wealth of information in dictionaries remains limited for the common user. Yet, the new possibilities of electronic media in terms of comfort, speed and flexibility (multiple inputs, polyform outputs) are enormous. Computational resources are not prone to the same limitations as paperbound dictionaries. The latter were limited in scope, being confined to a specific task (translation, synonyms, ...) for economical reasons, but this limitation is not justified anymore.

Today, by exploiting the advantages of the digital form, we can perform all tasks via one single resource, which may comprise a dictionary, a thesaurus and even more. The goal of this second CogALex workshop, which follows the first edition at COLING 2008 in Manchester, is to perform the groundwork for the next generation of electronic dictionaries, that is, to study the possibility of integrating the different resources, as well as to explore the feasibility of taking the users needs, knowledge and access strategies into account. To reach this goal, we have invited researchers from fields such as computational lexicography, psycholinguistics, cognitive psychology, language learning and ergonomics to address one or several of the following topics:

1. *Conceptual input* of a dictionary user. What is in the authors' minds when they are generating a message and looking for a word? Do they start from partial definitions, i.e. underspecified input (bag of words), conceptual primitives, semantically related words, something akin to synsets, or something completely different? What does it take to bridge the gap between this input, incomplete as it may be, and the desired output (target word)?
2. *Organizing the lexicon and indexing words*. Concepts, words and multi-word expressions can be organized and indexed in many ways, depending on the task and language type. For example, in Indo-European languages words are traditionally organized in alphabetical order, whereas in

Chinese they are organized by semantic radicals and stroke counts. The way words and multi-word expressions are stored and organized affects indexing and access. Since knowledge states (i.e. knowledge available when initiating search) vary greatly and in unpredictable ways, indexing must allow for multiple ways of navigation and access. Hence the question: what organizational principles allow the greatest flexibility for access?

3. *Access, navigation and search strategies* based on various entry types (modalities) and knowledge states. Words are composed of meanings, forms and sounds. Hence, access should be possible via any of these components: via meanings (bag of words), via forms, simple or compound ('hot, dog' vs. 'hot-dog'), and via sounds (syllables). Access should be possible even if input is given in an incomplete, imprecise or degraded form. Furthermore, to allow for natural and efficient access, we need to take the users' knowledge into account (search space reduction) and provide adequate navigational tools, metaphorically speaking, a map and a compass. How do existing tools address these needs, and what could be done to go further?

4. *NLP applications*: Contributors can also demonstrate how such enhanced dictionaries, once embedded in existing NLP applications, can boost performance and help to solve lexical and textual-entailment problems, such as those evaluated in SEMEVAL 2007, or, more generally, generation problems encountered in the context of summarization, question-answering, interactive paraphrasing or translation.

Quite a few of these issues are dealt with in the papers we received. The accepted papers present a rich selection of ideas on the crossroads of semantics, cognition, lexicography, and language learning, thereby emphasizing the interdisciplinary character of the workshop. These are the topics: generating semantic networks, encoding commonsense knowledge in WordNet, textual entailment, sentiment analysis, corpus-based extraction of conceptual classes, parsing of thesauri, term extraction, determining noun classifiers, requirements when using the dictionary of an authoring tool, and the problem of word access.

In sum, there is an active community of researchers working on cognitive aspects of the lexicon, and there is a real awareness concerning the importance of the problems presented in our call for papers.

We would like to thank all the people who in one way or another have helped us to make this workshop a success. Our special thanks go to Eduard Hovy for having accepted to give the invited presentation, and to the members of the program committee who did an excellent job in reviewing the submitted papers. Their reviews were important not only to assure a good selection of papers, but also for the authors, helping them to improve their work. We would also like to express our gratitude to the COLING organizers, in particular to the general workshop chairs and the publication chairs. Last but not least, we would like to thank our authors for their papers and presentations and the participants of the workshop for their interest and their contributions to the discussions.

Michael Zock and Reinhard Rapp

Organizers:

Michael Zock, LIF-CNRS, Marseille (France)
Reinhard Rapp, University of Tarragona (Spain)

Invited Speaker:

Eduard Hovy, Information Sciences Institute, University of Southern California (USA)

Program Committee:

Slaven Bilac, Google Tokyo (Japan)
Pierrette Bouillon, ISSCO, Geneva (Switzerland)
Dan Cristea, University of Iasi (Romania)
Katrin Erk, University of Texas (USA)
Olivier Ferret, CEA LIST (France)
Thierry Fontenelle, EU Translation Centre (Luxembourg)
Sylviane Granger Université Catholique de Louvain (Belgium)
Gregory Grefenstette, Exalead, Paris (France)
Ulrich Heid, IMS, University of Stuttgart (Germany)
Erhard Hinrichs, University of Tübingen (Germany)
Graeme Hirst, University of Toronto (Canada)
Eduard Hovy, ISI, University of Southern California, Los Angeles (USA)
Chu-Ren Huang, Hong Kong Polytechnic University (China)
Terry Joyce, Tama University, Kanagawa-ken (Japan)
Philippe Langlais, DIRO/RALI University of Montreal (Canada)
Marie-Claude L'Homme, University of Montreal (Canada)
Verginica Mititelu, RACAI, Bucharest (Romania)
Alain Polguère, ATILF - CNRS / Université Nancy 2 (France)
Reinhard Rapp, University of Tarragona (Spain)
Sabine Schulte im Walde, University of Stuttgart (Germany)
Gilles Sérasset, IMAG, Grenoble (France)
Serge Sharoff, University of Leeds (UK)
Anna Sinopalnikova, FIT, BUT, Brno (Czech Republic)
Carole Tiberius, Institute for Dutch Lexicology (The Netherlands)
Takenobu Tokunaga, TITECH, Tokyo (Japan)
Dan Tufis, RACAI, Bucharest (Romania)
Piek Vossen, Vrije Universiteit Amsterdam (The Netherlands)
Yorick Wilks, Oxford Internet Institute (UK)
Michael Zock, LIF-CNRS, Marseille (France)
Pierre Zweigenbaum, LIMSI-CNRS, Orsay (France)

Table of Contents

Conference Program

Sunday, August 22, 2010

9:00–9:15 Opening Remarks

Invited Keynote Presentation

9:15–10:30 *Distributional Semantics and the Lexicon*
Eduard Hovy

10:30–11:00 Coffee break

Session 1: Semantics and Cognition

11:00–11:30 *SemanticNet-Perception of Human Pragmatics*
Amitava Das and Sivaji Bandyopadhyay

11:30–12:00 *Exploiting Lexical Resources for Therapeutic Purposes: the Case of WordNet and STaRS.sys*
Gianluca E. Lebani and Emanuele Pianta

12:00–12:30 *Textual Entailmaint Recognition using Word Overlap, Mutual Information and Sub-path Set*
Yuki Muramatsu, Kunihiro Uduka and Kazuhide Yamamoto

12:30–13:00 *The Color of Emotions in Texts*
Carlo Strapparava and Gozde Ozbal

13:00–14:00 Lunch break

Sunday, August 22, 2010 (continued)

Session 2: Lexicography

14:00–14:30 *How to Expand Dictionaries by Web-Mining Techniques*
Nicolas Béchet and Mathieu Roche

14:30–15:00 *An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries*
Neculai Curteanu, Alex Moruz and Diana Trandabat

15:00–15:30 *Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: a Case Study for Italian*
Elisa Lavagnino and Jungyeul Park

15:30–16:00 Coffee break

Session 3: Word Access and Language Learning

16:00–16:30 *Computational Lexicography: A Feature-based Approach in Designing an E-dictionary of Chinese Classifiers*
Helena Gao

16:30–17:00 *In Search of the 'Right' Word*
Stella Markantonatou, Aggeliki Fotopoulou, Maria Alexopoulou and Marianna Mini

Keynote Presentation

17:00–17:45 *Lexical Access, a Search-Problem*
Michael Zock, Didier Schwab and Nirina Rakotonanahary

17:45–18:00 Wrap Up Discussion

18:00 End of the Workshop

INVITED KEYNOTE PRESENTATION

Distributional Semantics and the Lexicon

Eduard Hovy
Information Sciences Institute
University of Southern California
hovy@isi.edu

The lexicons used in computational linguistics systems contain morphological, syntactic, and occasionally also some semantic information (such as definitions, pointers to an ontology, verb frame filler preferences, etc.). But the human cognitive lexicon contains a great deal more, crucially, expectations about how a word tends to combine with others: not just general information-extraction-like patterns, but specific instantial expectations. Such information is very useful when it comes to listening in bad aural conditions and reading texts in which background information is taken for granted; without such specific expectation, one would be hard-pressed (and computers are completely unable) to form coherent and richly connected multi-sentence interpretations.

Over the past few years, NLP work has increasingly treated *topic signature word distributions* (also called 'context vectors', 'topic models', etc.) as a de facto replacement for semantics. Whether the task is wordsense disambiguation, certain forms of textual entailment, information extraction, paraphrase learning, and so on, it turns out to be very useful to consider a word(sense) as being defined by the distribution of word(senses) that regularly accompany it (in the classic words of Firth, "you shall know a word by the company it keeps"). And this is true not only for individual wordsenses, but also for larger units such as *topics*: the product of LDA and similar topic characterization engines is similar.

In this talk I argue for a new kind of semantics, which is being called Distributional Semantics. It combines traditional symbolic logic-based semantics with (computation-based) statistical word distribution information. The core resource is a single lexico-semantic lexicon that can be used for a variety of tasks, provided that it is reformulated accordingly. I show how to define such a semantics, how to build the appropriate lexicon, how to format it, and how to use it for various tasks. The talk pulls together a wide range of related topics, including Pantel-style resources like DIRT, inferences / expectations such as those used in Schank-style expectation-based parsing and expectation-driven NLU, PropBank-style word valence lexical items, and the treatment of negation and modalities. I conclude by arguing that the human cognitive lexicon has to have the same kinds of properties as the Distributional Semantics lexicon, given the ways people *do things with words*.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), page 1,
Beijing, August 2010

SemanticNet-Perception of Human Pragmatics

Amitava Das[1] and Sivaji Bandyopadhyay[2]
Department of Computer Science and Engineering
Jadavpur University
amitava.santu@gmail.com[1] sivaji_cse_ju@yahoo.com[2]

Abstract

SemanticNet is a semantic network of lexicons to hold human pragmatic knowledge. So far Natural Language Processing (NLP) research patronized much of manually augmented lexicon resources such as WordNet. But the small set of semantic relations like Hypernym, Holonym, Meronym and Synonym etc are very narrow to capture the wide variations human cognitive knowledge. But no such information could be retrieved from available lexicon resources. SemanticNet is the attempt to capture wide range of context dependent semantic inference among various themes which human beings perceive in their pragmatic knowledge, learned by day to day cognitive interactions with the surrounding physical world. SemanticNet holds human pragmatics with twenty well established semantic relations for every pair of lexemes. As every pair of relations cannot be defined by fixed number of certain semantic relation labels thus additionally contextual semantic affinity inference in SemanticNet could be calculated by network distance and represented as a probabilistic score. SemanticNet is being presently developed for Bengali language.

1 Historical Motivation

Semantics (from Greek "*σημαντικός*" - *semantikos*) is the study of meaning, usually in language. The word "semantics" itself denotes a range of ideas, from the popular to the highly technical. It is often used in ordinary language to denote a problem of understanding that comes down to word selection or connotation. We studied with various Psycholinguistics experiments to understand how human natural intelligence helps to understand general semantic from nature. Our study was to understand the human psychology about semantics beyond language. We were haunting for the intellectual structure of the psychological and neurobiological factors that enable humans to acquire, use, comprehend and produce natural languages. Let's come with an example of simple conversation about movie between two persons.

Person A: Have you seen the movie *'No Man's Land'?* How is it?
Person B: Although it is good but you should see *'The Hurt Locker'?*

May be the conversation looks very casual, but our intension was to find out the direction of the decision logic on the Person B's brain. We start digging to find out the nature of human intelligent thinking. A prolonged discussion with Person B reveals that the decision logic path to recommend a good movie was as the Figure 1. The highlighted red paths are the shortest semantic affinity distances of the human brain.

We call it semantic thinking. Although the derivational path of semantic thinking is not such easy as we portrait in Figure 1 but we keep it easier for understandability. Actually a human try to figure out the closest semantic affinity node into his pragmatics knowledge by natural intelligence. In the previous example Person B find out with his intelligence that *No Man's Land* is a war movie and got Oscar

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 2–11,
Beijing, August 2010

award. Oscar award generally cracked by Hollywood movies and thus Person B start searching his pragmatics network to find out a movie fall into war genre, from Hollywood and may be got Oscar award. Person B finds out the name of a movie *The Hurt Locker* at nearer distance into his pragmatics knowledge network which is an optimized recommendation that satisfy all the criteria. Noticeably Person B didn't choice the other paths like Bollywood, Foreign movie etc.

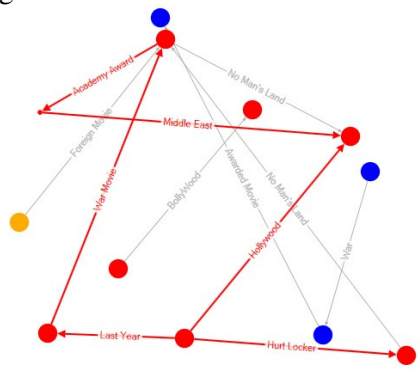


Figure 1: Semantic Thinking

And thus our aim was to develop a computational lexicon structure for semantics as human pragmatics knowledge. We spare long time to find out the most robust structure to represent pragmatics knowledge properly and it should be easy understandable for next level of search and usability.

We look into literature that probably direct to the direction of our ideological thinking. We found that in the year of 1996 Push Singh and Marvin Minsky proposed the field has shattered into subfields populated by researchers with different goals and who speak very different technical languages. Much has been learned, and it is time to start integrating what we've learned, but few researchers are widely versed enough to do so. They had a proposal for how to do so in their ConceptNet work. They developed lexicon resources like ConceptNet (Liu and Singh, 2004). ConceptNet-ConceptNet is a large-scale semantic network (over 1.6 million links) relating a wide variety of ordinary objects, events, places, actions, and goals by only 20 different link types, mined from corpus.

The present task of developing SemanticNet is to capture semantic affinity knowledge of human pragmatics as a lexicon database. We extend our vision from the human common sense (as in ConceptNet) to human pragmatics and have proposed semantic relations for every pair of lexemes that cannot be defined by fixed number of certain semantic relation labels. Contextual semantic affinity inference in SemanticNet could be calculated by network distance and represented as a probabilistic score. SemanticNet is being presently developed for Bengali language.

2 Semantic Roles

The ideological study of semantic roles started age old ago since Panini's *karaka* theory that assigns generic semantic roles to words in a natural language sentence. Semantic roles are generally domain specific in nature such as FROM_DESTINATION,TO_DESTINATION, DEPARTURE_TIME etc. Verb-specific semantic roles have also been defined such as EATER and EATEN for the verb eat. The standard datasets that are used in various English SRL systems are: PropBank (Palmer et al., 2005), FrameNet (Fillmore et al., 2003) and VerbNet (Kipper et al., 2006). These collections contain manually developed well-trusted gold reference annotations of both syntactic and predicate-argument structures.

PropBank defines semantic roles for each verb. The various semantic roles identified (Dowty, 1991) are Agent, patient or theme etc. In addition to verb-specific roles, PropBank defines several more general roles that can apply to any verb (Palmer et al., 2005).

FrameNet is annotated with verb frame semantics and supported by corpus evidence. The frame-to-frame relations defined in FrameNet are Inheritance, Perspective_on, Subframe, Precedes, Inchoative_of, Causative_of and Using. Frame development focuses on paraphrasability (or near paraphrasability) of words and multi-words.

VerbNet annotated with thematic roles refer to the underlying semantic relationship between a predicate and its arguments. The semantic tagset of VerbNet consists of tags as agent, patient, theme, experiencer, stimulus, instrument, location, source, goal, recipient, benefactive etc.

It is evident from the above discussions that no adequate semantic role set exists that can be defines across various domains. Hence pro-

posed SemanticNet does not only rely on fixed type of semantics roles as ConceptNet. For semantic relations we followed the 20 relations defined in ConceptNet. Additionally we proposed semantic relations for every pair of lexicons cannot be defined by exact semantic role and thus we formulated a probabilistic score based technique. Semantic affinity in SemanticNet could be calculated by network distance. Details could be found in relevant Section 8.

3 Corpus

Present SemanticNet has been developed for Bengali language. Resource acquisition is one of the most challenging obstacles to work with electronically resource constrained languages like Bengali. Although Bengali is the sixth[1] popular language in the World, second in India and the national language in Bangladesh.

There was another issue drive us long way to find out the proper corpus for the development of SemanticNet. As the notion is to capture and store human pragmatic knowledge so the hypothesis was chosen corpus should not be biased towards any specific domain knowledge as human pragmatic knowledge is not constricted to any domain rather it has a wide spread range over anything related to universe and life on earth. Additionally it must be larger in size to cover mostly available general concepts related to any topic. After a detail analysis we decided it is better to choose NEWS corpus as various domains knowledge like Politics, Sports, Entertainment, Social Issues, Science, Arts and Culture, Tourism, Advertisement, TV schedule, Tender, Comics and Weather etc are could be found only in NEWS corpus.

Statistics	NEWS
Total no. of news documents in the corpus	108,305
Total no. of sentences in the corpus	2,822,737
Avg no. of sentences in a document	27
Total no. of wordforms in the corpus	33,836,736
Avg. no. of wordforms in a document	313
Total no. of distinct wordforms in the corpus	467,858

Table 1: Bengali Corpus Statistics

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

Fortunately such corpus development could be found in (Ekbal and Bandyopadhyay, 2008) for Bengali. We obtained the corpus from the authors. The Bengali NEWS corpus consisted of consecutive 4 years of NEWS stories with various sub domains as reported above. For the present task we have used the Bengali NEWS corpus, developed from the archive of a leading Bengali NEWS paper[2] available on the Web. The NEWS corpus is quite larger in size as reported in Table 1.

4 Annotation

From the collected document set 200 documents have been chosen randomly for the annotation task. Three annotators (Mr. X, Mr. Y and Mr. Z) participated in the present task. Annotators were asked to annotate the theme words (topical expressions) which best describe the topical snapshot of the document.

The agreement of annotations among three annotators has been evaluated. The agreements of tag values at theme words level is reported in Table 2.

Annotators	X vs. Y	X Vs. Z	Y Vs. Z	Avg
Percentage	82.64%	71.78%	80.47%	78.3%
All Agree	75.45%			

Table 2: Agreement of annotators at theme words level

5 Theme Identification

Term Frequency (TF) plays a crucial role to identify document relevance in Topic-Based Information Retrieval. The motivation behind developing Theme detection technique is that in many documents relevant words may not occur frequently or irrelevant words may occur frequently. Moreover for the lexicon affinity inference, topic or theme words are the only strong clue to start with. The Theme detection technique has been proposed to resolve these issues to identify discourse level most relevant thematic nodes in terms of word or lexicon using a standard machine learning technique. The machine learning technique used here is Conditional Random Field (CRF)[3]. The theme word detection has been defined as a sequence

[2] http://www.anandabazar.com/

[3] http://crfpp.sourceforge.net

labeling problem using various useful depending features. Depending upon the series of input features, each word is tagged as either Theme Word (TW) or Other (O).

5.1 Feature Organization

The set of features used in the present task have been categorized as Lexico-Syntactic, Syntactic and Discourse level features. These are listed in the Table 3 below and have been described in the subsequent subsections.

Types	Features
Lexico-Syntactic	POS
	Frequency
	Stemming
Syntactic	Chunk Label
	Dependency Parsing Depth
Discourse Level	Title of the Document
	First Paragraph
	Term Distribution
	Collocation

Table 3: Features

5.2 Lexico-Syntactic Features

5.2.1 Part of Speech (POS)

It has been shown by Das and Bandyopadhyay, (2009), that theme bearing words in sentences are mainly adjective, adverb, noun and verbs as other POS categories like pronoun, preposition, conjunct, article etc. have no relevance towards thematic semantic of any document. The detail of the POS tagging system chosen for the present task could be found in (Das and Bandyopadhyay 2009).

5.3 Frequency

Frequency always plays a crucial role in identifying the importance of a word in the document or corpus. The system generates four separate high frequent word lists after function words are removed for four POS categories: adjective, adverb, verb and noun. Word frequency values are then effectively used as a crucial feature in the Theme Detection technique.

5.4 Stemming

Several words in a sentence that carry thematic information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in appropriate lists. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer based on stemming cluster technique has been used as described in (Das and Bandyopadhyay, 2010). This stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have the same root form are grouped in a finite number of clusters with the identified root word as cluster center.

5.5 Syntactic Features

5.5.1 Chunk Label

We found that Chunk level information is very much effective to identify lexicon inference affinity. As an example:

(সত্যজিত রায়ের)/NP (**মুক্তি**প্রাপ্ত **ছবি**গুলি)/NP (অনন্য)/NP (সাধারন)/JJP (।)/SYM
The **movies released** by Satyajit Roy are excellent.

In the above example two lexicons "**মুক্তি/release"** and "**ছবি/movie"** are collocated in a chunk and they are very much semantically neighboring in human pragmatic knowledge. Chunk feature effectively used in supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. In the task of identification of Theme expressions, chunk label markers play a crucial role. Further details of development of chunking system could be found in (Das and Bandyopadhyay 2009).

5.5.2 Dependency Parser

Dependency depth feature is very useful to identify Theme expressions. A particular Theme word generally occurs within a particular range of depth in a dependency tree. Theme expressions may be a Named Entity (NE: person, organization or location names), a common noun (Ex: accident, bomb blast, strike etc) or words of other POS categories. It has been observed that depending upon the nature of Theme expressions it can occur within a certain depth in the dependency tree in the sentences. A statistical dependency parser has

been used for Bengali as described in (Ghosh et al., 2009).

5.6 Discourse Level Features

5.6.1 Positional Aspect

Depending upon the position of the thematic clue, every document is divided into a number of zones. The features considered for each document are Title words of the document, the first paragraph words and the words from the last two sentences. A detailed study was done on the Bengali news corpus to identify the roles of the positional aspect features of a document (first paragraph, last two sentences) in the detection of theme words. The importance of these positional features has been described in the following section.

5.6.2 Title Words

It has been observed that the Title words of a document always carry some meaningful thematic information. The title word feature has been used as a binary feature during CRF based machine learning.

5.6.3 First Paragraph Words

People usually give a brief idea of their beliefs and speculations about any related topic or theme in the first paragraph of the document and subsequently elaborate or support their ideas with relevant reasoning or factual information. Hence first paragraph words are informative in the detection of Thematic Expressions.

5.6.4 Words From Last Two Sentences

It is a general practice of writing style that every document concludes with a summary of the overall story expressed in the document. We found that it is very obvious that every document ended with dense theme/topic words in the last two sentences.

5.6.5 Term Distribution Model

An alternative to the classical TF-IDF weighting mechanism of standard IR has been proposed as a model for the distribution of a word. The model characterizes and captures the informativeness of a word by measuring how regularly the word is distributed in a document. Thus the objective is to estimate that measures the distribution pattern of the k occurrences of the word w_i in a document d. Zipf's law describes distribution patterns of words in an entire corpus. In contrast, term distribution models capture regularities of word occurrence in subunits of a corpus (e.g., documents, paragraphs or chapters of a book). A good understanding of the distribution patterns is useful to assess the likelihood of occurrences of a theme word in some specific positions (e.g., first paragraph or last two sentences) of a unit of text. Most term distribution models try to characterize the informativeness of a word identified by inverse document frequency (IDF). In the present work, the distribution pattern of a word within a document formalizes the notion of theme inference informativeness. This is based on the Poisson distribution. Significant Theme words are identified using TF, Positional and Distribution factor. The distribution function for each theme word in a document is evaluated as follows:

$$f_d(w_i) = \sum_{i=1}^{n} (S_i - S_{i-1}) / n + \sum_{i=1}^{n} (TW_i - TW_{i-1}) / n$$

where n=number of sentences in a document with a particular theme word S_i=sentence id of the current sentence containing the theme word and S_{i-1}=sentence id of the previous sentence containing the query term, TW_i is the positional id of current Theme word and TW_{i-1} is the positional id of the previous Theme word.

5.6.6 Collocation

Collocation with other thematic words/expressions is undoubtedly an important clue for identification of theme sequence patterns in a document. As we used chunk level collocation to capture thematic words (as described in 5.5.1) and in this section we are introducing collocation feature as inter-chunk collocation or discourse level collocation with various granularity as sentence level, paragraph level or discourse level.

6 Theme Clustering

Theme clustering algorithms partition a set of documents into finite number of topic based groups or clusters in terms of theme words/expressions. The task of document clustering is to create a reasonable set of clusters

for a given set of documents. A reasonable cluster is defined as the one that maximizes the within-cluster document similarity and minimizes between-cluster similarities. There are two principal motivations for the use of this technique in the theme clustering setting: efficiency, and the **cluster hypothesis**.

The **cluster hypothesis** (Jardine and van Rijsbergen, 1971) takes this argument a step further by asserting that retrieval from a clustered collection will not only be more efficient, but will in fact improve retrieval performance in terms of recall and precision. The basic notion behind this hypothesis is that by separating documents according to topic, relevant documents will be found together in the same cluster, and non-relevant documents will be avoided since they will reside in clusters that are not used for retrieval. Despite the plausibility of this hypothesis, there is only mixed experimental support for it. Results vary considerably based on the clustering algorithm and document collection in use (Willett, 1988). We employ the **clustering** hypothesis only to measure inter-document level thematic affinity inference on semantics.

Application of the clustering technique to the three sample documents results in the following theme-by-document matrix, A, where the rows represent various documents and the columns represent the themes politics, sport, and travel.

$$A = \begin{bmatrix} election & cricket & hotel \\ parliament & sachin & vacation \\ governor & soccer & tourist \end{bmatrix}$$

The similarity between vectors is calculated by assigning numerical weights to these words and then using the cosine similarity measure as specified in the following equation.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \vec{q}_k . \vec{d}_j = \sum_{i=1}^{N} w_{i,k} \times w_{i,j} \text{ ---- (1)}$$

This equation specifies what is known as the dot product between vectors. Now, in general, the dot product between two vectors is not particularly useful as a similarity metric, since it is too sensitive to the absolute magnitudes of the various dimensions. However, the dot product between vectors that have been length normalized has a useful and intuitive interpretation: it computes the **cosine** of the angle between the two vectors. When two documents are identical they will receive a cosine of one; when they are orthogonal (share no common terms) they will receive a cosine of zero. Note that if for some reason the vectors are not stored in a normalized form, then the normalization can be incorporated directly into the similarity measure as follows.

Of course, in situations where the document collection is relatively static, it makes sense to normalize the document vectors once and store them, rather than include the normalization in the similarity metric.

$$s\left(\vec{q}_k, \vec{d}_j\right) = \frac{\sum_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^{N} w_{i,k}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}} \text{ ----(2)}$$

Calculating the similarity measure and using a predefined threshold value, documents are classified using standard bottom-up soft clustering *k-means* technique. The predefined threshold value is experimentally set as 0.5 as shown in Table 4.

ID	Theme	1	2	3
1	**প্রশাসন** (administration)	**0.63**	0.12	0.04
1	সুশাসন (good-government)	**0.58**	0.11	0.06
1	সমাজ (society)	**0.58**	0.12	0.03
1	আইন (law)	**0.55**	0.14	0.08
2	গবেষণা (research)	0.11	**0.59**	0.02
2	কলেজ (college)	0.15	**0.55**	0.01
2	**উচ্চশিক্ষা** (higher study)	0.12	**0.66**	0.01
3	জেহাদি (jehadi)	0.13	0.05	**0.58**
3	মসজিদ (mosque)	0.05	0.01	**0.86**
3	নয়াদিল্লী (New Delhi)	0.12	0.04	**0.65**
3	**কাশ্মীর** (Kashmir)	0.03	0.01	**0.93**

Table 4: Five cluster centroids (mean $\vec{\mu}_j$).

A set of initial cluster centers is necessary in the beginning. Each document is assigned to the cluster whose center is closest to the document. After all documents have been assigned, the center of each cluster is recomputed as the centroid or mean $\vec{\mu}$ (where $\vec{\mu}$ is

the clustering coefficient) of its members that is $\vec{\mu} = \left(1/\left|c_j\right|\right)\sum_{x \in c_j} \vec{x}$. The distance function is the **cosine vector** similarity function.

Table 4 gives an example of theme centroids by the *K-means* clustering. Bold words in Theme column are cluster centers. Cluster centers are assigned by maximum clustering coefficient. For each theme word, the cluster from Table 4 is still the dominating cluster. For example, **"প্রশাসন"** has a higher membership probability in cluster1 than in other clusters. But each theme word also has some non-zero membership in all other clusters. This is useful for assessing the strength of association between a theme word and a topic. Comparing two members of the cluster2, **"কাশ্মীর"** and **"নয়াদিল্লী",** it is seen that **"নয়াদিল্লী"** is strongly associated with cluster2 (p=0.65) but it has some affinity with other clusters as well (e.g., p =0.12 with the cluster1). This is a good example of the utility of soft clustering. These non-zero values are still useful for calculating vertex weight during Semantic Relational Graph generation.

7 Semantic Relational Graph

Representation of input text document(s) in the form of graph is the key to our design principle. The idea is to build a document graph $G=<V,E>$ from a given source document $d \in D$. At this preprocessing stage, text is tokenized, stop words are eliminated, and words are stemmed. Thus, the text in each document is split into fragments and each fragment is represented with a vector of constituent theme words. These text fragments become the nodes V in the document graph.

The similarity between two nodes is expressed as the weight of each edge *E* of the document graph. A weighted edge is added to the document graph between two nodes if they either correspond to adjacent text fragments in the text or are semantically related by theme words. The weight of an edge denotes the degree of the semantic inference relationship. The weighted edges not only denote document level similarity between nodes but also inter document level similarity between nodes. Thus to build a document graph *G*, only the edges with edge weight greater than some predefined threshold value are added to *G*, which basically constitute edges *E* of the graph *G*.

The Cosine similarity measure has been used here. In cosine similarity, each document *d* is denoted by the vector $\vec{V}(d)$ derived from *d*, with each component in the vector for each Theme words. The cosine similarity between two documents (nodes) *d1* and *d2* is computed using their vector representations $\vec{V}(d1)$ and $\vec{V}(d2)$ as equation (1) and (2) (Described in Section 6). Only a slight change has been done i.e. the dot product of two vectors $\vec{V}(d1) \bullet \vec{V}(d2)$ is defined as $\sum_{i=1}^{M} V(d1)V(d2)$. The Euclidean length of *d* is defined to be $\sqrt{\sum_{i=1}^{M} \vec{V}_i^{\,2}(d)}$ where *M* is the total number of documents in the corpus. Theme nodes within a cluster are connected by vertex, weight is calculated by clustering co-efficient of those theme nodes. Additionally inter cluster vertexes are there. Cluster centers are interconnected with weighted vertex. The weight is calculated by cluster distance as measured by cosine similarity measure as discussed earlier.

To better aid our understanding of the automatically determined category relationships we visualized this network using the Fruchterman-Reingold force directed graph layout algorithm (Fruchterman and Reingold, 1991) and the NodeXL network analysis tool (Smith et al., 2009)[4]. A theme relational model graph drawn by NoddeXL is shown in Figure 1.

8 Semantic Distance Measurement

Finally generated semantic relational graph is the desired SemanticNet that we proposed. Generated Bengali SemanticNet consist of almost 90K high frequent Bengali lexicons. Only four categories of POS (noun, adjective, adverb and verb) considered for present generation as reported in Section 5.2.1. In the generated Bengali SemanticNet all the lexicons are connected with weighted vertex either directly

[4] Available from http://www.codeplex.com/NodeXL

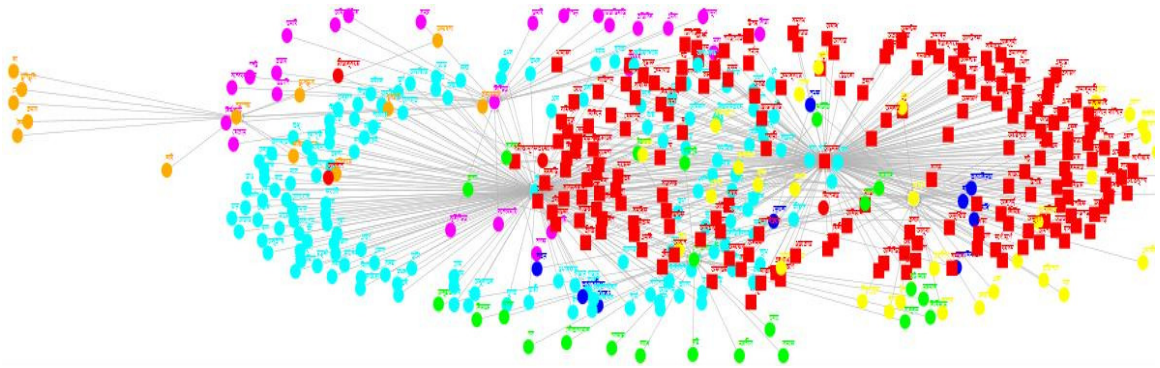


Figure 1: Semantic Relational Graph by NodeXL

or indirectly. Semantic lexicon inference could be identified by network distance of any two nodes by calculating the distance in terms of weighted vertex. We computed the relevance of semantic lexicon nodes by summing up the edge scores of those edges connecting the node with other nodes in the same cluster. As cluster centers are also interconnected with weighted vertex so inter-cluster relations could be also calculated in terms of weighted network distance between two nodes within two separate clusters. As an example:

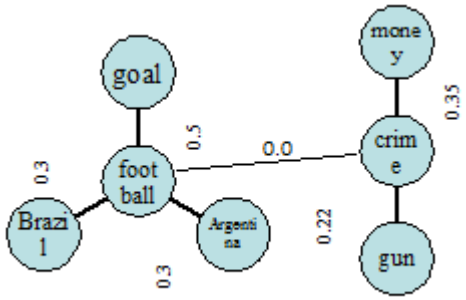


Figure 2: Semantic Affinity Graph

The lexicon semantic affinity inference from Figure 2 could be calculated as follows:

$$S_d(w_i, w_j) = \frac{\sum_{k=0}^{n} v_k}{k} \quad \text{----(1) or}$$

$$= \sum_{c=0}^{m} \frac{\sum_{k=0}^{n} v_k}{k} \times \prod_{c=0}^{m} l_c \text{---(2)}$$

where $S_d(w_i, w_j) =$ semantic affinity distance between two lexicons w_i and w_j. Equation (1) and (2) are for intra-cluster and inter-cluster semantic distance measure respectively. *k*=number of weighted vertex between two lexicons w_i and w_j. v_k is the weighted vertex between two lexicons. *m*=number of cluster centers between two lexicons. l_c is the distance between cluster centers between two lexicons.

For illustration of present technique let take an example:

$$\text{(Argentina, goal)} = \frac{0.5 + 0.3}{2} = 0.4$$

$$\text{(Gun, goal)} = \left(\frac{0.22}{1} + \frac{0.5}{1} \right) \times 0.0 = 0$$

It is evident from the previous example that the score based semantic distance can better illustrate lexicon affinity between *Argentina* and *goal* but is no lexicon affinity relation between *gun* and *goal*.

Instead of giving only certain semantic relations like WordNet or ConceptNet the present relative probabilistic score based lexicon affinity distance based technique can represent best acceptable solution for representing the human pragmatic knowledge. Not only ideologically rather the SemanticNet provide a good solution to any type of NLP problem. A detail analysis of Information retrieval system using SemanticNet is detailed in evaluation section.

Although every lexicon pair cannot be labeled by exact semantic role but we try to keep a few semantic roles to establish a crossroad from previous computational lexicon techniques to this new one. These semantic relations may be treated as a bridge to traverse SemanticNet by gathering knowledge from other resources WordNet and ConceptNet. Approximately 22k (24% of overall SemanticNet) lexicons are tagged with appropriate semantic roles by two processes as described below.

9 Semantic Role Assignment

Two types of methods have been taken to assign pair wise lexicon semantic affinity relations. First one is derived from ConceptNet. In the second technique sub-graph is identified consisting of a nearest *verb* and roles are assigned accordingly.

9.1 Semantic Roles from ConceptNet

A ConceptNet API[5] written in Java has been used to extract pair wise relation from ConceptNet. A Bengali-English dictionary (approximately 102119 entries) has been developed using the Samsad Bengali-English dictionary[6] used here for equivalent lookup of English meaning of each Bengali lexicon. Obtained semantic relations from ConceptNet for any lexicon English pair are assigned to source Bengali pair lexicons. As an example:

("Tree","Gree") ("গাছ","সবুজ")

- ❖ OftenNear
- ❖ PartOf
- ❖ PropertyOf
- ❖ IsA

9.2 Verb Sub-Graph Identification

It is an automatic process using manually augmented list of only 220 Bengali verbs. This process starts from any arbitrary node of any cluster and start finding any nearest verb within the cluster. The system uses the manually augmented list of verbs as partly reported in Table 5.

Verb	English Gloss	Probable Relations
হয়	Be	IsA
আছে	Have	CapableOf
থাকা	Have	CapableOf
তৈরি	Made	MadeOf
বসবাস	Live	LocationOf

Table 5: Semantic Relations

The semantic relation labels attached with every verb in the manually augmented list (as reported in Table 5) is then automatically assigned between each pair of lexicons.

[5] http://web.media.mit.edu/~hugo/conceptnet/

[6] http://dsal.uchicago.edu/dictionaries/biswas_bengali/

10 Evaluation

It is bit difficult to evaluate this type of lexicon resources automatically. Manual validation may be suggested as a better alternative but we prefer for a practical implementation based evaluation strategy.

For evaluation of Bengali SemanticNet it is used in Information Retrieval task using corpus from Forum for Information Retrieval Evaluation (FIRE)[7] ad-hoc mono-lingual information retrieval task for Bengali language. Two different strategies have been taken. First a standard IR technique with TF-IDF, zonal indexing and ranking based technique (Bandyopadhyay et al., 2008) has been taken. Second technique uses more or less same strategy along with query expansion technique using SemanticNet (Although the term SemanticNet was not mentioned there) as a resource (Bhaskar et al., 2010).

Only the following evaluation metrics have been listed for each run: mean average precision (MAP), Geometric Mean Average Precision (GM-AP), (document retrieved relevant for the topic) R-Precision (R-Prec), Binary preferences (Bpref) and Reciprical rank of top relevant document (Recip_Rank). The evaluation strategy follows the global standard as Text Retrieval Conference (TREC)[8] metrics. It is clearly evident from the system results as reported in Table 6 that SemanticNet is a better way to solve lexicon semantic affinity.

Scores	Bengali IR using	
	IR	SemanticNet
MAP	0.0200	0.4002
GM_AP	0.0004	0.3185
R-Prec	0.0415	0.3894
Bpref	0.0583	0.3424
Recip_Rank	0.4432	0.6912

Table 6: Information Retrieval using SemanticNet

Evaluation result shows effectiveness of developed SemanticNet in IR. Further analysis

[7] http://www.isical.ac.in/~clia/index.html

[8] http://trec.nist.gov/

revealed that general query expansion technique generally used WordNet synonyms as a resource. But in reality "হৃদয়" and "পরাণ" could not be clustered in one cluster though they represent same semantic of 'heart'. First one used in general context whereas the second one used only in literature. If there is any problem to understand Bengali let come with an example of English. Conceptually "you" and "thy" could be mapped in same cluster as they both represent the semantic of 2nd person but in reality "thy" simply refers to the literature of the great English poet Shakespeare. Standard lexicons cannot discriminate this type of fine-grained semantic differences.

11 Conclusion and Future Task

Experimental result of Information Retrieval using SemanticNet proves it is a better solution rather than any existing lexicon resources. The development strategy employs less human interruption rather a general architecture of Theme identification or Theme Clustering technique using easily extractable linguistics knowledge. The proposed technique could be replicated for any new language.

SemanticNet could be useful any kind of Information Retrieval technique, Information Extraction technique, and topic based Summarization and we hope for newly identified NLP sub disciplines such as Stylometry or Authorship detection and plagiarism detection etc.

Our future task will be in the direction of different experiments of NLP as mentioned above to profoundly establish the efficiency of SemanticNet. Furthermore we will try to develop SemanticNet for many other languages.

References

Bandhyopadhyay S., Das A., Bhaskar P.. English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008. In Working Note of Forum for FIRE-2008.

Bhaskar P., Das A.,Pakray P.and Bandyopadhyay S.(2010). Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010, In FIRE-2010.

Das A. and Bandyopadhyay S. (2009). Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII).

Das A. and Bandyopadhyay S. (2010). Morphological Stemming Cluster Identification for Bangla, In Knowledge Sharing Event-1: Task 3: Morphological Analyzers and Generators, January, 2010, Mysore.

Ekbal A., Bandyopadhyay S (2008). A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal. pages 173-182, 2008

Fillmore Charles J., Johnson Christopher R., and Petruck Miriam R. L.. 2003. Background to FrameNet. International Journal of Lexicography, 16:235–250.

Fruchterman Thomas M. J. and Reingold Edward M.(1991). Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164.

Ghosh A., Das A., Bhaskar P., Bandyopadhyay S.(2009). Dependency Parser for Bengali: the JU System at ICON 2009. In NLP Tool Contest ICON 2009, December 14th-17th, Hyderabad.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. Information Storage and Retrieval, 7, 217-240.

Kipper Karin, Korhonen Anna, Ryant Neville, and Palmer Martha. Extending VerbNet with Novel Verb Classes. LREC 2006.

Liu Hugo and Singh Push (2004). ConceptNet: a practical commonsense reasoning toolkit. BT Technology Journal, 22(4):211-226.

Palmer Martha, Gildea Dan, Kingsbury Paul, The Proposition Bank: A Corpus Annotated with Semantic Roles, Computational Linguistics Journal, 31:1, 2005.

Singh Push and Williams William (2003). LifeNet: a propositional model of ordinary human activity. In the Proc. Of DC-KCAP 2003.

Singh Push, Barry Barbara, and Liu Hugo (2004). Teaching machines about everyday life. BT Technology Journal, 22(4):227-240.

Smith Marc, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues, Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, and Eric Gleave. 2009. Analyzing (social media) networks with NodeXL. In C&T '09: Proc. Fourth International Conference on Communities and Technologies, LNCS. Springer.

Willerr, P. (1988). Recent trends in hierarchic document clustering: A critical review. Information Processing and Management, 24(5), 577-597.

Exploiting Lexical Resources for Therapeutic Purposes: the case of WordNet and STaRS.sys

Gianluca E. Lebani
Center for Mind/Brain Sciences
University of Trento
gianluca.lebani@unitn.it

Emanuele Pianta
HLT Group
Fondazione Bruno Kessler
pianta@fbk.eu

Abstract

In this paper, we present an on-going project aiming at extending the WordNet lexical database by encoding common sense featural knowledge elicited from language speakers. Such extension of WordNet is required in the framework of the STaRS.sys project, which has the goal of building tools for supporting the speech therapist during the preparation of exercises to be submitted to aphasic patients for rehabilitation purposes. We review some preliminary results and illustrate what extensions of the existing WordNet model are needed to accommodate for the encoding of commonsense (featural) knowledge.

1 Introduction

Electronic lexical resources such as WordNet and FrameNet are used for a great variety of natural processing tasks, ranging from query expansion, to word sense disambiguation, text classification, or textual entailment. Some of these resources are also used by human users as on-line dictionaries; see the Princeton WordNet[1] and the MultiWordNet[2] on-line sites. In this paper we describe a novel attempt to exploit the information contained in wordnets to build a tool designed to support the therapy of language disorders. In doing so, we will tackle also an interesting theoretical issue. Is the WordNet conceptual model apt to represent the common sense knowledge associated to concepts, which is partly lost in case of language disorders (aphasia) due to a brain damage? Note that, in cognitively oriented studies of the lexicon such knowledge is often represented in the form of featural descriptions elicited from speakers, such as *<a cat> is lazy*[3], *<camels> are found in deserts*, *<planes> fly* etc.

Anomia is the most pervasive and persistent of aphasia symptoms. It has been described as "a difficulty in finding high information words, both in fluent discourse and when called upon to identify an object of action by name" (Goodglass and Wingfield, 1997:3). The naming difficulties experienced by anomic patients can vary substantially, so that different "anomias" can be characterized as arising from either a mainly lexical or mainly semantic breakdown. Depending on the kind of anomia, therapeutic approaches can vary, so as to employ the more appropriate tasks and stimuli.

Computers can support the rehabilitation of language disorders in many ways: from assisting the administrative management to enhancing common assessment methods, from helping the clinician during the therapeutic session to alleviating the communicative difficulties of a patient by exploiting his unimpaired abilities (Petheram, 2004).

In these pages we introduce STaRS.sys (Semantic Task Rehabilitation Support system), a Computer Assisted Therapy (CAT) tool designed to support the therapist in the preparation of semantic exercises such as odd-one-out, yes/no attribute question answering, property generation and so forth. All these exercises are based on the kinds of information that are carried by featural

[1] http://wordnet.princeton.edu/
[2] http://multiwordnet.fbk.eu/

[3] Concepts and features will be printed in *italics courier new* font. When reporting a concept-feature pair, the concept will be further enclosed by *<angled brackets>*. Feature types and concept categories will be reported in *italics times new roman*.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 12–17,
Beijing, August 2010

descriptions. Such a scenario motivates the need for a lexical semantic resource which is richer and somehow more cognitively-oriented than the existing ones. We will argue that such needs can be satisfied by enhancing the WordNet model (WN: Fellbaum, 1998 ed) as implemented in the Italian MultiWordNet (MWN: Pianta et al, 2002) lexicon. Our project is developed in collaboration with the CIMeC's Center for Neuropsychological Rehabilitation (CeRiN), and focuses on Italian. We leave to the future the evaluation of whether and how our model can be expanded to other languages.

These pages are organized as follows: Sec. 2 shows the possibilities offered by the exploitation of STaRS.sys in a therapeutic context, and the lexical semantics requirements that such use poses. In Sec. 3 and 4 we illustrate specific issues related to the encoding of featural knowledge into the MWN model.

2 STaRS.sys in a therapeutic context

In this section we will illustrate the semantic requirements that the therapeutic use of STaRS.sys poses, and how we foresee the tool will be used in practical therapeutic scenarios.

2.1 Semantic requirements

An essential requirement of the STaRS.sys tool is the capability of managing the major variables that influence the performance of anomic patients in semantic therapeutic tasks (Raymer and Gonzalez-Rothi, 2002; Cree and McRae, 2003). Accordingly, we identified a minimum of five types of information which should be available for every lexical concept:

Conceptual Taxonomy. A fully-specified conceptual taxonomy is an essential requirement for our tool, in the light of the existence of patients affected by language disorders specific to certain semantic categories, such as *tools*, or *living beings* (Capitani et al, 2003).

Featural Descriptions. Featural descriptions are assumed to play a central role in the human semantic memory (Murphy, 2002) and will be represented here as `<concept> feature` couples, e.g. `<dog> has a tail`.

This information can be exploited for selecting sets of concepts which are relevant in a certain therapeutic context, e.g. concepts sharing a feature value ("red objects") or those for which a type of feature is particularly relevant (e.g. "animals with a peculiar fur").

Feature Types Classification. A grouping of FDs into feature types is needed for selectively working on feature types of interest, or for the estimation of semantic measures such as feature distinctiveness, semantic relevance, concept similarity and feature correlation (Cree and McRae, 2003; Sartori and Lombardi, 2004; Vinson et al, 2003). As we will see in the following sections, feature types can be mapped onto WordNet-like relations.

Prototypicality. A concept can be more or less representative of its category. Choosing and working on concepts with different levels of prototypicality can be informative, for both therapeutic and diagnostic purposes.

Word Frequency. Patients' performance can be affected by word frequency. Thereby, a critical skill for our tool is the ability to discriminate between words used with different frequencies.

2.2 Use Case Scenarios

By exploiting a lexical infrastructure encoding such semantic information, STaRS.sys can be used by a therapist for:

- retrieving concepts;
- retrieving information associated to concepts;
- comparing concepts.

These three functionalities can be illustrated by the preparation of three different tasks for a patient affected by, e.g., a semantic deficit selectively affecting animal concepts. Such a kind of patient would show comprehension and production difficulties restricted to concepts belonging to the *animal* category (Capitani et al, 2003). Plausibly, furthermore, his production problems would manifest both as naming failure in controlled conditions (i.e. in tests like the ones reported below) and as a difficulty/inability to retrieve the intended word in spontaneous speech (Semenza, 1999).

In the first scenario, the therapist looks for concepts that match given specifications in order to prepare a feature generation task. As an example, she submits to STaRS.sys a request for concepts of frequent use, referring to animals, associated to highly distinctive color features and having a high mean feature distinctiveness. The system returns concepts such as `zebra`, `tiger`

and *cow*. Finally the patient is asked to generate phrasal descriptions for these concepts.

In a second scenario, STaRS.sys is used to retrieve FDs for a given set of concepts. Right and wrong concept-feature couples are created to build a questionnaire, in which the patient is required to distinguish the right from the wrong pairs. For instance, the therapist submits to STaRS.sys a query for features of the concept *leopard* that are highly relevant and either perceptual or taxonomical, and obtains features such as *is yellow with black spots* and *is a cat*.

Finally, in the third scenario the therapist uses STaRS.sys to find concepts for an odd-one-out task. That is, she looks for triples composed of two similar concepts plus an incoherent one that has to be found by the patient. As an example, starting from the concept *lion*, she looks for animals that typically live in a similar/different natural habitat, and obtains similar concepts such as *leopard* and *cheetah*, and a dissimilar concept such as *wolf*.

3 WN as semantic lexical resource for STaRS.sys

The STaRS.sys application scenario motivates the need for a lexical semantic resources that:

R1: is cognitively motivated;

R2: is based on a fully-specified is-a hierarchy;

R3: is intuitive enough to be used by a therapist;

R4: allows for the encoding of featural properties and their association to concepts;

While designing the STaRS.sys tool, we made the hypothesis that a semantic lexical resource built according to the WN model could meet most of the above requirements.

In the WN model every concept is represented as a synset (set of synonyms) such as {hand, manus, hook, mauler, mitt, paw}. Such semantic units are organized in a network interconnected trough several relations. Examples of semantic relations include the *is-a* relation, e.g.{left_hand, left} *is-a* {hand, …}, and the *meronymy* relation, e.g. {hand, …} *has-part* {finger}.

At a first glance, WN seems to easily meet three of the above criteria. First, WN was initially conceived as a model of the human lexical memory. Second, WordNet implement extensive and systematic noun hierarchies. More specifically, a preliminary analysis of the Italian MWN nominal hierarchy has shown that the semantic categories which are relevant for rehabilitation purposes can be easily mapped onto MWN top level nodes (*tools*, *animals*, *people*). Third, WN is based on a conceptual model which is relatively simple and near to language use (as opposed to more sophisticated logics-based models). We expect that this feature will facilitate the use of STaRS.sys by therapists, which may not have all the formal logics awareness that is needed to use formal ontologies. Furthermore, MWN is manually developed trough an on-line Web application. We expect that such application can be used by therapists using STaRS.sys for the shared and community-based development/maintenance of the lexical resource they need.

A final motivation in favor of the choice of MWN is the fact that this Italian resource is strictly aligned at the semantic level to English and other European languages (e.g. Spanish, Portuguese, Romania, Hebrew). Thus, we can envisage that at least part of the semantic information which is encoded for Italian can be ported to the aligned languages and used for similar purposes.

4 Mapping featural descriptions into MWN

Our hypothesis about the usefulness of the WN model for the needs of STaRS.sys can be fully confirmed only if we find a way to encode in such a model all or most of the knowledge which is contained in feature descriptions elicited from Italian speakers (R4 in previous section). In more general terms we need to answer the following questions. Does MWN already contain all the information that is needed by the STaRS.sys requirements? If we need to extend the existing MWN, can we simply add new synsets and instances of existing relations, or do we need to add new relation types? Is the conceptual model of MWN or of any other WN variant powerful enough to encode all the information contained in feature descriptions?

A first simple approach to representing feature descriptions in MWN is associating feature descriptions to synset glosses. As a consequence, a MWN gloss, which is currently composed of a definition and a list of usage examples, all

crafted by lexicographers, would contain also a list of feature descriptions, elicited from language speakers.

This approach may be useful for some of the foreseen usages of STaRS.sys (e.g. retrieving feature descriptions from concepts), and can also be interesting for a generic use of MWN. However, to fully exploit the knowledge contained in FDs (e.g. for calculating concept similarity) it is necessary to encode that knowledge in a more explicit way; that is we need to map each FD in a *wordnet-like relation* between a *source* and a *target* concept. For instance, a pair such as `<cup> is used for drinking` can be represented as a *is_used_for* relation holding between the source concept {cup} and the target concept {drink}.

Encoding the source concept is relatively easy given that it is usually expressed as an isolated word that is used as stimulus for feature elicitation from subjects, e.g. "scimmia" ("monkey"). The only problematic aspect in this step may be the choice of the right sense which was meant when the word has been proposed to subjects. In some cases this may be not trivial, even if, in principle, stimulus words are supposed to be chosen so as to avoid ambiguities; see for instance the word "cipolla" ("onion"), which in MWN is ambiguous between the vegetable and food sense.

More complex is the encoding of the feature itself which is a free and possibly complex linguistic description (e.g. `likes eating bananas`). To fulfill our goal, we need to map such description in a wordnet-like relation and a target concept. Such goal can be accomplished in two steps.

4.1 Mapping feature types into MWN relations

Given the semantic requirements illustrated in Sec. 2.1, one the first steps in the development of the STaRS.sys tool has been the design of a classification of FDs in feature types; see Lebani and Pianta (2010). In a second moment, we realized that assigning a FD to a feature type is equivalent to assigning it to a wordnet relation, given that it is possible to create one-to-one mappings between features types and relations.

The adopted feature type classification has been designed so as to be (1) reasonably intuitive, (2) robust and (3) cognitively plausible. The cognitive plausibility requirement has been fulfilled by moving from an analysis of similar proposals put forwards in the experimental literature, or exploited in the therapeutic practice. As for the former, we considered research fields as distant as lexicography, theoretical linguistics and cognitive psychology. Examples of compatible proposals currently exploited in the therapeutic practice are the question type of Laiacona et al.'s (1993) semantic questionnaire, a type classification adopted by the therapists of the CIMeC's CeRiN (personal communication) and the Semantic Feature Analysis paradigm (Boyle and Coelho, 1995).

The resulting classification only considers concrete objects and is composed of 25 feature types. All of them (except the *is associated with* relations) belong to one of the following six relations) belong to one of the following six major classes: taxonomic properties, part-of- relations,

Feature Type	Example
has Portion	*<bread> cut into slices*
has Geographical Part	*<Africa> Egitto*
has Size	*<elephant> is big*
has Shape	*<clock> is round*
has Texture	*<eel> is slimy / <biscuit> is crunchy*
has Taste	*<lemon> is bitter*
has Smell	*<rose water> smells of rose*
has Sound	*<lighting> produces a thunder*
has Colour	*<lemon> is yellow*
is Used for	*<cup> is used for drinking*
is Used by	*<cleaver> is used by butchers*
is Used with	*<violin> is played with a bow*
Situation Located	*<jacket> used in occasions*
Space Located	*<camel> in the desert*
Time Located	*<pajamas> used at night*
has Origin	*<milk> comes from cows*
is Involved in	*<bird> eats seeds - is hunted*
has Attribute	*<subway> is fast*
has Affective Property	*<horror movie> is scary*
is Associated with	*<dog> man*

Table 1: STaRS.sys types not having a parallel wordnet semantic relation

perceptual properties, usage properties, locational properties and associated events and attributes.

A first version of this classification has been evaluated by asking 5 naïve Italian speakers to assign the appropriate type label to 300 concept[4]-feature pairs from a non-normalized version of the Kremer et al's (2008) norms. The inter-coder agreement between subjects (Fleiss' Multi-π = 0,73) validated the skeleton of our classification, at the same time suggesting some minor changes that have been applied to the classification proposed here. An evaluation of the improved classification involving therapists has been planned for the (very near) future.

Note that in order to map all of the feature types into wordnet relations we had to create a number of new relations which are not available in existing wordnets. The list of existing MWN relations used to encode STaRS.sys feature types includes five items: *hypernym, has_co-ordinate, has_part, has_member*, *has_substance*. The following table contains the list of the 20 additional relations, along with examples.

4.2 Encoding target concepts in MWN

A second step needed in order to fully represent the semantics of feature descriptions in MWN is the encoding of target concepts.

Target concepts can be expressed by a noun (e.g. `has a <neck>`), an adjective (e.g. `is <big>`) or a verb or a verbal construction (e.g. `is used for <drinking>`, `is used to <cut bread>`). In principle this is not problematic as WN encodes all these lexical categories.

What is problematic instead is the possible complexity of target concepts. Whereas in WN synsets are bound to contain only lexical units (with the few exceptions of the so called artificial nodes), the target of a featural description can be a free combination of words, for instance a noun modified by an adjective (e.g. `has a <long neck>`) , an adjective modified by an adverb (e.g. `is <very big>`) or a verb with an argument (e.g. `is used to <cut bread>`). For giving an idea of the phenomenon, consider that 27,6% of the features that composes the experimental sample in Lebani and Pianta (2010) contain target concepts expressed by free combination of words

The solution we adopted to solve this problem relies on the notion of *phraset* proposed by Bentivogli and Pianta (2003; 2004), that is a data structure used for encoding "sets of synonymous free combination of words (as opposed to lexical units) which are recurrently used to express a concept". In the original proposal, the authors introduced such a data structure to cope with lexical gaps in multingual resources or to encode alternative (linguistically complex) ways of expressing an existing concept. Phrasets can be associated to existing synsets to represent alternative (non lexical) ways of expressing lexicalized concepts, e.g. the Italian translations of "dishcloth":

Synset: {canovaccio, strofinaccio}
Phraset: {strofinaccio_per_i_piatti,
straccio_per_i_piatti}

where "strofinaccio per i piatti" and "straccio per i piatti" and are free combinations of words. In alternative, they can be used to represent lexical gaps, such as the Italian translation equivalent of "breadknife":

Synset: {GAP}
Phraset: {coltello_da_pane,
coltello_per_il_pane}

Phrasets can be annotated by exploiting the *composes/composed-of* lexical relation linking phraset with the synsets corresponding to the concepts that compose it. For instance the expression in the above phraset is linked by a *hypernym* and by a *composed-of* relation with the synset {coltello} (`knife`) and {pane} (`bread`). As far as FDs are concerned, the use of phrasets is compatible with the received view about the compositional nature of the human conceptual knowledge (Murphy, 2002).

Figure 1 shows how phrasets allow for representing the complex FD `<breadknife> is used to cut bread` in the MWN model.

5 Conclusion and future directions

This paper presents the preliminary results of a research aiming at exploiting and extending the WordNet conceptual model as an essential component of a tool for supporting the rehabilitation of patients with language disorders. A crucial

[4] In details, the subjects were submitted with concrete concepts belonging to one of the following categories: *mammals*, *birds*, *fruits*, *vegetables*, *body parts*, *clothing*, *manipulable tools*, *vehicles*, *furniture* and *buildings*.

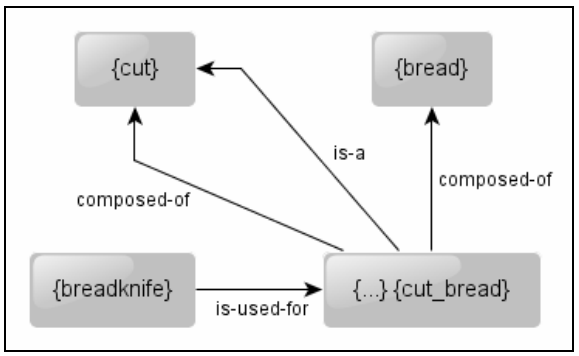


Figure 1: Representation of the concept-feature pair `<breadknife> is used to cut bread`

aspect for the use of wordnet-like resources in such a context is the possibility of representing lexical knowledge represented in the form of feature descriptions elicited from language speakers. Our work has illustrated the steps which are needed to encode feature descriptions in the WN model. To this purpose we introduced twenty new wordnet relations, and relied on phrasets for representing complex (non-lexicalized) concepts.

The study presented in these pages is a necessary theoretical step for the development of our tool. A practical evaluation of its feasibility is planned for the very near future, together with other (equally important but less relevant in this context) issues concerning both the population of our semantic knowledge base and the overall design of STaRS.sys.

Acknowledgements. We are grateful to Rita Capasso and Alessia Monti for the useful discussion of the application scenario sketched in these pages.

References

Luisa Bentivogli and Emanuele Pianta. 2003. Beyond Lexical Units: Enriching WordNets with Phrasets. *Proceedings of EACL 2003*: 67-70.

Luisa Bentivogli and Emanuele Pianta. 2004. Extending WordNet with Syntagmatic Information. *Proceedings of the 2nd International WordNet Conference*: 47-53.

Mary Boyle and Carl A. Coelho. 1995. Application of semantic feature analysis as a treatment for aphasia dysnomia. *American Journal of Speech-Language Pathology*, 4: 94-98.

Erminio Capitani, Marcella Laiacona, Brad Z. Mahon and Alfonso Caramazza. 2003. What are the Facts of Semantic Category-Specific Deficits? A Critical Review of the Clinical Evidence. *Cognitive Neuropsychology*, 20(3): 213-261.

George S. Cree and Ken McRae. 2003. Analyzing the Factors Underlying the Structure and Computation of the Meaning of Chipmunk, Cherry, Chisel, Cheese, and Cello (and Many Other Such Concrete Nouns). *Journal of Experimental Psychology: General*, 132 (2): 163-201.

Christiane Fellbaum. 1998 ed. *WordNet: an electronic lexical database*. The MIT Press.

Harold Goodglass and Arthur Wingfield. 1997. *Anomia: Neuroanatomical & Cognitive Correlates*. Academic Press.

Gerhard Kremer, Andrea Abel and Marco Baroni. 2008. Cognitively salient relations for multilingual lexicography. *Proceedings of COLING-CogALex Workshop 2008*: 94-101.

Marcella Laiacona, Riccardo Barbarotto, Cristina Trivelli and Erminio Capitani. 1993. Dissociazioni Semantiche Intercategoriali. *Archivio di Psicologia, Neurologia e Psichiatria*, 54: 209-248.

Gianluca E. Lebani and Emanuele Pianta. 2010. A Feature Type Classification for Therapeutic Purposes: a preliminary evaluation with non expert speakers. *Proceedings of ACL-LAW IV Workshop*.

Gregory L. Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.

Brian Petheram. 2004, ed. Special Issue on Computers and Aphasia. *Aphasiology*, 18 (3): 187-282.

Emanuele Pianta, Luisa Bentivogli and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. *Proceedings of the 1st International Conference on Global WordNet*.

Anastasia Raymer and Leslie Gonzalez-Rothi. 2002. Clinical Diagnosis and Treatment of Naming Disorders. In A.E. Hillis (ed) *The Handbook of Adult Language Disorders*. Psychology Press: 163-182.

Giuseppe Sartori and Luigi Lombardi. 2004. Semantic Relevance and Semantic Disorders. *Journal of Cognitive Neuroscience*, 16 (3): 439-452.

Carlo Semenza. 1999. Lexical-semantic disorders in aphasia. In G. Denes and L. Pizzamiglio (eds.). *Handbook of Clinical and Experimental Neuropsychology*. Psychology Press, Hove: 215-244.

David P. Vinson, Gabriella Vigliocco, Stefano Cappa and Simona Siri. 2003. The Breakdown of Semantic Knowledge: Insights from a Statistical Model of Meaning Representation. *Brain and Language*, 86: 347-365.

Textual Entailment Recognition using Word Overlap, Mutual Information and Subpath Set

Yuki Muramatsu
Nagaoka University of Technology
muramatsu@jnlp.org

Kunihiro Udaka
Nagaoka University of Technology
udaka@jnlp.org

Kazuhide Yamamoto
Nagaoka University of Technology
yamamoto@jnlp.org

Abstract

When two texts have an inclusion relation, the relationship between them is called entailment. The task of mechanically distinguishing such a relation is called recognising textual entailment (RTE), which is basically a kind of semantic analysis. A variety of methods have been proposed for RTE. However, when the previous methods were combined, the performances were not clear. So, we utilized each method as a feature of machine learning, in order to combine methods. We have dealt with the binary classification problem of two texts exhibiting inclusion, and proposed a method that uses machine learning to judge whether the two texts present the same content. We have built a program capable to perform entailment judgment on the basis of word overlap, i.e. the matching rate of the words in the two texts, mutual information, and similarity of the respective syntax trees (Subpath Set). *Word overlap* was calclated by utilizing BiLingual Evaluation Understudy (BLEU). *Mutual information* is based on co-occurrence frequency, and the *Subpath Set* was determined by using the Japanise WordNet. A Confidence-Weighted Score of 68.6% was obtained in the mutual information experiment on RTE. Mutual information and the use of three methods of SVM were shown to be effective.

1 Introduction

This paper can help solve textual entailment problems. Researchers of natural language processing have recently become interested in the automatic recognition of textual entailment (RTE), which is the task of mechanically distinguishing an inclusion relation. Text implication recognition is the task of taking a text (T) and a hypothesis (H), and judging whether one (the text) can be inferred from the other (hypothesis). Here below is an example task. In case of entailment, we call the relation to be 'true').

Example 1: Textual entailment recognition.
T: Google files for its long-awaited IPO.
H: Google goes public.
Entailment Judgment: True.

For such a task, large applications such as question answering, information extraction, summarization and machine translation are involved. A large-scale evaluation workshop has been conducted to stimulate research on recognition of entailment (Dagan et al., 2005). These authors divided the RTE methods into six methods. We focused on 3 methods of them.

Pérez and Alfonseca's method (Pérez and Alfonseca, 2005) used *Word Overlap*. This method is assumed to have taken place when words or sentences of the text and the hypothesis are similar, hence the relation should be true. Pérez and Alfonseca used the BLEU algorithm to calculate the entailment relationship. Glickman et al's. method was considered as using statistical lexical relations. These authors assumed that the possibility of entailment were high when the co-occurrence frequency of the word in the source and the target were high.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 18–27,
Beijing, August 2010

While this may be correct, we believe nevertheless that it problematic not to consider the co-occurrence of the hypothesis words. This being so, we proposed to use *mutual information*. Finally, Herrera et al's. method is based on Syntactic matching. They calculated the degree of similarity of the syntax tree.
We combined these three methods using machine learning techniques.

2 Related Works

Dagan et al. (Dagan et al, 2005) conducted research in 2005 on how to evaluate data of RTE; the authors insisted on the need of semantic analysis. As a first step, they considered the problem of textual entailment, proposing how to build evaluation data. Theu also organised and workshop on this topic. Their evaluation data are problems of binary classification of the texts to be compared. They used a sentence extracted from a newspaper corpus, and built a hypothesis from this text using one of seven methods: question answering, sentence comprehension, information extraction, machine translation, paraphrasing, information retrieval and comparable documents. They proposed a method of evaluation using RTE, and they introduced several RTE methods.

Odani et al. (Odani et al, 2005) did research on the construction of evaluation data in Japan, mentionning that there was a problem in the evaluation data of Dagan et al. For example, they stated that 'The evaluation data that he constructed are acting some factors. So it is difficult to discuss the problem'. Next, they did an RTE evaluation data using Japanese. The inference factors for judging entailment judgment were divided into five categories: inclusion, lexicon (words that can't be declined), lexicon (declinable words), syntax and inference. The subclassification was set for each classification, and Japanese RTE evaluation data was constructed. In addition, a dictionary and Web text were used for the entailment judgment. The authors were able to solve entailment judgment with words or phrases containing synonyms and/or a super-sub type relation. However, this classification lacks precision.

For example, they defined the term 'lexicon (words that cannot be declined)' as 'The meaning and the character of the noun that exists in text are data from which information on the truth of hypothesis is given'. Given this lack of clarity, we considered this method to be difficult to reproduce.

However, the evaluation data they built is general and available for public use. Regarding the research using the evaluation data of such RTE, there have been many reports in the workshop.

For example, Pérez and Alfonseca (Pérez and Alfonseca, 2005) assumed that the possibility of entailment was high when the text matched the hypothesis. The concordance rate of the text and the hypothesis was then calculated for judging the text and the hypothesis of the inclusion relation. In their research, they used BiLingual Evaluation Understudy (BLEU) to evaluate machine translation. An entailment judgment of 'true' was given when the BLEU score was above than a given threshold decided in advance. The evaluation data of Dagan et al. was used in the experiment, and its accuracy was about 50%. The evaluation data of comparable document types were the results with the highest accuracy. Hence the authors concluded that this method can be considered as a baseline of RTE. We dealt with it as *word overlap*.

Glickman et al. (Glickman et al, 2005) conducted research using co-occurring words. They assumed that the entailment judgment was 'true' when the probability of co-occurrence between the text and the hypothesis was high. In addition, the content word of the text with the highest co-occurrence probability was calculated from the content word of all of the hypotheses, and it was proposed as a method for entailment judgment. A Web search engine was used to calculate in the co-occurrence probability. This experiment yielded an accuracy of approximately 58%, while the evaluation data of comparable document types was about 83%. This being so, the authors concluded that they have been able to improve the results with the help of other deep analytical tools. We improved this method, and used it as mutual information.

Herrera et al. (Herrera et al., 2005) focused on syntactic similarity. They assumed that the entailment judgment was 'true' when the syntactic similarity of the text and the hypothesis was high. In addition, they used WordNet for considering identifiable expressions. The results

of the experiment yielded an accuracy of approximately 57%. We improved this method, and used it then as subpath set.

Prodromos Malakasiotis and Ion Androutsopoulos (Prodromos Malakasiotis and Ion Androutsopoulos, 2007) used Support Vector Machines. They assumed that the entailment judgment was 'true' when the similarity of words, POS tags and chunk tags were high. The results of the experiment yielded an accuracy of approximately 62%. However, they forgot to combine past RTE methods as feature of SVM.

The authors of this paper present a new RTE method. We propose to combine word overlap, mutual information and subpath sets. We dealt with SVM by using 3 methods equally as features, and we estimated higher precision than when using individual; independent methods.

3 Textual Entailment Evaluation Data

We used the textual entailment evaluation data of Odani et al. for the problem of RTE. This evaluation data is generally available to the public at the Kyoto University[1].

The evaluation data comprises the inference factor, subclassification, entailment judgment, text and hypothesis. Table 1 gives an example. The inference factor is divided into five categories according to the definition provided by Odani et al.: inclusion, lexicon (indeclinable word), lexicon (declinable word), syntax and inference. They define the classification viewpoint of each inference factor as follows:

Example 2: Classification criteria of inference factors

· Inclusion: The text almost includes the hypothesis.

· Lexicon (Indeclinable Word): Information of the hypothesis is given by the meaning or the behaviour of the noun in the text.

· Lexicon (Declinable Word): Information of the hypothesis is given by the meaning or the behaviour of the declinable word in the text.

· Syntax: The text and the hypothesis have a relation of syntactic change.

· Inference: Logical form.

They divided the data into 166 subclasses, according to each inference factor. The entailment judgment is a reliable answer in the text and the hypothesis. It is a difficult problem to entailment judgment for the criteria answer. Therefore, when they reported on the RTE workshop, they assumed the following classification criteria:

Example 3: Classification criteria of entailment determination.

· ◎(T_{alw}): When the text is true, the hypothesis is always true.

· ○(T_{alm}): When the text is true, the hypothesis is almost true.

· △(F_{may}): When the text is true, the hypothesis may be true.

· ×(F_{alw}): When the text is true, the hypothesis is false.

In terms of the text and the hypothesis, when we observed the evaluation data, the evaluation data accounted for almost every sentence in both the texts and the hypotheses, and also the hypotheses were shorter than the texts.

There is a bias in the number of problems evaluated by the inference factor and by the subclassification. The number of evaluation data open to the public now stands at 2471.

Inference Factor	Sub-Classification	Entailment Judgment	Text	Hypothesis
Lexicon (Indeclinable Word)	Behavior	◎	Toyota opened a luxury car shop.	Lexus is a luxury car.

Table 1: RTE Evaluation data of Odani et al.

[1] http://www.nlp.kuee.kyoto-u.ac.jp/nl-resource

4 Proposal Method

Up now, a number of methods have been proposed for RTE. However, when the previous methods were combined, the performances were hard to judge. Hence, we used each method as a feature of machine learning, and combined them then.

The input text and the hypothesis were considered as a problem of binary classification ('true' or 'false'). Therefore, we employed support vector machines (Vapnik, 1998), which are often used to address binary classification problems (in fact, we implemented our system with Tiny SVM). With this method we achieved higher precision than with individual independent methods.

Figure 1 shows our proposed method.

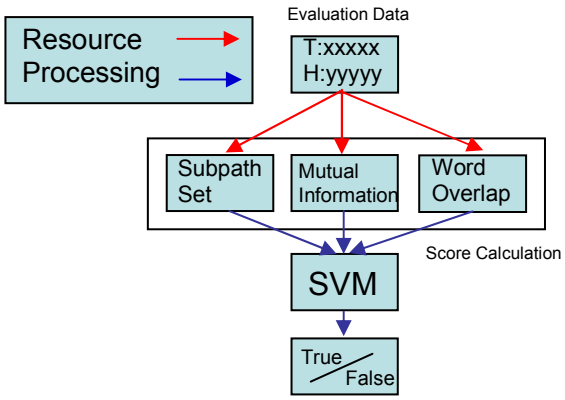


Figure 1: Our Proposed Method

In the following sections, we will describe the three features used in machine learning.

4.1 Word Overlap

It is assumed that when words or sentences of the text and the hypothesis are similar, the relation should be true. Pérez and Alfonseca used a BLEU algorithm to calculate the entailment between the text and the hypothesis. BLEU is often used to evaluate the quality of machine translation. Panieni et al. provided the following definition of BLEU. In particular, the BLEU score between length r of the sentence B and length c of the sentence A is given by the formulas (1) and (2):

$$Bleu(A,B) = BP \exp(\sum_{i=1}^{n} \log(p_i)/n) \quad (1)$$

$$BP = exp\left(1 - max\{1, r/c\}\right) \quad (2)$$

where p_i represents the matching rate of n-gram. The n-gram of this method was calculated as word n-gram. We assumed n = 1 and used the public domain program NTCIR7[2]. Here is an example of the calculation.

Example 4: Calculation by BLEU.
T:月は地球の衛星である。(The moon is Earth's satellite.)
H:月は地球の周りにある。(The moon is around the Earth.)
BLEU:0.75

We estimated n = 1 for the following reasons:

1. The reliability of word overlap is not high when n is large.
2. The calculated result of BLEU often becomes 0 when n is large.

First, we will explain the reason 1 mentioned above. The report of Kasahara et al. (Kasahara et al., 2010) is a reproduction of the one provided by Pérez et al (Pérez et al., 2005). They prepared an original RTE evaluation set of reading comprehension type, and proposed a new RTE system using a BLEU algorithm. When they experimented by increasing the maximum number of elements n of word n-gram from 1 to 4, the optimum maximum number of elements n is 3. They proposed the following analysis: if the hypothesis is shorter than the text, with n = 4, then the frequency is low in word 4-gram. However, the accidental coincidence of the word 4-gram significantly affected BLEU. When n is large, the reliability of the word overlap decreases.

Next, as an explanation of reason 2, when the length of the targeted sentence is short, the numerical result of BLEU sometimes becomes 0. For example, the number of agreements of 4-gram becomes 0 when calculating with n = 4, and the BLEU value sometimes becomes 0.

[2] http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/

Such calculations accounted for approximately 69% of the Odani et al. evaluation set.

4.2 Mutual Information

Glickman et al. (Glickman et al. 2005) assumed that the possibility of entailment is high when the co-occurrence frequency of the word in the text and the hypothesis is high. Therefore, they proposed a method of total multiplication, by searching for the word with the highest co-occurrence frequency from all the words of the hypothesis, as shown in formulas (3) and (4):

$$P(Trh=1|t)=\prod_{u\in h}\max_{V\in t} lep(u,v) \quad (3)$$

$$lep(u,v)\approx\frac{n_{u,v}}{n_v} \quad (4)$$

$P(Trh=1|t)$ expresses the probability of entailment between the text and the hypothesis. In these formulas, u is the content word of the hypothesis (noun, verb, adjective or unknown word); v the content word of the text; n represents the number of Web search hits; $n_{u,v}$ is the number of hits when the words u and v are searched on the Web. But, when the content word of the text is low frequency, the numerical result of the *lep(u, v)* increases for $P(Trh=1|t)$. We believe that it was a problem not to take into account the co-occurrence of the hypothesis words. In addition, their method to handle long sentences and reaching the conclusion 'false' is problematic. This is why, we considered Rodney et al's. method (Rodney et al. 2006) and proposed the use of mutual information, which is calculates on the basis of the formulas (5) and (6):

$$P(Trh=1|t)=\frac{1}{\mathbf{u}}\prod_{u\in h}\max_{V\in t} lep(u,v) \quad (5)$$

$$lep(u,v)\approx-\log\frac{p(n_{u,v})}{p(n_u)\cdot p(n_v)} \quad (6)$$

u is the number of the content words of the hypothesis. Hence, 1/**u** averages product of max *lep(u,v)*. This being so we considered that this model can do entailment judgments independantly of the length of the hypothesis.

It searches for the word of the text considering that the mutual information reaches the maximum value from each of the hypothesis words. When $P(Trh=1|t)$ is higher than an arbitrary threshold value, it is judged to be 'true', and 'false' in the opposite case. Glickman assumed the co-occurrence frequency to be the number of Web-search hits. However, we estimated that the reliability of the co-occurrence frequency was low, because the co-occurrence of the Web search engine was a wide window. This is why, we used the Japanese Web N-gram[3]. In particular, we used 7-gram data, and calculated the co-occurrence frequency $n_{u,v}$, frequency n_u and n_v of the word. $p(n_i)$ was calculated by (?) the frequency n_i divided the number of all words. Japanese Web N-gram was made from 20,036,793,177 sentences, including 255,198,240,937 words. The unique number of 7-gram is 570,204,252.

To perform morphological analysis, we used Mecab[4], for example:

Example 5: Calculation by mutual information.
T:この部屋はクーラーが効いている。(The air conditioner works in this room.)
H:涼しい。(It is cool.)
Mutual Information:10.0

$$P(Trh=1|t)=\frac{1}{1}\prod_{u\in}\max_{V\in t} lep(u,v) \quad (7)$$

$$lep(u,v)=-\log\frac{p\left(n_{\text{涼しい,クーラー(cool, the air conditioner)}}\right)}{p\left(n_{\text{涼しい(cool)}}\right)\cdot p\left(n_{\text{クーラー(the air conditioner)}}\right)}\approx 10.0 \quad (8)$$

This method actually standardises the result by dividing by the maximum value of *lep(u, v)*. As a result, p reaches the value 1 from 0. We used the discounting for n_u, n_v,and $n_{u,v}$, because a zero-frequency problem had occurred when calculating the frequency. There are some methods for discounting. We used the additive method reported by Church and Gale (Church and Gale, 1991). They compared some discounting methods by using the newspaper corpus. The addition method is shown as follows.

$$P(w)=\frac{C(w)+1}{N+V} \quad (9)$$

[3] http://www.gsk.or.jp/catalog/GSK-2007-C/
[4] http://mecab.sourceforge.net/

The additive method assumed N to be the number of all words in a corpus. *C(w)* is the frequency of word *w* in the corpus. V is a constant to adjust the total of the appearance probability to 1. It is equal to the unique number of words w. The additive method is very simple, it adds a constant value to occurrence count *C(w)*. The method of adding 1 to the occurrence count is called Laplace method also.

4.3 Subpath Set

Herrera et al. (Herrera et al., 2005) parsed the hypothesis and the text, and they calculated the degree of similarity of the syntax tree from both. Our method also deals with the degree of similarity of the syntax tree. The tree kernel method of Collins and Duffy （M. Collins and N. Duffy, 2002）shows the degree of similarity of the syntax tree; however, it requires much time to calculate the degree of similarity. Therefore, we employed the subpath set of Ichikawa et al. This latter calculates partial routes from the root to the leaf of the syntax tree. Our method assumes the node to be a content word (noun, verb, adjective or unknown word) in the syntax tree, while the branch is a dependency relation. For parsing we relied on Cabocha[5] .

The frequency vector was assumed to comprise a number of partial routes, similar to the approach of Ichikawa et al. (Ichikawa et al., 2005). The number of partial routes is unique. However, even if the same expression is shown for the word with a different surface form, it is not possible to recognise it as the same node. Therefore, we used the Japanese version of WordNet (Bond et al., 2009), in which a word with a different surface can be treated as the same expression, because Japanese WordNet contains synonyms. The same expressions of our method were hypernym words, hyponym words and synonym words in Japanese Word Net, because RTE sometimes considered the hierarchical dictionary of the hypernym and the hyponym word to be the same expression. However, our hypernym and hyponym words were assumed to be a parent and a child node of the object word, as shown in Figure 3.

[5] http://chasen.org/~taku/software/cabocha/

Example 6: Calculation by subpath set.

T:キャンペーン中なので、ポイントが 2 倍付く。

(T:The point adheres by the twice because it is campaigning.)

H:キャンペーン中なので、ポイントが普段より 2 倍付く。

(H:The point adheres usually by the twice because it is campaigning.)

Subpath:0.86

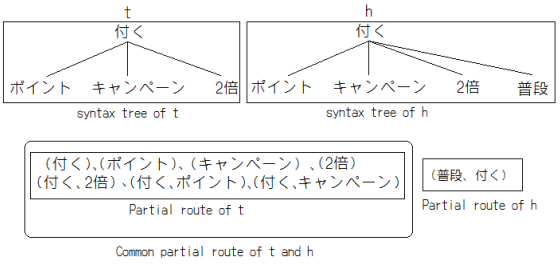


Figure 2: Partial route chart of subpath set.

The number of partial routes is 7, and 6 partial routes overlap in T and H. So, the subpath is 0.86 (6/7).

5 Evaluation

The textual entailment evaluation data of Odani et al., described in Section 3, was used in the experiment. The entailment judgment of four values is manually given to the textual entailment evaluation data. In our experiment we considered 'T_{alw}' and 'T_{alm}' to be 'true' and 'F_{may}' and 'F_{alw}' as 'false'. The evaluation method used was a Confidence-Weight Score (CWS, also known as Average Precision), proposed by Dagan et al.. As for the closed test, the threshold value with the maximum CWS was used.

$$Accuracy = Correct / All \quad (10)$$

$$CWS = \frac{1}{k} \sum_{1 \le i \le All} r_k \cdot precision(k) \quad (11)$$

$$precision(k) = \frac{1}{k} \sum_{1 \le i \le k} r_i \quad (12)$$

All = Number of all evaluation data. Correct = Number of correct answer data. If *k* is a correct answer, *rk* = 1. If *k* is an incorrect answer, *rk* = 0.

When the Entailment judgment annotated in evaluation data matches with the Entailment judgment of our method, the answer is true.

The threshold of the Closed test was set beforehand ($0 \leqq th \leqq 1$). When it was above the threshold, it was judged "true". When it was higher than the threshold, it was judged "false". SVM was used to calculate the value of three methods (word overlap, mutual information and subpath set) as the features for learning data, was experimented.

Open test was experimented 10-fold cross-validations. 9 of the data divided into 10 were utilized as the learning data. Remaining 1 was used as an evaluation data. It looked for the threshold that CWS becomes the maximum from among the learning data. It experimented on the threshold for which it searched by the learning data to the evaluation data. It repeats until all data that divides this becomes an evaluation data, averaged out. (Or we experimented Leave-one-out cross validation.)

Using the SVM, experiments were conducted on the numerical results of Sections 4.1 to 4.3 as the features.

The textual entailment evaluation data numbered 2472: 'T_{alw}': 924, 'T_{alm}': 662, 'F_{may}': 262 and 'F_{alw}': 624, and there were 4356 words. The total number of words was 43421. Tables 2 and 3 show the results of the experiment, which focused respectively on the closed and open tests,. When the 'true' textual entailment evaluation data 'T_{alw}' only and 'T_{alw} and T_{alm}' was used, mutual information achieved the best performance. When the true data 'T_{alm}' only was used, SVM achieved the best performance.

6 Discussion

In this section, we discuss the relation between each 3 method value assumed to be the criterion of judgment and CWS in the closed test. When the 'true' evaluation data was assumed to be 'T_{alm}' only in the open test, the result of SVM exceeded the results of the closed test. We then consider the relation between SVM and CWS.

6.1 Close Test of Ward Overlap

We believe that the results of the experiments of word overlap were more effective than other methods, because they achieved the best performance excluding 'T_{alm}' and 'T_{alw} and T_{alm}' in 3 methods. Figure 3 shows the relation to CWS when BLEU value changes.

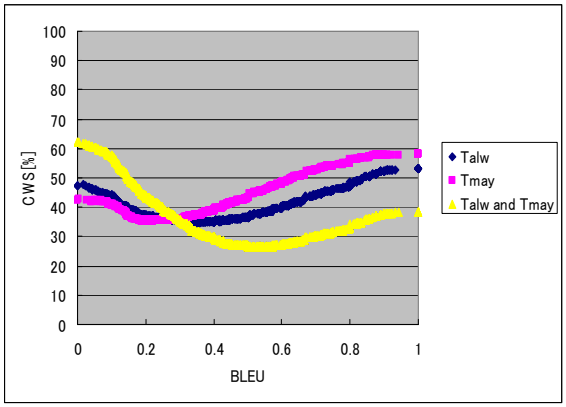


Figure 3: Results of the closed test of the RTE experiments by word overlap.

The tendency shown in Figure 3 did not change much when the relation between the threshold value and CWS was observed, even though the 'true' evaluation data was changed.

	CWS					
	Closed Test			Open Test		
	T_{alw}	T_{alm}	T_{alw} and T_{alm}	T_{alw}	T_{alm}	T_{alw} and T_{alm}
Word Overlap	53.0%	57.9%	62.1%	39.0%	60.2%	59.3%
Mutual Informaition	**55.9%**	52.9%	**68.6%**	**53.4%**	55.6%	**67.4%**
Subpath Set	54.5%	57.0%	61.8%	45.0%	59.7%	61.1%
SVM	51.4%	**61.2%**	63.5%	49.9%	**61.9%**	64.1%

Table 2: Results of the RTE experiments

However, the entailment judgment of the word overlap method becomes nearly 'false' when the BLEU value is 1 (or 'true' when BLEU score is 0.) Table 3 shows the entailment judgment when the BLEU value is 0 or 1.

We assumed that BLEU value that CWS becomes the maximum depends on the ratio of number of T and F in the evaluation set. However, when true condition is "T_{alw}" only, T is more than F (T:924,F:886). And when true condition is "T_{alm}" only, F is more than T (T:662,F:886). For this reason, The possibility of our assumption is low because both true conditions are BLEU value that CWS becomes the maximum is 1.

6.2 Close Test of Mutual Information

We believe that the results of the experiments of mutual information were more effective than other methods, because they achieved the best performance excluding 'T_{alm}' in 3 methods. Figure 4 shows the relation to CWS when mutual information value changes.

The tendency shown in Figure 4 did not change much when the relation between mutual information value and CWS was observed, even though the 'true' evaluation data was changed. When mutual information values are from 0.2 (or 0.3) to 1, CWS increased. However, the entailment judgment of the mutual information method becomes almost 'true' when mutual information score is near 1 (or 'false' when mutual information score value is near 0.)

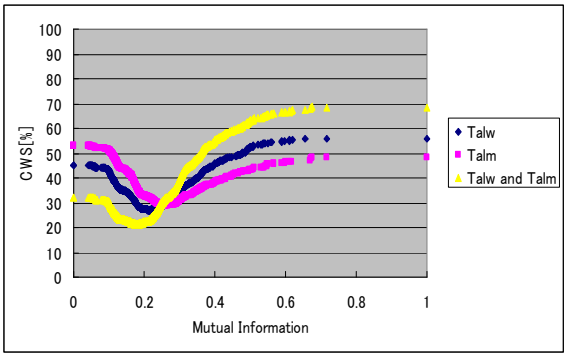


Figure 4: Results of the closed test of the RTE experiments by mutual information.

Table 4 shows the entailment judgment when the mutual information value is near 0 or 1. Our results showed most entailment judgment results to be almost 'true' (or almost 'false') for the optimal threshold value in the evaluation data. Therefore, we considered that the method of RTE using mutual information should be reviewed.

6.3 Close Test of Subpath Set

We believe that the results of the experiments of subpath set were not better than other methods. Figure 5 shows the relation to CWS when subpath set (SS) value changes.

The tendency shown in Figure 5 changed much when the relation between the threshold value and CWS was observed, even though the 'true' evaluation data was changed. When the true conditions are "T_{alw}" and "T_{alm}", the tendencies were very near.

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (Bleu=1)	5	919	12	874	53.0
	T_{alm} (Bleu=1)	0	662	12	874	57.9
	T_{alw} and T_{alm} (Bleu=0)	1586	0	886	0	62.1

Table 3: Entailment judgment in closed test of word overlap (T=True, F=False).

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (MI=0.72)	924	0	884	2	55.9
	T_{alm} (MI=0)	0	2	662	884	52.9
	T_{alw} and T_{alm} (MI=0.68)	1586	0	884	2	68.6

Table 4: Entailment judgment in closed test of mutual information (T=True, F=False, MI=mutual information).

However, when the true conditions were "T_{alw}" and "T_{alw} and T_{alm}", the tendencies were different. The tendency of "T_{alw}" was rising. The tendency of "T_{alw} and T_{alm}" was dropping until the subpath set value was 0.2. The entailment judgment of the mutual information method becomes almost 'true' when subpath set value was near 1)

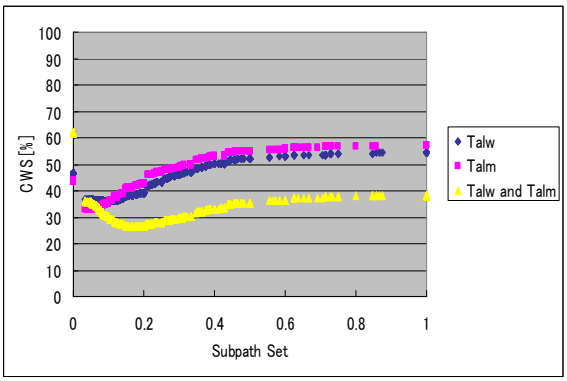


Figure 5: Results of the closed test of the RTE experiments by subpath set.

Table 5 shows the entailment judgment when the threshold value is near 0 or 1. Our results showed most entailment judgment results to be almost 'true' (or almost 'false') for the optimal subpath set value in the evaluation data.

6.4 Open Test of SVM

The open tests were conducted in 10-fold cross-validation , and the experimental result is their average. Figure 6 shows the related chart 10-fold cross-validation.

When the true data were assumed to be 'T_{alm}' only, the maximum value of CWS was 70.3%. As a result, the result of 10–fold cross validation exceeded the closed test.

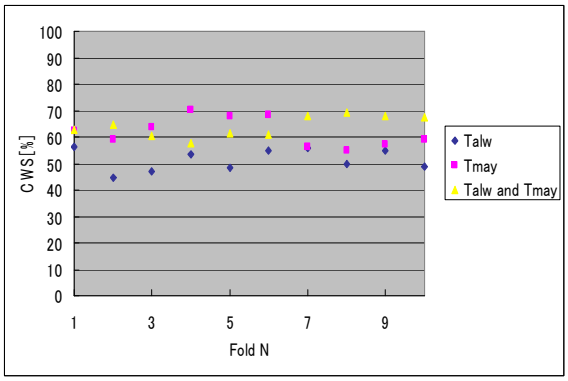


Figure 6: Results of the open test of the RTE experiments by SVM.

When the true data was assumed to be 'T_{alw}' only, the minimum value of CWS was 42.7%. We focused on the difference between the maximum and minimum value in 10-fold cross-validation. When the true answer was assumed to be 'T_{alm}', the difference between the maximum and minimum value is the greatest (15.3 points) in the open tests, and 'T_{alw} and T_{alm}' was the lowest with 11.6 points.

We believe that when the result 'T_{alm}' was 'true', it was consequently more unstable than 'T_{alw} and T_{alm}', because there was a larger amount of evaluation data 'T_{alw} and T_{alm}'.

7 Conclusion

We built a Japanese textual entailment recognition system based on the past methods of RTE. We considered the problem of RTE as a problem of binary classification, and built a new model of RTE for machine learning. We proposed machine learning to consider the matching rate of the words of the text and the hypothesis, using mutual information and similarity of the syntax tree. The method of using mutual information and the use of three methods of SVM tunrned out to be effective.

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (SS=1)	9	915	14	872	54.5
	T_{alm} (SS=1)	1	661	14	872	57.0
	T_{alw} and T_{alm} (SS=0)	1586	0	886	0	61.8

Table 5: Entailment judgment in closed test of subpath set (T=True, F=False, SS=subpath set).

In the future, we will consider changing the domain of the evaluation data and the experiment. Moreover, we will propose a new method for the feature of machine learning.

We will also consider to expand WordNet. Shnarch et al. (Shnarch et al., 2009) researched the extraction from Wikipedia of lexical reference rules, identifying references to term meaning triggered by other terms. They evaluated their lexical reference relation for RTE. They improved previous RTE methods. We will use their method for ours in order to expand Japanese WordNet. We believe that this can help us improve our method/results.

References

Michitaka Odani, Tomohide Shibata, Sadao Kurohashi, Takayuki Nakata, Building data of japanese Text Entailment and recognition of inferencing relation based on automatic achieved similar expression. *In Proceeding of 14th Annual Meeting of the Association for Natural Language Processing,* pp. 1140-1143, 2008 (in Japanese)

Diana Pérez and Enrique Alfonseca. Application of the Bleu algorithm for recognising textual entailment. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 9-12, 2005

Oren Glickman, Ido Dagan and Moshe Koppel. Web Based Probabilistic Textual Entailment. *In Proceedings of the PASCAL Recognizing Textual Entailment Challenge*, pp. 33-36, 2005

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. Enhancing the Japanese WordNet. *In the 7th Workshop on Asian Language Resources*, in conjunction with ACL-IJCNLP, pp. 1-8, 2009

Hiroshi Ichikawa, Taiichi Hashimoto, Takenobu Tokunaka and Hodumi Tanaka. New methods to retrieve sentences based on syntactic similarity. *Information Processing Society of Japan SIGNL Note*, pp39-46, 2005(in Japanese)

Kishore Panieni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, 2002

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognizing Textual Entailment Challenge. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 1-8,2005

Jesús Herrera, Anselmo Peñas and Felisa Verdejo, Textual Entailment Recognition Based on Dependency Analysis and WordNet. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 21-24, 2005

Kaname Kasahara, Hirotoshi Taira and Masaaki Nagata, Consider of the possibility Textual Entailment applied to Reading Comprehension Task consisted of multi documents. *In Proceeding of 14th Annual Meeting of the Association for Natural Language Processing,* pp. 780-783, 2010 (in Japanese)

M. Collins and N. Duffy. Convolution kernel for natural language. *In Advances in Neural Information Proccessing Systems (NIPS)*, volume 16, pages 625–632, 2002.

Prodromos Malakasiotis and Ion Androutsopoulos. Learning Textual Entailment using SVMs and String Similarity Measures. *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 42-47, 2007

Vladimir N. Vapnik, The Statisitcal Learning Theory. Springer, 1998.

Church, K. W. & Gale, W. A.. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, volume 5, 19-54.

Rodney D. Nielsen, Wayne Ward and James H. Martin. Toward Dependency Path based Entailment. *In Proceedings of the second PASCAL Recognizing Textual Entailment Challenge*, pp. 44-49, 2006

Eyal Shnarch, Libby barak, Ido Dagan. Extracting Lexical Reference Rules from Wikipedia. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 450-458, 2009

The Color of Emotions in Texts

Carlo Strapparava and **Gozde Ozbal**
FBK-irst
strappa@fbk.eu, gozbalde@gmail.com

Abstract

Color affects every aspect of our lives. There has been a considerable interest in the psycholinguistic research area addressing the impact of color on emotions. In the experiments conducted by these studies, subjects have usually been asked to indicate their emotional responses to different colors. On the other side, sensing emotions in text by using NLP techniques has recently become a popular topic in computational linguistics. In this paper, we introduce a semantic similarity mechanism acquired from a large corpus of texts in order to check the similarity of colors and emotions. Then we investigate the correlation of our results with the outcomes of some psycholinguistic experiments. The conclusions are quite interesting. The correlation varies among different colors and globally we achieve very good results.

1 Introduction

In our daily speech, we frequently make use of colors in order to increase our expressiveness by invoking different emotions. For instance, we usually stress the *redness* of someone's face for the implication of his/her anger or excitement, or we use phrases including the color *black* to refer to a depressed mood. On the other hand, the color *pink* is mostly used with positive connotations such as 'to see everything in pink light', where the meaning is related to optimism and happiness.

Actually, the parts of the nervous system which are responsible for emotional arousal are affected as soon as a color is perceived. Thus, the term *color emotion* has lately been used to represent the emotions arousing in human beings when they percept a color (Xin et al., 2004).

The correlation of color and emotion has been the focus of a lot of psycholinguistic studies so far. In the experiments conducted by these studies, subjects have usually been asked to indicate their emotional responses to different colors so that some general results stating which color arouses what kind of emotion could be obtained.

Sensing emotion, or in other words, affective sensing in text by using Natural Language Processing (NLP) techniques is recently a very popular topic in computational linguistics. There exist several studies focusing on the automatic identification of emotions in text with the help of both knowledge-based and corpus-based methods. Thus it is conceivable to explore whether state-of-the-art corpus analysis techniques can give support to psycholinguistic experiments.

Considering that psycholinguistic experiments are very costly since a lot of resources are required for both the setup and evaluation phases, employing a corpus-based approach for affective sensing could be much more efficient for all analysis to be held in the future, if this technique was proven to give reasonable results.

In this paper, we employ a semantic similarity mechanism automatically obtained in an unsupervised way from a large corpus of texts in order to check the similarity of color and emotion via computational analysis methods. We adopt the psycholinguistic experiments as references, with which we compare our results to find out if there is a correlation between the two approaches.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 28–32,
Beijing, August 2010

The paper is organized as follows. In Section 2, we introduce some related work focusing on the association of color and emotion only from a psycholinguistic point of view, since this topic has not been addressed by computational analysis techniques so far. In Section 3, we describe the methodology for implementing a similarity between colors and emotions, in particular how to represent an emotion in a latent semantic space. We present the evaluation of our approach and make a comparison with the results of psycholinguistic experiments in Section 4. Lastly, we report the conclusions and possible future work in Section 5.

2 Background

As we mentioned previously, there has been a considerable interest in the psycholinguistic research area addressing the impact of color on emotions.

(Zentner, 2001) mainly addressed the question of whether young children could show reliable color preferences. This study also tried to make a comparison with the results obtained with adults and older children. Subjects' color preferences were obtained by asking them to choose the one that they prefer among an array of colored cardboard rectangles. As an alternative way to represent musical information for providing feedback on emotion expression in music, (Bresin, 2005) suggested to use a graphical non-verbal representation of expressivity in music performance by exploiting color as an index of emotion. And for the purpose of determining which colors were most suitable for an emotional expression, some experiments were conducted, where subjects rated how well several colors and their nuances corresponded to music performances expressing different emotions. (Kaya, 2004) tried to investigate and discuss the associations between color and emotion by conducting experiments where college students were asked to indicate their emotional responses to principal, intermediate and achromatic colors, and the reasons for their choices.

There exist also some research investigating whether the color perception is related to the region of the subject. For instance, (Gao et al., 2007) analyzed and compared the color emotions of people from seven regions in a psychophysical experiment, with an attempt to clarify the influences of culture and color perception attributes on color emotions. This study suggested that it might be possible to compose a color emotion space by using a restricted number of factors. As for (Soriano and Valenzuela, 2009), this study tried to find out why there was often a relationship between color and emotion words in different languages. In order to achieve this, a new experimental methodology called the Implicit Association Test (IAT) was used to explore the implicit connotative structure of the Peninsular Spanish color terms in terms of Osgood's universal semantic dimensions explained in (Adams and Osgood, 1973). The research conducted by (Xin et al., 2004) tried to compare the color emotional responses that were obtained by conducting visual experiments in different regions by using a set of color samples. A quantitative approach was used in this study in an attempt to compare the color emotions among these regions. (Madden et al., 2000) focused on the possible power of color for creating and sustaining brand and corporate images in international marketing. This study tried to explore the favorite colors of consumers from different countries, the meanings they associated with colors, and their color preferences for a logo.

The study that we will use for evaluating our results is a work which focused on the topic "emotional responses to color used in advertisement" (Alt, 2008). During the experiments, this study conducted a survey where the subjects were required to view an advertisement with a dominant color hue, and then select a specific emotional response and a positive/negative orientation related with this color. More than 150 subjects participated in this study, roughly equally partitioned in gender. There are two main reasons why we preferred to use this study for our evaluation procedure. Firstly, the presentation and organization of the results provide a good reference for our own experiments. In addition, it focusses on advertisement, which is one of the applicative fields we want to address in future work.

3 Methodology

Sensing emotions from text is an appealing task of natural language processing (Pang and Lee,

2008; Strapparava and Mihalcea, 2007): the automatic recognition of affective states is becoming a fundamental issue in several domains such as human-computer interaction or sentiment analysis for opinion mining. Indeed, a large amount of textual material has become available form the Web (e.g. blogs, forums, social networks), raising the attractiveness of empirical methods analysis on this field.

For representing the emotions, we exploit the methodology described in (Strapparava and Mihalcea, 2008). The idea underlying the method is the distinction between *direct* and *indirect* affective words.

For direct affective words (i.e. words that directly denote emotions), authors refer to the WORDNET AFFECT (Strapparava and Valitutti, 2004) lexicon, a freely available extension of the WORDNET database which employs some basic emotion labels (e.g. anger, disgust, fear, joy, sadness) to annotate WORDNET synsets.

For indirect affective words, a crucial aspect is building a mechanism to represent an emotion starting from affective lexical concepts and to introduce a semantic similarity among generic terms (and hence also words denoting colors) and these emotion representations.

Latent Semantic Analysis is used to acquire, in an unsupervised setting, a vector space from the British National Corpus[1]. In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition on the term-by-document matrix representing the corpus. LSA has the advantage of allowing homogeneous representation and comparison of words, word sets (e.g. synsets), text fragments or entire documents. For representing word sets and texts by means of a LSA vector, it is possible to use a variation of the *pseudo-document* methodology described in (Berry, 1992). This variation takes into account also a *tf-idf* weighting schema. In practice, each document can be represented in the LSA space by summing up the normalized LSA vectors of all the terms contained in it. Therefore a set of words (and even all the words labeled with a particular emotion) can be represented in the LSA space, performing the pseudo-document technique on them.

As stated in (Strapparava and Mihalcea, 2008), each emotion can be represented in various ways in the LSA space. The particular one that we are employing is the 'LSA Emotion Synset' setting, which has proved to give the best results in terms of fine-grained emotion sensing. In this setting, the synsets of direct emotion words, taken form WORDNET AFFECT, are considered.

For our purposes, we compare the similarities among the representations of colors and emotions in the latent similarity space.

4 Experiments

For the experiments in this paper, we built an LSA vector space on the full BNC corpus using 400 dimensions. To compare our approach with the psycholinguistic experiments reported in (Alt, 2008), we represent the following emotions: anger, aversion/disgust, fear, happiness/joy, and sadness. And we consider the colors Blue, Red, Green, Orange, Purple, Yellow. Table 1 reports the rankings of emotions according to colors from (Alt, 2008).

Color	*Ranking of Emotions*				
	Anger	Aversion/ Disgust	Fear	Joy	Sadness
Blue	5	2	4	1	3
Red	1	4	2	3	5
Green	5	2	3	1	4
Orange	4	2	3	1	5
Purple	5	2	4	1	3
Yellow	5	2	4	1	3

Table 1: Emotions ranked by colors from psycholinguistic experiments

In Table 2 we report our results of ranking emotions with respect to colors using the similarity mechanism described in the previous section. To evaluate our results with respect to the psycholinguistic reference, we use Spearman correlation coefficient. The resulting correlation between two approaches for each color is reported in Table 3.

We can observe that the global correlation is rather good (0.75). In particular, it is very high

[1]BNC is a very large (over 100 million words) corpus of modern English, both spoken and written (see `http://www.hcu.ox.ac.uk/bnc/`). Other more specific corpora could also be considered, to obtain a more domain oriented similarity.

Color	*Ranking Emotions using Similarity with Color*				
	Anger	Aversion/ Disgust	Fear	Joy	Sadness
Blue	4	2	3	1	5
Red	4	3	2	1	5
Green	4	2	3	1	5
Orange	4	2	3	1	5
Purple	5	2	3	1	4
Yellow	4	2	3	1	5

Table 2: Emotions ranked by similarity with colors

Color	*Correlation*
Blue	0.7
Red	0.3
Green	0.9
Orange	1.0
Purple	0.9
Yellow	0.7
Total	0.75

Table 3: Correlation

for the colors Orange, Green and Purple, which implies that the use of language for these colors is quite in accordance with psycholinguistic results. The results are good for Blue and Yellow as well, while the correlation is not so high for Red. This could suggest that Red is a quite ambiguous color with respect to emotions.

5 Conclusions

There are emotional and symbolic associations with different colors. This is also revealed in our daily use of language, as we frequently make references to colors for increasing our expressiveness by invoking different emotions. While most of the research conducted so far with the aim of analyzing the relationship between color and emotion was based on psycholinguistic experiments, the goal of this study was exploring this association by employing a corpus-based approach for affective sensing.

In order to show that our approach was providing reasonable results, we adopted one of the existing psycholinguistic experiments as a reference. Following that adoption, we made a comparison between the results of this research and our own technique. Since we have observed that these two results were highly correlated as we expected, we would like to explore further this direction. Certainly different cultures can play a role for variant emotional responses (Adams and Osgood, 1973). Thus, as a next step, we are planning to investigate how the perception of color by human beings varies in different languages by again conducting a computational analysis with NLP techniques. Employing this approach could be very useful and efficient for the design of applications related to the fields of multimedia, automatic advertisement production, marketing and education (e.g. e-learning environments)

In addition, based on our exploration about the color perception of emotions from a corpus-based point of view, we suggest that "visual" information regarding objects and events could be extracted from large amounts of text, using the same kind of techniques proposed in the present paper. This information can be easily exploited for creation of dictionaries or used in dynamic visualization of text such as kinetic typography (Strapparava et al., 2007). As a concrete example, our approach can be extended to discover the association of colors not only with emotions, but also with indirect affective words in various languages. We believe that the discovery of this kind of relationship would allow us to automatically build colorful dictionaries, which could substantially help users with both interpretation and memorization processes.

References

Adams, F. M. and C. E. Osgood. 1973. A cross-cultural study of the affective meanings of colour. *Journal of cross-cultural psychology*, 4:135—156.

Alt, M. 2008. Emotional responses to color associated with an advertisement. Master's thesis, Graduate College of Bowling Green State University, Bowling Green, Ohio.

Berry, M. 1992. Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49.

Bresin, R. 2005. What is the color of that music performance? In *Proceedings of the International Computer Music Conference (ICMA 2005)*, pages 367–370.

Gao, X.P., J.H. Xin, T. Sato, A. Hansuebsai, M. Scalzo, K. Kajiwara, S. Guan, J. Valldeperas, M. Lis Jose,

and M. Billger. 2007. Analysis of cross-cultural color emotion. *Color research and application*, 32(223—229).

Kaya, N. 2004. Relationship between color and emotion: a study of college students. *College Student Journal*, pages 396–405.

Madden, T. J., K. Hewett, and S. Roth Martin. 2000. Managing images in different cultures: A cross-national study of color meanings and preferences. *Journal of International Marketing*, 8(4):90–107.

Ortony, A., G. L. Clore, and M. A. Foss. 1987. The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53:751–766.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Soriano, C. and J. Valenzuela. 2009. Emotion and colour across languages: implicit associations in spanish colour terms. *Social Science Information*, 48:421–445, September.

Strapparava, C. and R. Mihalcea. 2007. SemEval-2007 task 14: Affective Text. In *Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval 2007)*, pages 70–74, Prague, June.

Strapparava, C. and R. Mihalcea. 2008. Learning to identify emotions in text. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, New York, NY, USA. ACM.

Strapparava, C. and A. Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086.

Strapparava, C., A. Valitutti, and O. Stock. 2007. Dances with words. In *Proc. of the 20^{th} International Joint Conference on Artificial Intelligence (IJCAI-07)*, Hyderabad, India, January.

Xin, J.H., K.M. Cheng, G. Taylor, T. Sato, and A. Hansuebsai. 2004. A cross-regional comparison of colour emotions. part I. quantitative analysis. *Color Research and Application*, 29:451—457.

Zentner, M. R. 2001. Preferences for colors and color-emotion combinations in early childhood. *Developmental Science*, 4(4):389–398.

How to Expand Dictionaries with Web-Mining Techniques

Nicolas Béchet
LIRMM, UMR 5506, CNRS,
Univ. Montpellier 2
France
bechet@lirmm.fr

Mathieu Roche
LIRMM, UMR 5506, CNRS,
Univ. Montpellier 2
France
mroche@lirmm.fr

Abstract

This paper presents an approach to enrich conceptual classes based on the Web. To test our approach, we first build conceptual classes using syntactic and semantic information provided by a corpus. The concepts can be the input of a dictionary. Our web-mining approach deals with a cognitive process which simulates human reasoning based on the enumeration principle. The experiments reveal the interest of our approach by adding new relevant terms to existing conceptual classes.

1 Introduction

Concepts have several definitions; one of the most general describes a concept 'as the mind's representation of a thing or an item' (Desrosiers-Sabbath, 1984). In a domain such as ours, i.e. ontology building, semantic webs, and computational linguistics, it seems appropriate to stick to the Aristotelian approach to a concept, and consider it as a set of knowledge (gathered information) on common semantic features. The choice of the features and how the knowledge is gathered depend on criteria we will explain below.

In this paper, we deal with the building of conceptual classes, which can be defined as gathering semantically close terms. First, we suggest building specific conceptual classes by focusing on knowledge extracted from corpora.

Conceptual classes are shaped by the study of syntactic dependencies between corpus terms (as described in section 2). Dependencies tackle relations such as Verb/Subject, Noun/Noun Phrase Complements, Verb/Object, Verb/Complements, and sometimes Sentence Head/Complements. In this paper, we focus on the Verb/Object dependency because it is representative of a field. For instance, in computer science, the verb 'to load' takes as objects, nouns of the conceptual class software (L'Homme, 1998). This feature also extends to 'download' or 'upload', which have the same verbal root.

Corpora are rich sources of terminological information that can be mined. A terminology extraction of this kind is similar to a Harris-like distributional analysis (Harris, 1968) and many works in the literature have been the subject of distributional analysis to acquire terminological or ontological knowledge from textual data (e.g (Bourigault and Lame, 2002) for law, (Nazarenko *et al*., 2001; Weeds *et al*., 2005) for medicine).

After building conceptual classes (section 2), we describe an approach to expand concepts by using a Web search engine to discover new terms (section 3). In section 4, experiments conducted on real data enable us to validate our approach.

2 Building Conceptual Classes

2.1 Principle

In our approach, a class can be defined as a gathering of terms with a common field. In this paper, we focus on objects of verbs judged to be semantically close by using a measure. These objects are thus considered as instances of conceptual classes. The first step in building conceptual classes consists in extracting Verb/Object syntactic relations as explained in the following section.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 33–37,
Beijing, August 2010

2.2 Mining for Verb/Object relations

Our corpora are in French since our team is mostly devoted to French-based NLP applications. However, the following method can be used for any other language, provided a reliable dependency parser is available. In our case, we use the SYGFRAN parser developed by (Chauché, 1984). As an example, in the French sentence *"Thierry Dusautoir brandissant le drapeau tricolore sur la pelouse de Cardiff après la victoire."* (translation: 'Thierry Dusautoir brandishing the three colored flag on Cardiff lawn after the victory'), there is a verb-object syntactic relation: *"verb: brandir (to brandish), object: drapeau (flag)",* which is a good candidate for retrieval. The second step of the building process corresponds to the gathering of common objects related to semantically close verbs.

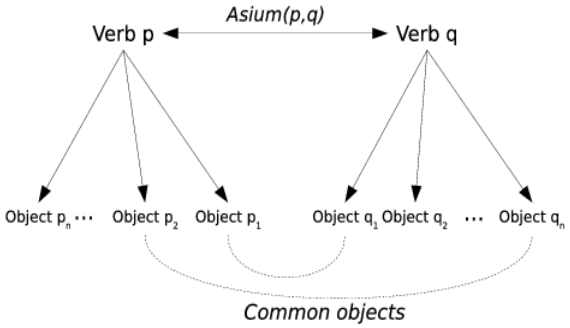


Figure 1: Common and complementary objects of the verbs "to consume" and "to eat"

Assumption of Semantic Closeness. The underlying linguistic hypothesis is the following: Verbs with a significant number of common objects are semantically close.

To measure closeness, the ASIUM score (Faure and Nedellec, 1999; Faure, 2000) is used (see figure 1). This type of work is similar to distributional analysis approaches such as that of (Bourigault and Lame, 2002).

As explained in the introduction, the measure considers two verbs to be close if they have a significant number of common features (objects).

Let p and q be verbs with their respective $p_1,...,p_n$ and $q_1,...,q_m$ objects. $NbOC_p(q_i)$ is the number of occurrences of q_i objects from q that are also objects of p (common objects). $NbO(q_i)$ is the number of occurrences of q_i objects of q verb. The Asium measure is then:

$$Asium\ (p,q) = \frac{\log_{Asium}(\sum NbOC_q(p_i)) + \log_{Asium}(\sum NbOC_p(q_i))}{\log_{Asium}(\sum NbO(p_i)) + \log_{Asium}(\sum NbO(q_i))}$$

Where $log_{Asium}(x)$ is equal to:

- for $x = 0$, $log_{Asium}(x) = 0$
- else $log_{Asium}(x) = log(x) + 1$

Therefore, conceptual classes instances are the common objects of close verbs, according to the ASIUM proximity measure.

The following section describes the acquisition of new terms starting with a list of terms/concepts obtained with the global process summarized in this section and detailed in (Béchet *et al*., 2008).

3 Expanding conceptual classes

3.1 Acquisition of candidate terms

The aim of this approach is to provide new candidates for a given concept. It is based on enumeration on the Web of terms that are semantically close. For instance, with a query (string) "bicycle, car, and", we can find other vehicles. We propose to use the Web to acquire new candidates. This kind of method uses information regarding the "popularity" of the web and is independent of a particular corpus.

Our method of acquisition is quite similar to that of (Nakov and Hearst, 2008). These authors propose to query the Web using the Google search engine to characterize the semantic relation between a pair of nouns. The Google star operator among others, is used to that end. (Nakov and Hearst, 2008) refer to the study of (Lin and Pantel, 2001) who used a Web mining approach to discover inference rules missed by humans.

To apply our method, we first consider the common objects of semantically close verbs, which are instances of reference concepts (e.g. vehicle). Let N concepts $C_{i\in\{1,N\}}$ and their respective instances $I_j(C_i)$. For each concept C_i, we submit to a search engine the following queries: "$I_{jA}(C_i)$, $I_{jB}(C_i)$, and" and "$I_{jA}(C_i)$, $I_{jB}(C_i)$, or" with jA and $jB \in \{1, ..., NbInstanceC_i\}$ and $jA \neq jB$.

The search engine returns a set of results from which we extract new candidate instances of a concept. For example, if we consider the query: "bicycle, car, and", one page returned by a search engine gives the following text:
Listen here for the Great Commuter Race (17/11/05) between bicycle, car and **bus***, as part of...*

Having identified the relevant features in the result returned (in bold in our example), we add the term "bus" to the initial concept "vehicle". In this way, we obtain new candidates for our concepts. The process can be repeated. In order to automatically determine which candidates are relevant, the candidates are filtered as shown in the following section.

3.2 Filtering of candidates

The quality of the extracted terms can be validated by an expert, or automatically by using the Web to check if the extracted candidates (see section 3.1) are relevant. The principle is to consider a relevant term if it is often present with the terms of the original conceptual class (kernel of words). Thus, our aim is to validate a term "in the context". From that point of view, our method is close to that of (Turney, 2001), which queries the Web via the AltaVista search engine to determine appropriate synonyms for a given term. Like (Turney, 2001), we consider that information concerning the number of pages returned by the queries can give an indication of the relevance of a term.

Thus, we submit to a search engine different strings (using citation marks). A query consists of the new candidate and both terms of the concept. Formally, our approach can be defined as follows. Let N concepts $C_i \in \{1, N\}$, their respective instances $I_j(C_i)$ and the new candidates for a concept C_i, $N_{ik \in \{1, NbNI(Ci)\}}$. For each C_i, each new candidate N_{ik} is sent as a query to a Web search engine. In practice the three terms are separated either by a comma or the word "or" or "and"[1]. For each query, the search engine returns a number of results (i.e. number of web pages). Then, the sum of these results is calculated using all possible combinations of "or", "and", or of the three words (words of the kernel plus candidate word to enrich it). Below is an example with the kernel words "car", "bicycle" and the candidate "bus" to test (using Yahoo):

- "car, bicycle, and bus": 71 pages returned
- "car, bicycle, or bus": 268 pages returned
- "bicycle, bus, and car": 208 pages returned
- and so forth

Global result: 71 + 268 + 208...

The filtering of candidates consists in selecting the k first candidates by class (i.e. with the highest sum), they are added as new instances of the initial concept. We can reiterate the acquisition approach by including these new terms. The acquisition/filtering process can be repeated several times.

In the next section, we present experiments conducted to evaluate the quality of our approach.

4 Experiments

4.1 Evaluation protocol

We used a French corpus from the Yahoo site (http://fr.news.yahoo.com/) composed of 8,948 news items (16.5 MB) from newspapers. Experiments were performed on 60,000 syntactic relations (Béchet *et al.*, 2008; Béchet *et al.*, 2009) to build original conceptual classes. We manually selected five concepts (see Figure 2). Instances of these concepts are the common objects of verbs defining the concept (see section 2.2).

Concepts	Organisme /Administration	Fonction	Objets symboliques	Sentiment	Manifestation de protestation
	(Civil Service)	*(work)*	*(symbols)*	*(feeling)*	*(protest)*
Instances	parquet *(prosecution)*	négociateur *(negotiator)*	drapeau *(flag)*	mécontentement *(discontent)*	protestation *(remontrance)*
	mairie *(city hall)*	cinéaste *(filmmaker)*	fleur (flower)	souhait *(wish)*	grincement *(grind)*
	gendarme *(policeman)*	écrivain *(writer)*	spectre *(specter)*	déception *(disappointment)*	indignation *(indignation)*
	préfecture *(prefecture)*	orateur *(public speaker)*		désaccord *(disagreement)*	émotion *(emotion)*
	pompier *(fireman)*			désir *(desire)*	remous *(swirl)*
	O.N.U. *(U.N.)*				tollé *(collective protest)*
					émoi *(commotion)*
					panique *(panic)*

Figure 2: The five selected concepts and their instances.

[1] Note that the commas are automatically removed by the search engines.

For our experiments, we use an API of the search engine Yahoo! to obtain new terms. We apply the following post-treatments for each new candidate term. They are initially lemmatized. Therefore, we only keep the nouns, after applying a PoS (Part of Speech) tagger, the TreeTagger (Schmid, 1995).

After these post-treatments, we manually validate the new terms using three experts. We compute the precision of our approach to each expert. The average is calculated to define the quality of the terms. Precision is defined as follows.

$$Precision = \frac{\text{Number of relevant terms given by our system}}{\text{Number of terms given by our system}}$$

In the next section, we present the evaluation of our method.

4.2 Experimental results

Table 1 gives the results of the term acquisition method (i.e. for each acquisition step, we apply our approach to filter candidate terms). For each step, the table lists the degree of precision obtained after expertise:

- All candidates. We calculate the precision before the filtering step.
- Filtered candidates. After applying the automatic filtering by selecting k terms per class, we calculate the precision obtained. Note that the automatic filtering (see section 3.2) reduces the number of terms proposed, and thus reduces the recall[2].

	Precision		*Terms number*
Steps #	*All terms*	*Filtered terms*	*(without filter)*
1	0.69	0.83	29
2	0.69	0.77	47
3	0.56	0.65	103

Table 1: Results obtained with k=4 (i.e. automatic selection of the k first ranked terms by the filtering approach).

[2] The recall is not calculated because in an unsupervised context it is difficult to estimate.

Finally Table 1 shows the number of terms generated by the acquisition system.

These results show that a significant number of terms can be generated (i.e. 103 words). For example, for the concept 'feeling', using the initial terms given in figure 1, we obtained the following eight French terms (in two steps): "horreur (horror), satisfaction (satisfaction), déprime (depression), faiblesse (weakness), tristesse (sadness), désenchantement (disenchantment), folie (madness), fatalisme (fatalism)".
This approach is appropriate to produce new relevant terms to enrich conceptual classes, in particular when we select the first terms (k = 4) returned by the filtering system. In a future work, we plan to test other values of the automatic filtering. The precision obtained in the first two steps was high (i.e. 0.69 to 0.83). The third step returned lower scores; noise was introduced because we were too "far" from the initial kernel words.

5 Conclusion and Future Work

This paper describes an approach for conceptual enrichment classes based on the Web. We apply the "enumeration" principle to find new terms using Web search engines. This approach has the advantage of being less dependent on the corpus. Note that as the use of the Web requires validation of candidates, we propose an automatic filtering method to select relevant terms to add to the concept. In a future work, we plan to use other statistical web measures (e.g. Mutual Information, Dice measure, and so forth) to automatically validate terms.

References

Béchet, N., M. Roche, and J. Chauch´e. 2008. How the ExpLSA approach impacts the document classification tasks. In *Proceedings of the International Conference on Digital Information Management, ICDIM'08*, pages 241–246, University of East London, London, United Kingdom.

Béchet, N., M. Roche, and J. Chauché. 2009. Towards the selection of induced syntactic relations. In *European Conference on Information Retrieval (ECIR)*, Poster, pages 786–790.

Bourigault, D. and G. Lame. 2002. Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du droit. In *TAL*, pages 43–51.

Chauché, J. 1984. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of COLING, Standford University, California*, pages 11–15.

Desrosiers-Sabbath, R. 1984. *Comment enseigner les concepts.* Presses de l'Université du Québec.

Faure, D. and C. Nedellec. 1999. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI*, pages 329–334.

Faure, D. 2000. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM.* Ph.D. thesis, Université Paris-Sud, 20 Décembre.

Harris, Z. 1968. *Mathematical Structures of Language.* John Wiley & Sons, New-York.

L'Homme, M. C. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. In *Cahiers de Lexicologie 73*, pages 61–84.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Nakov, Preslav and Marti A. Hearst. 2008. Solving relational similarity problems using the web as a corpus. In ACL, pages 452–460.

Nazarenko, A., P. Zweigenbaum, B. Habert, and J. Bouaud. 2001. Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology,* pages 327–351.

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceed ngs of the ACL SIGDAT-Workshop, Dublin*.

Turney, P.D. 2001. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of ECML'01, Lecture Notes in Computer Science*, pages 491–502.

Weeds, J., J. Dowdall, G. Schneider, B. Keller, and D. Weir. 2005. Weir using distributional similarity to organise biomedical terminology. In *Proceedings of Terminology*, volume 11, pages 107–141.

An Optimal and Portable Parsing Method for Romanian, French, and German Large Dictionaries

Neculai Curteanu
Institute of Computer Science,
Romanian Academy,
Iaşi Branch
ncurteanu@yahoo.com

Alex Moruz
Institute of Computer Science, Romanian Academy;
Faculty of Computer Science,
"Al. I. Cuza" University, Iaşi
mmoruz@info.uaic.ro

Diana Trandabăţ
Institute for Computer Science, Romanian Academy;
Faculty of Computer Science,
"Al. I. Cuza" University, Iaşi
dtrandabat@info.uaic.ro

Abstract

This paper presents a cross-linguistic analysis of the largest dictionaries currently existing for Romanian, French, and German, and a new, robust and portable method for Dictionary Entry Parsing (DEP), based on Segmentation-Cohesion-Dependency (SCD) configurations. The SCD configurations are applied successively on each dictionary entry to identify its lexicographic segments (the first SCD configuration), to extract its sense tree (the second configuration), and to parse its atomic sense definitions (the third one). Using previous results on **DLR** (The Romanian Thesaurus – new format), the present paper adapts and applies the SCD-based technology to other four large and complex thesauri: **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), and **GWB** (Göthe-Wörterbuch). This experiment is illustrated on significantly large parsed entries of these thesauri, and proved the following features: (1) the SCD-based method is a completely *formal grammar-free* approach for dictionary parsing, with efficient (weeks-time adaptable) modeling through sense hierarchies and parsing portability for a new dictionary. (2) SCD-configurations separate and run sequentially and independently the processes of lexicographic segment recognition, sense tree extraction, and atomic definition parsing. (3) The whole DEP process with SCD-configurations is *optimal*. (4) SCD-configurations, through sense marker classes and their dependency hypergraphs, offer an unique instrument of lexicon construction comparison, sense concept design and DEP standardization.

1 Introduction

The general idea behind parsing a large dictionary can be reduced to transforming a raw text entry into an indexable linguistic resource. Thus, for each dictionary entry, a structured representation of its senses has to be created, together with a detailed description of the entry's form: i.e. morphology, syntax, orthography, phonetics, lexical semantics, etymology, usage, variants etc.

The aim of this paper is to present an efficient *dictionary entry parsing* (DEP) method, based on Segmentation-Cohesion-Dependency (SCD) configurations (Curteanu, 2006), applied on a set of five large and complex dictionaries: **DLR** (The Romanian Thesaurus – new format), **DAR** (The Romanian Thesaurus – old format), **TLF** (Le Trésor de la Langue Française), **DWB** (Deutsches Wörterbuch – GRIMM), and **GWB** (Göthe-Wörterbuch).

The paper is structured in 8 sections: Section 2 presents the state of the art in DEP, with an emphasis on the comparison between the proposed method and other dictionary parsing strategies, before detailing the SCD-based proposed method in Section 3. The following sections present the application of the proposed method to the five dictionaries. The paper ends with a discussion on

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 38–47,
Beijing, August 2010

comparative results and development directions concerning optimality, portability, standardization, and dictionary networks.

2 Dictionary Entry Parsing

Natural language text parsing is a complex process whose prerequisite essential stage is a thorough modeling of the linguistic process to be developed, *i.e.* the structures and relations aimed to constitute the final result of the analysis. Similarly, for DEP, the semantics of the lexical structures, the sense markers, and the hierarchies (dependencies) between sense structures must be specified.

Standard approaches to dictionary entry parsing (referred to from now on as *standard* DEP), such as the one used by (Neff and Boguraev, 1989), the *LexParse* system presented in (Hauser and Storrer, 1993; Kammerer, 2000; Lemnitzer and Kunze, 2005), or lexicographic grammars, as those presented in (Curteanu & Amihaesei, 2004; Tufis et al., 1999), recognize the sense / subsense definitions in a strictly sequential manner, along with the incremental building of the entry sense tree. The interleaving of the two running processes is the main source of errors and inefficiency for the whole DEP process.

Both the *standard* DEP (Figure 1) and our proposed method based on SCD-*configurations* (Figure 2) involve the following *three* running *cycles* and *four* essential *phases* for extracting the sense-tree structure from a dictionary:

[A1], **[B1]** – parsing the *lexicographic segments* of an entry;

[A2], **[B2]** – parsing the *sense-description segment* of the dictionary entry, at the level of explicitly defined senses, until and not including the contents of the atomic definitions / senses; at this stage, the *sense-tree* of the sense-description segment is built having (sets of) atomic senses / definitions in their leaf-nodes.

[A3], **[B3]** – parsing the *atomic definitions* / *senses*.

Phase_1 := Sense-*i* Marker Recognition;
Phase_2 := Sense-*i* Definition Parsing;
Phase_3 := Attach Parsed Sense-*i* Definition to Node-*i*;
Phase_4 := Add Node-*i* to EntrySense-Tree.

The parsing cycles and phases of existing approaches, called *standard* DEP, are summarized by the pseudo-code in Fig. 1, where *MarkerNumber* is the number of markers in the dictionary-entry marker sequence and *EntrySegmentNumber* is the number of lexicographic segments of the parsed entry.

[A1]. For *s* from 1 to *EntrySegmentNumber*
If(**Segment-*s* = Sense-Segment**)
[A2]. For *i* from 0 to *MarkerNumber*
Phase_1 Sense-*i* Marker Recognition;
Phase_2 Sense-*i* Definition Parsing;
[A3]. *If*(Success)
Phase_3 Attach Parsed Sense-*i* Defi nition to Node-*i*;
Phase_4 Add Node-*i* to **Entry Sense Tree;**
[/A3]. *Else* Fail and Stop.
[/A2]. EndFor
Output: EntrySenseTree *with* **Parsed Sense Definitions**
(only if **all** sense definitions are **parsed**).
Else **Segment-*s* Parsing;**
Continue
[/A1]. EndFor
Output: Entry parsed segments, including the Sense-Segment (only if all definitions in the Sense-Segment are parsed).

Fig. 1. Standard dictionary entry parsing

The main drawback of the classical, *standard* DEP, is the embedding of the parsing cycles, [A1] [A2] [A3] ... [/A3] [/A2] [/A1], derived from the intuitive, but highly inefficient parsing strategy based on the general Depth-First searching. After presenting the SCD-based dictionary parsing method, section 3.2. compares the parsing cycles and phases of standard DEP to the ones of SCD-based DEP.

3 Parsing with SCD Configurations

The SCD *configuration*(*s*) method is a procedural, recognition-generation computational device, that is distinct from the traditional and cumbersome *formal grammars*, being able to successfully replace them for several tasks of natural language parsing, including text free parsing (Curteanu, 2006) and thesauri parsing (Curteanu et al., 2008). For SCD-based parsing, the semantics and the linguistic modeling of the text to be analyzed should be clearly specified at each parsing level, and implemented within the following components of each SCD configuration (hereafter, SCD-*config*):

- A set of *marker classes*: a *marker* is a boundary for a specific linguistic category (*e.g.* **A.**, **I.**, **1.**, **a).**, etc.). Markers are joined into *marker classes*, with respect to

their functional similarity (*e.g.* {**A.**, **B.**, **C.**, …}, {**1.**, **2.**, **3.**, …}, {**a).**, **b).**, …});

- A *hypergraph-like hierarchy* that establishes the dependencies among the marker classes;
- A *searching* (*parsing*) *algorithm*.

Once an SCD configuration is defined, parsing with the SCD configuration implies identifying the markers in the text to be parsed, constructing the *sequences* of markers and categories, recognizing the marked text structures (spans within the bounding markers) corresponding to the SCD configuration semantics, and classifying them according to the marker sequences within the pre-established hierarchy assigned to that SCD configuration. The last step settles the dependencies and correlations among the parsed textual structures. Identifying the lexicographic segments of an entry, the syntactic and semantic structure of each segment, the senses, definitions and their corresponding markers, is the result of an in-depth lexical semantics analysis. Designing the classes and the hypergraph structure of their dependencies are essential cognitive aspects of working with SCD configurations, and need to be pre-established for each dictionary.

Within the parsing process, each SCD *configuration*, *i.e.* marker classes, hierarchy, and searching algorithm, is completely commanded by its *attached* semantics. The semantically-driven parsing process, either for free or specialized texts, consists in a number of SCD configurations applied sequentially (in cascade), each one on a different semantic level. The semantic levels (each one driving an SCD configuration) are *subsuming* each other in a top-down, monotonic manner, starting with the most general semantics of the largest text span, until the most specific level.

3.1 SCD Configurations for DEP

The SCD-based process for DEP consists in three SCD *configurations*, applied sequentially on the levels and sublevels of the dictionary entry, where each level should be monotonic at the lexical-semantics *subsumption* relation.

The task of applying the SCD *configurations* to DEP requires knowing the semantics of the corresponding classes of sense and definition markers, together with their hierarchical representation.

The first SCD configuration (**SCD-*config*1**) is devoted to the task of obtaining the partition of the entry *lexicographic segments* (Hauser & Storrer, 1993). Since usually there are no dependency relations between lexicographic segments, SCD-*config*1 is not planned to establish the dependency relations (cycle **[A1]** in Fig. 1, or cycle **[B1]** in Fig. 2).

The *second* important *task* of DEP is to parse each lexicographic segment according to its specific semantics and linguistic structure. The most prominent lexicographic segment of each entry is the *sense-description one*, the central operation being the process of extracting the *sense tree* from the segment. This is the purpose of the *second* SCD *configuration* (denoted **SCD-*config*2**), corresponding exactly to the DSSD parsing algorithm in (Curteanu et al., 2008), which, for the **DLR** sense trees, has a precision of 91.18%. In order to refine the lexical-semantics of primary senses, one has to descend, under secondary senses, into the definitions and definition examples, which constitute the text spans situated between two sequentially-related nodes of the parsed sense tree. This SCD configuration is represented as cycle **[B2]** in Fig. 2.

The *third step* of DEP parsing (cycle **[B3]** in Fig. 2) is represented by the configuration **SCD-*config*3**, needed to complete the DEP. SCD-*config*3 consists in a specific set of marker classes for the *segmentation* at dictionary definitions, the hypergraph-like hierarchy of the classes of markers for these sense definitions, and the parsing algorithm to establish the dependencies among atomic senses / definitions. As a prerequisite groundwork, an *adequate modeling* of the sense definitions is needed and *the segmentation* of *definitions* is implemented as an essential step to establish the dependency-by-subsumption among the sense types of the considered thesaurus. The final result of the entry parsing process should be the sequential application of the SCD-*config*1, SCD-*config*2, and SCD-*config*3 configurations.

3.2 A Structural Analysis: Standard DEP *vs.* SCD Configurations

A pilot experiment of parsing with SCD configurations was its application to the **DLR** thesaurus parsing (Curteanu et al., 2008); the process of *sense tree building* has been completely detached

and extracted from the process of *sense definition parsing*.

The sense-tree parsing with *SCD-based* DEP cycles and phases is summarized in pseudo-code in Fig. 2 and comparative Table 1 below.

[B1]. For *s* from 1 to *EntrySegmentNumber*
Segment-*s* Parsing;
If(**Segment-*s* = Sense-Segment**)
Standby on **Sense-Segment Parsing;**
Else Continue
[/B1]. EndFor
Output: Entry parsed segments, not including the Sense-Segment;
[B2]. For *i* from 0 to *MarkerNumber*
Phase_1 Sense-*i* Marker Recognition;
Assign (Unparsed) Sense-*i* Definition to **Node-*i*;**
Phase_4 Add Node-*i* to **EntrySenseTree;**
Standby on **Sense-*i* Definition Parsing;**
[/B2]. EndFor
Output: EntrySenseTree (with **unparsed** sense definitions)**.**
Node-*k* = Root(**EntrySenseTree**);
[B3]. *While* not all nodes in **EntrySenseTree** are visited
Phase_2 Sense-*k* Definition Parsing;
If(Success)
Phase_3 Attach Parsed Sense-*k* Definition to Node-*k*;
Else **Attach Sense-*k* Parsing Result to Node-*k*;**
Node-*k* = getNextDepth FirstNode(**EntrySenseTree**)
Continue
[/B3]. EndWhile.
Output: EntrySenseTree (with **parsed** or **unparsed** definitions)**.**
Output: Entry parsed segments, including the Sense-Segment.

Fig. 2. SCD-based dictionary entry parsing

Standard DEP	*SCD-based DEP*
(Phase_1; **Phase_2** **(Phase_3;** **Phase_4)** **)**	**(Phase_1;** **Phase_4)** **(Phase_2;** **Phase_3)**

Table 1: Dictionary parsing phases in standard DEP and SCD-based DEP

Table 1 presents the ordering of the dictionary parsing phases in the *standard* DEP strategy (the four phases are embedded) and the SCD-based DEP strategy (the phases are organized in a linearly sequential order).

Since the process of *sense tree construction* (cycle **Phase_1** + **Phase_4**) has been made completely detachable from the *parsing* of *the* (atomic) *sense definitions* (cycle **Phase_2** + **Phase_3**), the whole SCD-based DEP process is much more efficient and robust. An efficiency *feature* of the SCD-based parsing technique is that, working exclusively on sense marker sequences, outputs of **[B2]** and **[B3]** cycles in Fig. 2 (*i.e.* sense trees) are obtained either the sense definition parsing process succeeds or not, either correct or not!

These properties of the new parsing method with SCD configurations have been effectively supported by the parsing experiments on large Romanian, French, and German dictionaries.

4 Romanian DLR Parsing

The study of the application of SCD-configuration to DEP started with the analysis of the **DLR** parsing (Curteanu et al., 2008). Fig. 3 presents the hierarchy of SCD-*config*2 for **DLR** sense marker classes,

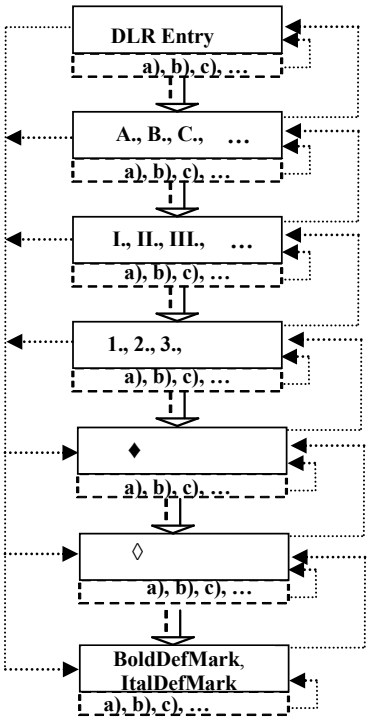


Fig. 3. Hierarchy of DLR marker classes

devoted to *sense tree parsing*. The *dashed arrows* point to the upper or lower levels of **DLR** sense marker hierarchy, from the *literal enumeration* layer-embedded level. The *continuous-dashed arrows* in Fig. 3 point downwards from the higher to the lower priority levels of **DLR** marker class hypergraph. Because of its special representation characteristics, the literal enumeration is illustrated on a layer *attached* to the hie-

rarchy level (dashed line) to which it belongs, on *each* of the sense levels.

A detailed description of the **DLR** sense tree extraction with SCD-*config*2 (denoted as DSSD algorithm) is found in (Curteanu et al., 2008).

4.1 DLR Parsing: Problems and Results

The three SCD-*configurations* establish the dependencies among **DLR** senses (SCD-*config*1-2) and definitions (SCD-*config*3). However, **DLR** is encoded by default *inheritance* rules of senses (definitions), acting on all the node levels of the sense / definition trees.

The sense tree parser (output of SCD-*config*2) was tested on more than 500 dictionary entries of large and medium sizes. The success rate was 91.18%, being computed as a perfect match between the output of the program and the gold standard. Furthermore, it is worth noting that an entry with only one incorrect parse (*i.e.* one node in the sense tree attached incorrectly) was considered to be erroneously parsed in its entirety, an approach which disregards all the other correctly attached nodes in that entry.

A first source of parsing errors is the non-monotony of the marker values: "**A.** [**B.** missing] … **C.** …"; "**2.** [instead of **1.**]... **2.** …"; "…**a)**… **b)** … **c)** … **b)** [instead of **d)**] …". Another major source of parsing errors comes from the inherent ambiguity in deciding which is the regent and which is the dependent (sub)sense in marker sequences as "**1. a) b) c)** ◊ [◊]…".

For evaluating SCD-*config*3, 52 dictionary entries of various sizes were used as a gold standard, totaling a number of approximately 2000 chunks and 22,000 words. The results are given in Table 2. Upon further analysis of the evaluation results, the most frequent errors were found to be due to faulty *sigle* (abbreviation of the source of examples) segmentation. A detailed analysis of the error types for the DLR dictionary is discussed in (Curteanu et al., 2009).

Evaluation Type	Precision	Recall	F-measure
Exact Match	84.32%	72.09%	77.73%
Overlap	92.18%	91.97%	92.07%

Table 2: Evaluation results for segmentation of DLR atomic sense elements

Correcting the acquisition of *sigles* leads to a 94.43% *f-measure* for *exact match* (the number of correctly identified sense and definition units) and a 98.01% *f-measure* for *overlap* (the number of correctly classified words in each entry). To achieve the **DLR** parsing process completely, the last operation to be executed is to establish the dependency relations between atomic senses / definitions, under all the sense nodes in the computed sense-tree of the entry. Currently, the **DLR** is parsed almost completely, including at atomic senses / definitions, the lexicographic segments and sense-trees being obtained with a correctness rate above 90% for explicitly marked sense definitions.

5 Romanian DAR Parsing

The structure of the main lexicographical segments in **DAR** is outlined below:

I. The *French Translation* segment, denoted *FreSeg*, contains the French translations of the lemma and the main sense hierarchy of the entry word. The translation of the sense structure into Romanian and the complete description of the sense tree in **DAR** are in a subsumption relation. In some cases, the French translation may not exist for specific Romanian lemmas.

II. The *general description* segment (*RomSeg*) is composed of several paragraphs and contains morphologic, syntactic, semantic, or usage information on the entry word. *RomSeg* usually starts with the entry word in italics (otherwise, the entry word occurs in the first row of the first paragraph).

III. The third segment of a **DAR** entry, called *SenseSeg*, is the *lexical-semantic description* of the entry word. *SenseSeg* is the main objective of the lexicographic analysis of the entry parsing in order to obtain its sense tree.

IV. The fourth segment of a **DAR**, *NestSeg*, contains one or more "*nests*", which are segments of text describing morphological, syntactic, phonological, regional, etc. variants of an entry, sometimes with an attached description of the specific senses. The structure of the **DAR** *nest* segment is similar to that of a typical **DAR** entry, and the recursive nature of **DAR** entries comes from the sense parsing of *nest* segments.

V. The fifth segment of **DAR** entries, denoted *EtymSeg*, contains etymological descriptions of the entry word and is introduced by an *etymology-dash* (long dash "–"). Among the five segments of a **DAR** entry, the only compulsory ones are *FreSeg* and *SenseSeg*. The other three segments are optional in the entry description, depending on each entry word.

5.1 DAR Marker Classes and Hierarchy

The priority ordering of **DAR** marker classes is:
1. Capital letters (*LatCapLett_Enum*): A., B., ...
2. Capital roman numerals (*LatCapNumb_Enum*): I., II., ...
3. Arabic numerals (*ArabNumb_Enum*): 1^0., 2^0. These markers introduce the *primary senses*, in a similar manner to those in **DLR**.
4. For introducing *secondary senses*, **DAR** uses the same sense markers used in **DLR** for definitions of type *MorfDef*, *RegDef*, *BoldDef*, *ItalDef*, *SpecDef*, *SpSpecDef*, and *DefExem*, and a set of markers specific to **DAR**: ||, |, #, †.
5. According to the level of the lexical-semantic description, **DAR** uses *literal enumeration* on *two levels*: **(5.a)** lowercase Latin letters (*LatSmallLett_Enum*): a.), b.), … **(5.b)** a *LatSmallLett_Enum* can have another enumeration, using lowercase Greek letters (*GreSmallLett_Enum*): α.), β.), γ.), …

The hierarchies for sense markers in DAR are given in Fig. 4.

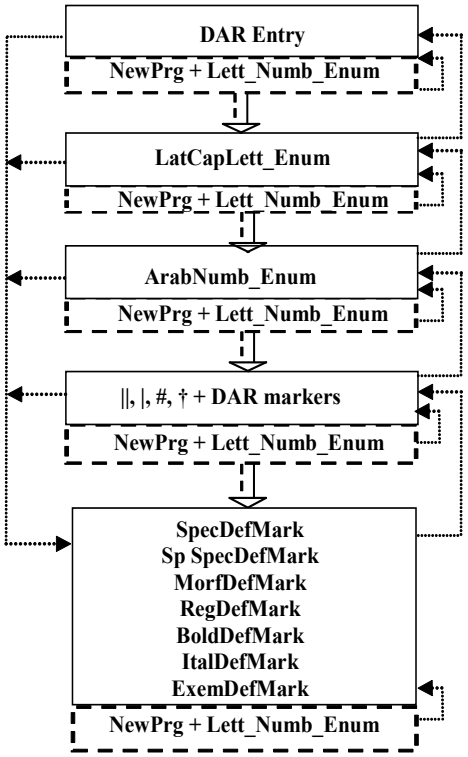
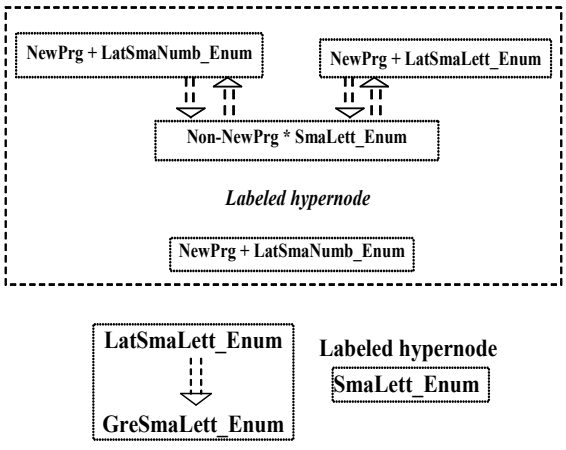


Fig. 4: Dependency hypergraph for DAR

5.2 Special problems in DAR parsing

A first difficulty for parsing the **DAR** lexicographic segments is the occurrence of the *New_Paragraph* (*NewPrg*). For *NewPrg* marker recognition we used an *implicit level of enumeration*, *LatSmaNum_Enum* discriminating the cases when *NewPrg* is (or not) accompanied by another type of **DAR** sense marker.



The second difficult problem in **DAR** parsing is the process of refining the *NewPrg* marker with literal enumerations (*LatSmallLett_Enum*), which can be in turn refined by an implicit enumeration using *NewPrg*. This has been solved by interpreting the sense levels for enumerations according to their context.

Using SCD configurations, we have parsed 37 **DAR** entries of large and medium sizes. The results of the **DAR** parsing have been evaluated manually, and we estimated the parsing precision as really promising, taking into account the difficult problems raised by **DAR**. The lack of a gold standard for **DAR** entries did not allow performing an automatic evaluation at this moment.

DAR Entry Parsing (Excerpt):

```
- <entry>
- <lexsegm value="FreSeg." class="0">
- <sense value="LARG, -Ă" class="1">
  <definition>adj., s. a. şi f. I. 1°. {i}Large, vaste.{/i} 2°. (Fig.)
{i}Large, ample, majestueux. Largement. {/i}3°. {i}Au large, à
…{i}Femme légère, dessalée{/i}.</definition>
- </sense >
- </lexsegm >
```

```
- <segm value="SenseSeg." class="0">
- <sense value="I." class="8">
- <definition>  A d j. şi a d v. </definition>
- <sense value="1°." class="12">
- <definition>
  A d j. (În opoziţie cu î n g u s t) Extins în toate direcţiile;
…{i}Larg {/i}{i}= {/i}largus.
  <SRCCITE source="ANON. CAR.">ANON.
CAR.</SRCCITE>  {i}Calea ceaia largă.{/i}
  <AUTHCITE source="EV." author="CORESI" sigla="CORESI,
EV." …
  </definition>
</sense>
- <sense value="2°." class="12">
- <definition>  F i g. (Literar, după fr.) Mare, amplu, …
  <AUTHCITE source="C. I." volume="II" …</AUTHCITE> …
  </definition>
- <sense value="‖" class="20">
  <definition>Adv. {i}Musafirul... se urca …</definition>
  </sense>
  </sense>
- <sense value="3°." class="12">
- <definition>  (În funcţiune predicativă,…)
…  …  …
</definition>
  </sense>
  </sense>
- <sense value="II." class="8">
- <definition>  S u b s t. </definition>
- <sense value="1°. " class="12">
- <definition>  S. a. Lărgime. {b}Inimii închise...  </definition>
  </sense>
…  …  …
- <sense value="2°." class="12">
- <sense value=" NewPrg " class="13">
  <definition>{i}LĂRGÍME{/i} s f. v. {b}larg{/b}.</definition>
  </sense>
- <sense value="NewPrg" class="13">
  <definition>{i}LĂRGĂMÂNT{/i} † S. A. V.
{i}larg{/i}.</definition>
…  …  …
  </sense>
  </sense>
  </lexsegm>
  </entry>
```

6 French TLF Parsing

The French **TLF**, a very well-organized and structured thesaurus, provides both similarities and distinctive characteristics in comparison to the Romanian **DLR**. The structure of **TLF** lexicographic segments, obtained with the SCD-*config*1, is relatively simple. A **TLF** entry commences with the *sense-description segment*, optionally (but very frequently) followed by a package of "*final*" *segments*, introduced by specific labels, as in the pattern:

REM. 1. … **2.** … **3.** …
PRONONC. ET ORTH. – … Eng.: …
ÉTYMOL. ET HIST. I. ... **1. a)** … **b)** … **2.** …
3. … **II.** …
STAT. Fréq. abs. littér.: … **Fréq. rel. littér.:** …
DÉR. 1. …**2.** …**3. a)** … **b)** … **Rem. a)** … **b)** …
BBG. …

As one may notice, some *final segments* can enclose particular sense descriptions, similarly to those met in the proper sense-description segment. The sense markers in **TLF** resemble to those in **DLR**, but there are also significant differences. The dependency hypergraph of the **TLF** marker classes is the following:

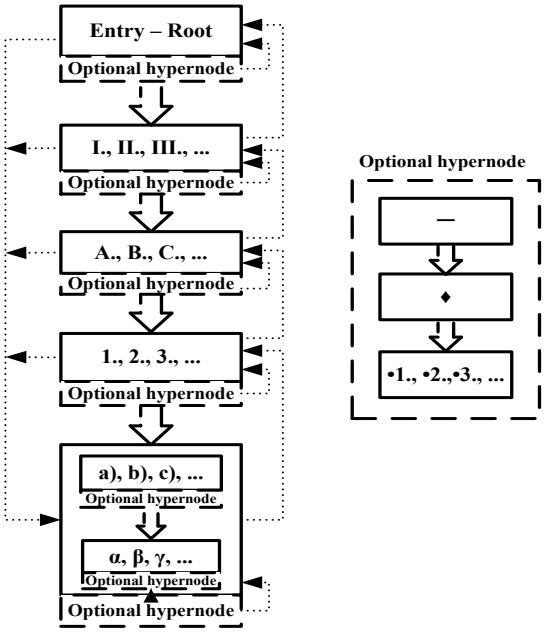


Fig. 5. Dependency hypergraph of TLF sense marker classes

Cross-linguistic hints involving **TLF** entry parsing with SCD configurations: (*a*) A new sense marker (compared to **DLR**) is "–" (*inheritance- dash*), aiming to signal the presence of an inherited sense. (*b*) When "–" occurs *after* another **TLF** marker, the "–" role is to inherit a parent sense (either regent or not) from the sense-tree. (*c*) When "–" begins at new paragraph (*NewPrg*), its role is of a intermediary subsense, inheriting the meaning of the regent (sub)sense. (*d*) Another new **TLF** marker is "•1., •2., …" (indexed, small red-ball), defining the *new* **TLF** sense concept: *Indexed Examples to Definitions* for the whole entry (denoted *IdxDefExem*). (*e*) The literal enumeration with Latin small letters (*LatSmaLett_Enum*) is refined with Greek small letters (*GreSmaLett_Enum*). (*f*) In **TLF**, only the *filled diamond* "♦" marker is present (as secondary sense); the *empty diamond* "◊" is missing. (*g*) Some primary senses ("**I.**", "**A.**") in

TLF receive reversed priorities (Fig. 5) in the marker class hierarchy as compared to **DLR**.

6.1 TLF Parsing Results

For **TLF**, we processed 31 significant entries (TLFi, 2010) of medium and large dimensions (entries of 5 pages each, A4 format, in average) with the parser based on SCD configurations. The parsing results have been evaluated manually, the correctness rate being above 90%. One of the reasons of these very promising results for **TLF** parsing may be the regularity and standardization of the **TLF** entry texts. An automatic, precise evaluation of the SCD-based parser was not possible since we are missing currently a gold-corpus of **TLF** entries.

TLF Entry Parsing (Excerpt):

```
- <entry>
- <lexsegm value="SenseSeg." class="0">
- <sense value="ANNONCER" class="1">
+ <definition> - <sense value="I." class="2">
- <definition>  <i>Emploi trans.</i> … …
  </definition>
- <sense value="A." class="3">
  <definition>[Le suj. désigne une pers.]</definition>
- <sense value="1." class="4">
- <definition>
  [L'obj. désigne un événement] Faire connaître …
  </definition>
- <sense value="a)" class="5">
- <definition>
  [L'événement concerne la vie quotidienne] …
  <i>Annoncer qqc. à qqn, annoncer une bonne</i> … …
  </definition>
- <sense value="circle" class="10">
- <definition>
  1. À la mi-novembre, Costals …
  <b>annonça</b>  son retour pour le 25. Dans la lettre …
  </definition>
  </sense>
- <sense value="circle" class="10">
- <definition>
  2. Électre, fille d'un père puissant, réduite…
  <b>annonce</b> … … …
  </definition>
  </sense>
  </sense>
- <sense value="b)" class="5">
- <definition>
  <i>JEUX (de cartes). Faire une annonce.</i> …
  </definition>
- <sense value="circle" class="10">
- <definition>
  3. Celui qui la détient la belote …
  <b>annonce</b>  alors :  <i>belote,</i>
… …
  </definition>
  </sense>
  </sense>
… … …
  </lexsegm>
- <lexsegm value="FinSeg." class="0">
- <sense value="-" class="5">
- <definition>  <b>ÉTYMOL. ET HIST.</b> … …
  <i>Ca</i>  1080  <i>anuncier</i>
… … …
  </definition>
  </sense>
…. … …
- <definition>
  <b>BBG.</b>  ALLMEN 1956. BRUANT 1901. …
  <b>ARRIVÉE, subst. fém.</b>
  </definition>
… … …
  </sense>
  </lexsegm>
  </entry>
```

7 Lexicographic Segments and Sense Markers in German DWB and GWB

The German **DWB** entries comprise a complex structure of the lexicographic segments, which provide a non-uniform and non-unitary composition (Das Woerterbuch-Netz, 2010). One special feature is that **DWB** and **GWB** lexicographic segments are composed of two parts: a first (optional) *root-sense* subsegment, and the segment *body*, which contain the explicit sense markers, easily recognizable. For **DWB**, the parsing of lexicographic segments is not at all a comfortable task since they are defined by three distinct means:

(A) *After* the *root-sense* of a **DWB** entry, or *after* the *root-sense* of a lexicographic segment, (a list of) italicized-and-spaced key-words are placed to constitute the *label* of the lexicographic segment *that follows*. Samples of such key-word labels for **DWB** lexicographic segments are: "*Form, Ausbildung und Ursprung*", "*Formen*", "*Ableitungen*", "*Verwandtschaft*","*Verwandtschaft und Form*", "*Formelles und Etymologisches*", "*Gebrauch*", "*Herkunft*", "*Grammatisches*", etc., or, for **DWB** sense-description segment: "*Bedeutung und Gebrauch*" (or just "*Bedeutung*"). In the example below, they are marked in grey.

GRUND, *m., dialektisch auch f. gemeingerm. wort; fraglich ist … poln. russ. slov. nlaus.* grunt *m.* *f o r m u n d h e r k u n f t* .

1) *für das verständnis der vorgeschichte des wortes ist die z w i e g e s c h l e c h t i g k e i t …*

H. V. SACHSENHEIM *spiegel* 177, 30;

städtechron. 3, 51, 14. … … … drey starcke grund 6, 290. *b e d e u t u n g* . *die bedeutungsgeschichte des wortes … …*

I. grund *bezeichnet die feste untere begrenzung eines dinges.*

A. *grund von gewässern; seit ältester zeit belegbar: profundum* (*sc. mare*) crunt *ahd. gl.* 1, 232, 18;

1) *am häufigsten vom meer* (*in übereinstimmung mit dem anord. gebrauch*): *...*

(B) *The second* way to specify the **DWB** current lexicographic segments is to use their labels as key-words immediately *after* the primary sense markers.

(C) The *third* (and most frequent) way to identify the lexical description segment(s) of a **DWB** entry is simply the lack of a segment label at the beginning of the sense description segment. By default, after the entry *root-sense segment* (which can be reduced to the Latin translation of the German lemma) the sense-description segment comes, without any "*Bedeutung*" label, introducing explicit sense markers and definitions.

7.1 German DWB and GWB Dependency Hypergraphs. Parsing Results

Without coming into details (see the marker class dependency hypergraphs in Fig.6 and Fig.7), one can say with a good measure of truth that a general resemblance hold between **DAR**

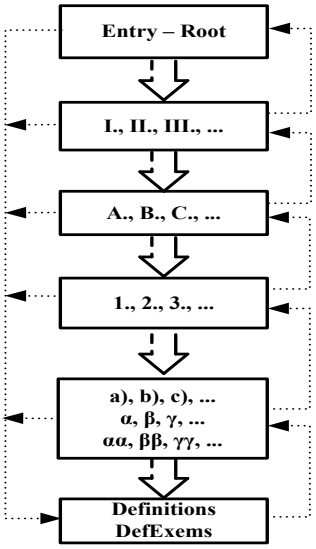


Fig. 6. DWB dependency hypergraph

and **DWB**, and **TLF** and **GWB**, respectively. The sense markers in **DWB** are usual, with the remark that sense refinement by literal enumeration is realized on three levels: *LatSmaLett_Enum* (**a)**, **b)**, …), *GreSmaLett_Enum* (**α)**, **β)**, …), and *GreDoubleSmaLett_Enum* (**αα)**, **ββ)**, …).

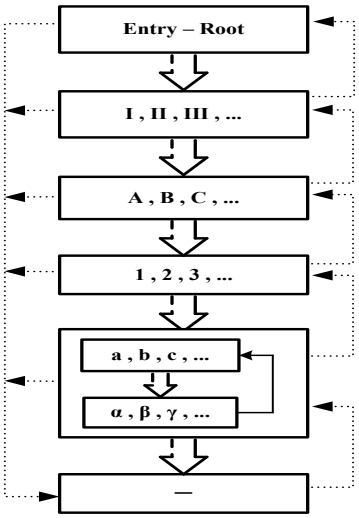


Fig. 7. GWB dependency hypergraph

A number of 17 very large **DWB** entries have been parsed only with SCD-*config*1 and SCD-*config*2. We appreciate on this small but significant excerpt of **DWB** entries that parsing of the sense description segment at sense trees is performed with a high precision, but delimitation of the lexicographic (sub-)segments and labels is a more difficult problem. The lack of a **DWB** entry gold corpus did not allow a precise, automated evaluation of the parser.

8 Directions: Optimal Lexicon Design, Standardization, Lexicon Networks

The special features of parsing with **SCD configurations** (SCD-*configs*) are: • SCD-configs is a completely *formal grammar-free* approach which involves simple, efficient (weeks-time adaptable), thus portable modeling and programs. • In all currently existing DEP methods, the sense tree construction of each entry is, more or less, recursively embedded and mixed within the definition parsing procedures. • SCD-*configs* provides a lexical-semantics refinement level on each SCD-*config*. • SCD-*configs* separate and run sequentially, on independent levels (*viz.* configurations), the processes of lexicographic segment recognition, sense tree extraction, and atomic definition parsing. • This makes the whole DEP process with SCD-*configs* to be *optimal*. • The sense marker classes and their dependency hypergraphs, specific to each thesaurus, offer an unique instrument of lexicon con-

struction comparison, sense concept design and standardization. With the SCD parsing technique, one can easily compare the sense categories, their marking devices, the complexity and recursiveness measure of the sense dependency hypergraphs for each thesaurus.

The cross-linguistic analysis of the five large thesauri showed the necessity of a careful lexical-semantics modeling of each dictionary. Equally important, many semantic and lexicographic concepts such as sense markers and definitions, (indexed) examples to definitions, sense and source references etc. can be similar, adaptable, and transferable between corresponding SCD-configurations of different thesauri.

The SCD-*configs* analysis pointed out the need of a more general and adequate terminology for the lexical-semantics notions. *E.g.*, comparing the Romanian and French thesauri with the German ones, we decided that, while preserving the definition type labels *MorfDef*, *DefExem*, *SpecDef* and *SpSpecDef*, we should change the *RegDef* into *GlossDef*, *BoldDef* into *IdiomDef*, *ItalDef* into *CollocDef*, and add the **TLF** *IdxDefExem* (an indexed *DefExem*) to the sense concept set.

The future experiments will continue with new thesauri parsing: Russian, Spanish, Italian, but the true challenge shall be oriented towards Chinese / Japanese thesauri, aiming to establish a thorough lexical-semantics comparison and a language-independent, portable DEP technology based on SCD configurations. A further development would be to align the Romanian thesauri sense and definition types to TEI P5 standards (XCES, 2007), and to design an optimal and cross-linguistic compatible *network of Romanian electronic dictionaries*, similar to a very good project of dictionary network, *i.e.* the German Woerterbuch-Netz (with links to **TLFi** entries too), whose twelve component lexicons include **DWB** and **GWB.**

Acknowledgement. The present research was partly financed within the **eDTLR** grant, PNCDI II Project No. 91_013/18.09.2007.

References

DLR revision committee. (1952). *Coding rules for DLR* (in Romanian). Romanian Academy, Institute of Philology, Bucharest.

Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., Dănilă, E. (2007). *The Digital Form of the Thesaurus Dictionary of the Romanian Language*. In Proc. of the 4th SpeD 2007.

Curteanu, N., and E. Amihăesei. (2004). *Grammar-based Java Parsers for DEX and DTLR Romanian Dictionaries*. ECIT-2004, Iasi, Romania.

Curteanu, N. (2006). *Local and Global Parsing with Functional* (F)X-*bar Theory and SCD Linguistic Strategy*. (I.+II.), Computer Science Journal of Moldova, Academy of Science of Moldova, Vol. 14 no. 1 (40):74-102; no. 2 (41):155-182.

Curteanu, N., D. Trandabăţ, A. M. Moruz. (2008). *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*, Proceedings of CogAlex Workshop, COLING 2008, ISBN 978-1-905593-56-9, :55-63.

Curteanu, N., Moruz, A., Trandabăţ, D., Bolea, C., Spătaru, M., Husarciuc, M. (2009). *Sense tree parsing and definition segmentation in eDTLR Thesaurus*, in Trandabăţ et al. (Eds.), Proc. of the Workshop "Linguistic Resources and Instruments for Romanian Language Processing", Iasi, Romania, "Al.I.Cuza" University Publishing House, ISSN 1843-911X, pp. 65-74, (in Romanian).

Das Woerterbuch-Netz (2010): http://germazope.uni-trier.de/Projects/WBB/woerterbuecher/

Hauser, R., and A. Storrer. (1993). *Dictionary Entry Parsing Using the LexParse System*. Lexikographica (9): 174-219.

Kammerer, M. (2000). *Wöterbuchparsing Grundsätzliche Überlegungen und ein Kurzbericht über praktische Erfahrungen*, http://www.matthias-kammerer.de/content/WBParsing.pdf

Le Trésor de la Langue Française informatisé (2010). http://atilf.atilf.fr/tlf.htm

Lemnitzer, L., and C. Kunze. (2005). *Dictionary Entry Parsing*, ESSLLI 2005.

Neff, M., and B. Boguraev. (1989). *Dictionaries, Dictionary Grammars and Dictionary Entry Parsing*, Proc. of the 27th ACL Vancouver, British Columbia, Canada, :91 – 101.

Tufiş, Dan. (2001). From Machine Readable Dictionaries to Lexical Databases, RACAI, Romanian Academy, Bucharest, Romania.

XCES TEI Standard, Variant P5. (2007). http://www.tei-c.org/Guidelines/P5/

Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: a Case Study for Italian

Elisa Lavagnino
LCI - Télécom Bretagne &
CeRTeM - Università degli Studi di Genova
elisa.lavagnino
@telecom-bretagne.eu

Jungyeul Park
LINA
Université de Nantes
jungyeul.park
@univ-nantes.fr

Abstract

This paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages. At present, we mainly focus on English and the major Romance languages such as French, Spanish, Portuguese, and Italian. This paper is a case study for Italian language. We present how to build automatically conceptual structure of automatically extracted multi-word terms from domain specific corpora for Italian. We show the experimental results for extracting multi-word terms from two domain corpora ("natural area" and "organic agriculture"). Since this work is still ongoing, we discuss our future direction at the end of the paper.

1 Introduction

Great progress has been recently obtained on using text analysis to extract terms in a specific field. The study of texts helps in finding and organizing textual segments representing conceptual units. A corpus is a collection of texts stored in an electronic database. Texts have been selected to be representative of a particular goal. A corpus must be balanced in quality and quantity contents: in order to be representative of a domain, texts have to cover all the possible communicative situations. Generally, in a specialised domain, users share contents and they normally can understand and communicate with each others without ambiguities. However, when different communities get in touch the possibility of misunderstanding arises because of terminological variation. This variation can be detected at a conceptual level or at the formal one. Our approach tries to overcome this problem by collecting different text typologies. Texts may be extracted from different sources which can be classified as their specialisation level, their contents, their pragmatic application, etc. In our case, we are interested in using different texts, in order to analysis the result of automatic extraction in different communicative situations to improve its functioning.

A term can be simple if composed by one word, or complex if composed by several words. This paper focuses on extracting and conceptually structuring multi-word terms for Italian. Collet (2000) affirmed that a complex term (multi-word term in our terminology) is a complex unit whose components are separated by a space and are syntactically connected. The resulting unit denominates a concept which belongs to the language for special purposes (LSP). Texts on any domain are easily available on the Web these days. To create a corpus representing a field, materials should be, however analysed and re-elaborated in order to resolve eventual problems arising the transfer of data. In particular, a corpus have to be processed in order to classify the composing units. This classification represents the first step towards terminological extraction. Terminologists must often look through many texts before finding appropriate ones (Agbago and Barrire, 2005). L'Homme (2004) presents guidelines for choosing terminology such as domain specificity, language originality, specialization level, type, date, data

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 48–55,
Beijing, August 2010

evaluation.[1]

Since interaction between domains increases consistently, domain specificity is a crucial point to considerer during the creation of a corpus. Text typologies and communicative situations reflect their peculiarity to terms. A concept can be represented differently if the level of specialisation of a text or the context changes. Here, we consider the context as the frame in which the communication takes place. For example, the domain of "natural area", Italian language is really interesting because terms register a high level of variations due to the different contexts.

The LSP changes as the society evolves. Terms can register the diachronic variation due to the development of a certain domain. The evolution of a domain influences also the terminologies which form LSP. Terminological evolution generates variations in the conceptual representation which should be observed in order to detect terms and their variants and to establish relations between them. For example, the domain of "organic agriculture" is now evolving and changing because of political choices. This affects the terminology and the eventual creation of new forms. The affix *bio-* which can be use as a variant of almost all multi-word terms concerning the biological production such as *metodo di produzione biologica* ('method of organic production') becomes *metodo bio* and *prodotto biologico* ('organic product') becomes *prodotto bio* or just *bio*.

In this paper, we present an approach for extracting automatically multi-word terms (MWT) from domain specific corpora for Italian. We also try to conceptually structure them, that is we build the 'conceptual' structure of variations of multi-word terms where we can learn dynamics of terms (Daille, 2002). Conceptual structure in this paper limits to the semantic relationships between terms such as **Hyperonomy**, **Antony**, **Set of**, and **Result** between multi-word terms and we currently implement only hyperonomy relations.

Actually, this paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages. At present, we mainly focus on English and the major Romance languages such as French, Spanish, Portuguese, and Italian. This paper is a case study for Italian language. The remaining of this paper is organized as follows: We explain how to automatically extract and conceptually structure multi-word terms from domain specific corpora in the next section. We also describe some implementation issues and current advancement. Since this work is still on-going, we discuss our future direction in the last section.

2 Automatically Extracting and Conceptually Structuring Multi-Word Terms

2.1 ACABIT

To extract automatically multi-word terms from domain specific corpora and conceptually structure them for Italian, we adapt existing ACABIT which is a general purpose term extractor. It takes as input a linguistically annotated corpus and proposes as output a list of multi-word term candidates ranked from the most representative of the corpus to the least using the log-likelihood estimation.[2] ACABIT is currently available for English and French as different programs for each language. Fundamentally, ACABIT works as two stages: `stat` and `tri`. At the `stat`, it allows us to identify multi-word terms in corpora to calculate the statistic. At the `tri`, it allows us to sort and conceptually structure them based on base terms. For the moment, we reimplement universal `stat` for major Romance languages. We explain the more detailed issues of our reimplementation of ACABIT for Italian in Section 2.3.

2.2 Base Term and its Variations

For automatic multi-word term identification, it is necessary to define first the syntactic structures which are potentially lexicalisable (Daille, 2003). We refer to these complex sequences as **base terms**. For Italian, the syntactic structure of base terms is as follows (where Noun_1 is a head):

[1] The translated text is adapted from Agbago and Barrire (2005)

[2] `http://www.bdaille.fr`

Noun$_1$ Adj *area protetta* ('protected area'), *azienda agricola* ('agricultural company')

Noun$_1$ Noun$_2$ *zona tampone* ('buffer area')

Noun$_1$ *di* (Det) Noun$_2$ *sistema di controllo* ('control system'), *conservazione dei biotopi* ('biotope conservation')

Besides these base term structures, there is also [Noun$_1$ *à* V$_{inf}$] for example for French. For Italian, there might be [Noun$_1$ *da* V$_{inf}$] such as *prodotto biologico da esportare* ('organic product to export') which is rather phraseology and not a term. Consequently, we define only three base term structures for Italian for now.

ACABIT for Italian should spot variations of base terms and puts them together. For example, there are **graphical variations** such as case differences and the presence of an optional hyphen inside of base term structures, **inflexional variations** where *aree protette* ('protected areas') should be considered as the variation of *area protetta* ('protected area'), or **shallow syntactic variations** which only modifies function words of the base terms, such as optional character of the preposition and article such as *sistema di informazione* and *sistema informativo* ('information system').

To conceptually structure identified multi-word terms, ACABIT for Italian should put together syntactic variations which modify the internal structure of the base term: internal modification and coordination. Internal modification variations introduce the modifier such as the adjective in [Noun$_1$ *di* Noun$_2$] structure or a nominal specifier inside of [Noun$_1$ Adj] structure. For example, *qualità ambientale* ('environmental quality') and *elevata qualità ambientale* ('high environmental quality') for [Noun$_1$ *di* Noun$_2$] structure and *ingrediente biologico* ('organic ingredient') and *ingrediente d'origine biologico* ('organic origin ingredient') for [Noun$_1$ Adj] structure . Coordination variations coordinate or enumerate the base term structure, for example *habitat ntaurali* ('natural habitat') and *habitat naturali e quasi naturali* ('natural and almost natural habitat')

2.3 Implementation

To keep consistent with the original ACABIT and to take an advantage of by directly using a certain part of existing modules, we use the input and the output formats of ACABIT. The input format of ACABIT requires the lemmatized forms of words for detecting inflexional variations of multi-word terms. For example, putting together inflexional variations such as *area protetta* and *aree protette* ('protected area(s)') is easily predictable by using their lemmatized forms. The original version of ACABIT for French uses BRILL's POS tagger[3] for POS tagging and FLEMM[4] for restoring morpho-syntactic information and lemmatized forms. And for English, it uses BRILL's POS tagger and CELEX lexical database[5] as a lemmatiser.

Since we are reimplementing ACABIT for multiple languages and we want to use the homogeneous preprocessing for ACABIT, we use TREETAGGER[6] which annotates both of part-of-speech tags and lemma information as preprocessor for . Moreover, TREETAGGER is available for several languages. We, then adapt the result of TREETAGGER for the input format for ACABIT. We use French POS tagger's tagset (Étiquettes de Brill94 Français INALF/CNRS) for every language, we convert TREETAGGER tagset into BRILL's tagset.[7]

Figure 1 shows the example of the input format of ACABIT in XML makes use of which conforms to Document Type Definition (DTD) in Figure 2. In Figure 1, POS tags are followed by morpho-syntactic information and the lemmatized form of a token in each <PH>.[8] TREETAGGER provide only lemmatized forms with POS information, instead of providing its main

[3] http://www.atilf.fr

[4] http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm

[5] http://www.ldc.upenn.edu/

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] http://www.lirmm.fr/~mroche/Enseignements/FdD_M2P_old/Etiqueteur/tags.html#francais_inalf

[8] For the convenience of the notations, accented characters are sometimes presented as 'e and 'a for è and à, respectively in the Figure.

morphological features such as gender, number, person and case as FLEMM in the previous version of ACABIT. We simply introduce dummy morphological information because it is not actually used in ACABIT. Note that è/SYM/è in Figure 1 is not correctly POS-tagged by TREE-TAGGER. It is one of flexional forms of *essere* ('be') instead of the symbol (SYM). However, we do not perform any post-processing to correct errors and we leave it as it is analyzed for the moment.

In Figure 2, <CORPUS> is for the name of the corpus, <RECORD> is for different texts which are usually from separate files, <INFO> has a format like <INFO>00/CAR/00 -/-0001800/SBC/0001800</INFO> with the year of text creation 00 and the file identification 0001800. <TITLE> is for the title, <AB> is for the text body, and <PH NB="*num*"> is for sentence identification.

ACABIT proposes as output a list of multi-word terms ranked from the most representative of the corpus using log-likelihood estimation (Dunning, 1993) and their variations in the corpus. It also shows the semantic relation between multi-word terms. The example of the output is given in Figure 3. A base term, for example *area protetto* ('protected area') is put together with its syntactic variations "*area naturale protetto* ('natural protected area') and *area marino protetto* ('marine protected area'). We can rewrite them as **Hyperonomy** (*area naturale protetto*) = *area protetto*" or **Hyperonomy** (*area marino protetto*) = *area protetto* because ACABIT identifies that *area protetto* is a hypernym of *area naturale protetto* and *area marino protetto* as <MODIF>ied terms of <BASE> terms . Likewise, a base term *prodotto biologico* ('organic product') has its syntactic variation: internal modification such as *prodotto non biologico* ('non-organic product'), *prodotto alimentare non biologico* ('non-organic alimentary product'), and *prodotto ittico biologico* ('organic fishing product'), and coordination like *prodotto biologico e non biologico* ('organic and non-organic product'). Moreover, there are **Antonym** relation described as LINK type="Neg" between the base terms and some of its syntactic variations such as *prodotto non biologico* and *prodotto alimentare non biologico*. Note that output of ACABIT in Figure 3 only contains canonical forms of multi-word terms.

2.4 Experiments

Creation of domain specific corpora: For our experiments we crawl two domain corpora of "natural area" domain which consists of 17,291 sentences and 543,790 tokens from *Gli E-Quaderni*[9] and 47,887 sentences and 1,857,914 tokens from *Parchi*[10]. We also crawl in the Internet to create the corpus of "organic agriculture" which consists of 5,553 sentences and 150,246 tokens from *National legislations* and *European legislations* for organic agriculture[11].

Automatic evaluation: Table 1 shows the statistics of experimental results from each domain. Since our domain corpora are mutually related, we count the common multi-word terms and there are 600 unique terms (base terms + variations) shared in both corpora. This is 18.74% of the number of terms in "organic agriculture". Figure 4 shows example of these common terms.

2.5 Current advancement

Till now, we reimplement only stat for multiple languages. To conceptually struture them, we still borrow tri of the previous ACABIT. We have not implemented yet full features of stat for Italian neither because of the lack of morpho-syntactic rules.

For example, the preposition inside of the term of [Noun$_1$ *di* Noun$_2$] structure might be equivalent to a prefix-added Noun$_2$ such as *deterioramento dopo la raccolta* ('rot after harvest') vs. *deterioramento post-raccolta* ('post-harvesting rot'). Likewise, the morphological derivation

[9] http://www.parks.it/ilgiornaledei parchi/e-quaderni-federparchi.html

[10] http://www.parks.it/federparchi /rivista/

[11] http://www.sinab.it/index.php?mod =normative_politiche&smod=comunitarie &m2id=189&navId=196 and http://www.sinab.it/index.php?mod=normative_politiche &smod=nazionali&m2id=189&navId=197, respectively.

```
<?xml version="1.0" encoding="UTF-8"?>
<CORPUS>
<RECORD>
<INFO>00/CAR/00 -/- 0001800/SBC/0001800</INFO>
<TITLE> </TITLE>
<AB>
<PH NB="0"> La/DTN:_:s/la presente/ADJ:_:p/presente Ricerca/SBP:_:s/Ricerca
'e/SYM/'e frutto/SBC:_:s/frutto di/PREP/di un/DTN:_:s/un lavoro/SBC:_:s/lavoro
realizzato/ADJ2PAR:_:s/realizzare da/PREP/da una/DTN:_:s/una
pluralit'a/ADJ:_:s/pluralit'a di/PREP/di soggetti/SBC:_:p/soggetto -/SYM/-
pubblici/ADJ:_:p/pubblico ,/, privati/ADJ:_:p/privato ,/, del/DTN:_:s/del
mondo/SBC:_:s/mondo della/DTN:_:s/della ricerca/SBC:_:s/ricerca e/COO/e
dell'/DTN:_:s/dell' associazionismo/SBC:_:s/associazionismo -/SYM/-
sul/DTN:_:s/sul tema/SBC:_:s/tema agricoltura/SBC:_:s/agricoltura ,/,
ambiente/SBC:_:p/ambiente ,/, aree/SBC:_:p/area protette/ADJ:_:p/protetto ,/,
occupazione/SBC:_:p/occupazione ./.
</PH>
...
</AB>
</RECORD>

<RECORD>
...
</RECORD>
</CORPUS>
```

Figure 1: Example of the input of ACABIT

```
<!ELEMENT CORPUS (RECORD)*>
<!ELEMENT RECORD (DATE?, TITLE?, INFO?, AB)>
<!ELEMENT DATE (#PCDATA)>
<!ELEMENT INFO (#PCDATA)>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT AB (PH)*>
<!ELEMENT PH (#PCDATA)>
<!ATTLIST PH NB CDATA #IMPLIED>
```

Figure 2: DTD for the input format of ACABIT

Domain	Total # of extracted multi-word terms	Unique # of terms (base terms + variations)	Unique # of terms (base terms + variations) without hapax
"Natural Area"	34,665	21,119 (16,182+4,937)	4,131 (3,724+407)
	120,633	63,244 (46,421+16,823)	12,674 (11,481+1,193)
"Organic Agriculture"	10,071	3,201 (2,509+692)	1,737 (1,431+306)

Table 1: Experimental results

```
<?xml version="1.0" encoding="UTF-8"?>
<LISTCAND>
...
<SETCAND new_ident="3" loglike="4839.794" freq="183">
<LINK type="Neg" old_ident1="3" old_ident2="3_0"></LINK>
<LINK type="Neg" old_ident1="3" old_ident2="3_1"></LINK>
        <CAND old_ident="3_0">
        <NA freq="38">
                <MODIF> <TERM> prodotto non biologico </TERM>
                </MODIF>
        </NA>
        </CAND>
        <CAND old_ident="3_1">
        <NA freq="4">
                <MODIF> <TERM> prodotto alimentare non biologico </TERM>
                </MODIF>
        </NA>
        </CAND>
        <CAND old_ident="3">
        <NA freq="2">
                <COORD> <TERM> prodotto biologico e non biologico </TERM>
                </COORD>
        </NA>
        <NA freq="1">
                <MODIF> <TERM> prodotto ittico biologico </TERM>
                </MODIF>
        </NA>
        <NA freq="138">
                <BASE> <TERM> prodotto biologico </TERM>
                </BASE>
        </NA>
        </CAND>
</SETCAND>
...
<SETCAND new_ident="6" loglike="6757.769" freq="260">
    <CAND old_ident="6">
    <NA freq="234">
        <BASE> <TERM> area protetto </TERM>
        </BASE>
    </NA>
    <NA freq="23">
        <MODIF> <TERM> area naturale protetto </TERM>
        </MODIF>
    </NA>
    <NA freq="3">
        <MODIF> <TERM> area marino protetto </TERM>
        </MODIF>
    </NA>
    </CAND>
</SETCAND>
<SETCAND new_ident="881" loglike="1855.26" freq="39">
    <CAND old_ident="881">
    <NA freq="39">
        <BASE> <TERM> pratica agricolo </TERM>
        </BASE>
    </NA>
    </CAND>
</SETCAND>
...
</LISTCAND>
```

Figure 3: Example of the output

attivitá economiche sostenibili (‘economical sustainable activity’)
conservazione del paesaggio (‘landscape preservation’)
danno ambientale (‘environmental damage’)
elemento naturalistico (‘naturalistic element’)
equilibrio naturale (‘natural equilibrium’)
denominazione d’origine protetta (‘protected origin denomination’)
denominazione d’origine controllata (‘controlled origin denomination’)

Figure 4: Example of common terms shared in both “natural area” and “organic agriculture”

of Noun_2 in [Noun_1 *di* Noun_2] structure might imply a relational adjective such as *acidità del sangue* (‘acidity of the blood’) vs. *acidità sanguigna* (‘blood acidity’). Figure 5 shows examples of rules of morpho-syntatic variations between noun and adjectival endings for Italian, which they are independently provided as external properties file for Italian. In Figure 5, endings `-zione` (nominal) and `-tivo` (adjectival) mean that if there are adjective ended with `-tivo` like *affermativo*, the system searches for the morphological derivation of a noun ended with `-zione` like *affermazione* and put them together. Only partial rules of morpho-syntatic variations for Italian are presently integrated. We try to find the exhaustive list in near future.

3 Discussion, Conclusion and Future Work

In general, manual retrieval and validation of terms is labor intensive and time consuming. The automatic or semi-automatic methods which works on text in order to detect single or multi-word terms relevant to a subject field is referred to as term extraction. Term extraction produces the raw material for terminology databases. It is a process which is likely to produce significant benefits in terms individuation. The reasons which justify term extractions are:

1. building glossaries, thesauri, terminological dictionaries, and knowledge bases; automatic indexing; machine translation; and corpus analysis rapidly.

2. Indexing to automatize information retrieval or document retrieval.

3. Finding neologism and new concepts.

Term extraction systems are usually categorized into two groups. The first group is represented by the linguistically-based or rule-based approaches use linguistic information such as POS and chunk information to detect stop words and to select candidate terms to predefined syntactic patterns. The second group is represented by the statistical corpus-based approaches select n-gram sequences as candidate terms. The terms are selected by applying statistical measures. Recently, these two approach are combined.

We implement ACABIT for Italian, which uses the combined method to extract multi-word terms and structure them automatically. We introduce base term structures and their linguistic variation such as graphical, inflexional, and shallow syntactic variations. We also consider the modification of the structure of base terms such as internal modification using adjective and coordinate variations. We evaluate on two domain specific corpora mutually related “natural area” and “organic agriculture” to extract multi-words terms and we find 600 unique terms shared in both copora. This paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages and this is a case study for Italian language. For the moment, we reimplement universal `stat` for major Romance languages. Most of previous work on extracting terms, especially for multiple languages are focusing on single-word terms and they are also often based on statistical approach with simple morphological patterns, for example Bernhard (2006), and Velupillai and Dalianis (2008).

Nominal ending	Adjectival ending	Examples
-zione	-tivo	*affermazione* ('affirmation') / *affermativo* ('affirmative')
-zione	-ante	*comunicazione* ('communication') / *comunicante* ('communicable')
-logia	-metrico	*ecologia* ('ecology') / *econometrico* ('econometric')
-gia	-gico	*enologia* ('enology') / *enologico* ('enologic')
-a	-ante	*cura* ('treat') / *curante* ('treating')
-	-bile	*cura* ('treat') / *curabile* ('treatable')
-ia	-peutico	*terapia* ('therapy') / *terapeutico* ('therapeutic')
-	-le	*vita* ('life') / *vitale* ('vital')
-	-tico	*acqua* ('water') / *acquatico* ('aquatic')

Figure 5: Example of rules of morpho-syntatic variations (noun-adjective)

Since this work is still on-going, we consider only **Hyperonomy** relations as the conceptual relation where a relative adjective modifies inside of the base term with [Noun_1 Adj] or [Noun_1 *di* Noun_2] structures. We also consider **Antonym** only with negative adverbs like *non*. There are still **Antonym** (e.g. *solubilità micellare* ('micellar solubilization') vs. *insolubilità micellare* ('micellar insolubilisation')), **Set of** (e.g. *piuma d'anatra* ('duck feather') vs. *piumaggio dell'anatra* ('duck feathers')), **Result** (e.g. *filettaggio del salmone* ('salmon filleting') vs. *filetto di salmone* ('salmon fillet')) relationships. ACABIT for French detects conceptual relations by using morphological conflating which implements stripping-recording morphological rules. We are planning to add these conceptual relationships in ACABIT for Italian in near future.

Acknowledgment

The authors would like to thank Béatrice Daille who kindly provide to us with ACABIT, for her valuable remarks on an earlier version of this paper. We also thank the four anonymous reviewer for their constructive comments.

References

Agbago, Akakpo and Caroline Barrière. 2005. Corpus Construction for Terminology. *Corpus Linguistics 2005*. Birmingham, United Kingdom, July 14-17, 2005.

Bernhard, Delphine. 2006. Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, April 3-7, 2006.

Collet, Tanja. 2000. *La réduction des unités terminologiques complexes de type syntagmatique*. Ph.D. Dissertation. Université de Montréal.

Daille, Béatrice. 2002. *Découvertes linguistiques en corpus*. Habilitation à diriger des recherches. Université de Nantes.

Daille, Béatrice. 2003. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan. July 7-12, 2003.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

L'Homme, Marie-Claude. 2004. *La terminologie : principes et techniques*, Les Presses de l'Université de Montréal.

Velupillai and Dalianis (2008).

Velupillai, Sumithra and Hercules Dalianis. 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. Manchester, United Kingdom. August 23, 2008.

Williams, Geoffrey Clive. 2003. From meaning to words and back: Corpus linguistics and specialised lexicography. *ASp* 39-40. http://asp.revues.org/1320.

Computational Lexicography: A Feature-based Approach in Designing an E-dictionary of Chinese Classifiers

Helena H. Gao
Nanyang Technological University
helenagao@ntu.edu.sg

Abstract

Chinese noun classifiers are obligatory as a category in association with nouns. Conventional dictionaries include classifiers as lexical entries but explanations given are very brief and thus hardly helpful for L2 learners. This paper presents a new design of an e-dictionary of Chinese classifiers. The design is based on both theoretical studies of Chinese classifiers and empirical studies of Chinese classifier acquisition by both children and adults. My main argument with regards to Chinese classifier acquisition is that cognitive strategies with a bottom-up approach are the key to the understanding of the complexity of classifier and noun associations. The noun-dependent semantic features of classifiers are evidence to support my argument. These features are categorically defined and stored in a separated database in an e-learning environment linked to the e-dictionary. The aim of making such a design is to provide a platform for L2 learners to explore and learn with a bottom-up approach the associations of classifiers with nouns. The computational agent-based model that automatically links noun features to that of classifiers is the technical part of the design that will be described in detail in the paper. Future development of the e-dictionary will be discussed as well.

1 Introduction

Noun classifiers are a typical feature of Chinese that distinguishes itself from many other languages. In simple terms, a classifier is a morpheme or word used to classify a noun according to its inherent semantic features. Noun classifiers in Chinese are obligatory as a category of its own and used to specify a noun when it is used with a determiner or a numeral. In other words, A Chinese classifier is never used independently. It must occur before a noun with a numeral (e.g., *yi* 'one', *liang* 'two', *san* 'three') and/or a determiner (e.g., *zhe* 'this', *nei* 'that'), or certain quantifiers (e.g., *ji* 'how many', *mei* 'every'). Such a combination is referred to as *a classifier phrase*.

However, the definition of Chinese classifiers is not a simple one. There are different types of classifiers in terms of their semantic functions. Some of them carry the unique features of the Chinese language; others are representatives of classifier languages, and yet all of them have the grammatical functions of measure words, which are a universal category in all languages. Due to the complexity of classifier functions, different definitions and classifications have been found. However, generally speaking, classifiers refer to common properties of noun objects across domains and common relations of objects in the world, rather than to categories having to do solely with language-internal relations (Lucy, 1992). Some researchers take a functional approach and define Chinese classifiers based on their grammatical functions. For example, Chao (1968) divides classifiers into nine categories. They are "classifiers or individual measures", "classifiers associated with v-o", "group measures", "partitive measures", "container measures", "temporary measures", "standard measures", "quasi-measures or autonomous measures", and "measures for verbs of action". From his classification we can see that he does not distinguish the concept of a classifier from that of a measure word. The advantage of such a classification is its inclusion of all the three types of classifiers mentioned above and being able to define them all as measure words, but the disadvantage is that those that are Chinese specific noun classifiers are all treated under the universal concept of measure words. This may be easy for learners to understand the grammatical functions of Chinese classifiers but the ontological nature of noun objects that classifiers are associated with are

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 56–65,
Beijing, August 2010

largely ignored. In recent decades, researchers have started to take a cognitive approach to understand the links between nouns and classifiers and found it necessary to make a distinction between classifiers and measure words. For instance, Tai & Wang (1990:38) state that "A classifier categorizes a class of nouns by picking out some salient perceptual properties, either physically or functionally based, which are permanently associated with entities named by the class of nouns; a measure word does not categorize but denotes the quantity of the entity named by a noun. " This definition makes a clear distinction between a classifier and a measure word, which is assumed to be helpful for L2 learners to have a better understanding of the cognitive basis of a classifier system. This is because there are no measure words in English or other European languages that can also function as classifiers in the same sense as Chinese classifiers. A recent study done by Gao (2010) has shown that Swedish adult learners of Chinese had a lower proficiency in classifier application than their general Chinese proficiency and that most of them were not aware of the difference between the concept of a classifier and that of a measure word.

Other previous studies on classifiers include descriptive and experimental studies of classifier systems of natural languages. For example, some descriptive studies make typological surveys of classifier systems in different languages (e.g. Allan, 1977; Lyons, 1977; Goddard, 1998); others provide semantic analysis of classifiers and their associated nouns (e.g. Downing, 1993; Huang & Ahrens, 2003; Matsumoto, 1993), and some also propose that there is an ontological base on which classifiers and nouns are associated with (Sowa 2000; Philpot et al., 2003; Nichols et al., 2005).

Experimental studies using computer technology to apply findings of classifier knowledge to natural language processing (NLP) have provided a new approach for the semantic analysis of classifiers (e.g. Nirenburg & Raskin, 2004; Hwang et al., 2007, Quek, 2010) and for computer-assisted language learning (e.g. Guo & Zhong, 2005). However, no e-learning systems developed so far are found to be able to guide second language learners to use the semantic properties to understand the links between classifiers and their associated nouns.

The emergence of computer-assisted language learning (CALL) provides language learners with a user-friendly and flexible e-learning tool. CALL incorporates technology into the language learning process and also applies itself across a broad spectrum of teaching styles, textbooks, and courses (Donaldson & Haggstrom, 2006). Its bidirectional and individualized features makes it possible for learners to use it effectively to improve different aspects of language skills (e.g. Mallon 2006; Chang et al., 2008).

My idea of designing the e-dictionary of Chinese classifiers is similar to that of CALL. Empirical studies have shown that classifier learning is a big challenge for L2 learners of Chinese. My argument with regards to Chinese classifier acquisition is that cognitive strategies with a bottom-up approach are the key to the understanding of the complexity of classifier and noun associations. Therefore, the design of the e-dictionary has a focus on guiding learners to explore the cognitive foundations of classifier-noun relations. The e-learning system implemented in the e-dictionary is designed to promote self-paced accelerated learning. It consists of a database of the decomposed semantic features of classifiers and their associated nouns. These well-defined unique and non-unique features will help learners to take a cognitive approach to explore case by case the matched pairs of classifiers and nouns. Currently the e-dictionary has included 168 noun classifiers and 680 nouns, of which 80 classifiers and 560 nouns have been analysed and entered into the e-learning database. My aim is to define and include all Chinese classifiers and their associated nouns[1] and eventually link them to the e-learning system.

2 Multi-categorization of Classifiers

In cognitive linguistics, categories are defined by groups of features and relationships within a same family. From this viewpoint, the

[1] Eleven classifier dictionaries are consulted (see References). The number of classifiers listed in them ranges from 143 to 422 and the number of associated nouns is from 388 to 8609. However, if we follow Tai and Wang's (1990) definition of classifiers, 178 of them are true classifiers.

occurrence of a noun with a particular classifier is dependent upon the categorical features of both nouns and classifiers. However, the internal semantic network of categories may be ambiguous due to historical and social factors, which makes categorization dependent on not only noun referents' intrinsic properties but also their functional and human perceptual ones. In other words, classifier and noun associations encode as well human cognitive understandings of the real world entities. As a result, classifiers are found to be able to link nouns cross-categorically. That is, one single classifier can associate itself with a number of nouns from different noun categories and at the same time one single noun from certain categories can be associated with not one but two classifiers. This multiple-categorization nature of classifiers complicates the classification of classifiers and nouns for the purpose of providing an effective learning strategy. It is also virtually impossible for linguists to build a meta-theory for a systematic organization of any clear logical classifier-noun categories and thus hard for lexicographers to find an effective way to illustrate the semantic connections between classifiers and nouns. However, one thing we are clear about is that the main obstacles in classifier acquisition are that the inhabited meaning associations in the nature of classifiers are opaque and that the complex classifier associations with nouns have caused noun categorizations to be linguistically unconventional. Yet, from a cognitive viewpoint, these associations and categorizations can provide cognitive motivations to learners if we can provide a learning tool that allows them to pay attention to the pragmatic use of classifiers on a cognitive basis.

3 Semantic Decomposition of Classifiers and Nouns

Table 1 is a demonstration of the semantic features of some most commonly used noun classifiers and their associated nouns. A total of 168 classifiers are collected and sorted out according to the number of noun categories each classifier is associated with. One special feature of this dictionary design is that the classifiers' associated nouns are grouped into categories based on the real-world entities as noun referents. Currently I have defined the following 11 categories in the e-dictionary: "nature, humans & body parts", "animals", "vegetables & fruits", "man-made objects", "buildings", "clothing", "food", "furniture", "tools" and "vehicles". A hierarchy of noun classifiers is built up according to the number of noun categories they enter into. For instance, the classifier *liang* occurs only in the "vehicles" category, (e.g. car, lorry, bicycle, etc.). Out of the 168 classifiers, 149 occur in fewer than 3 noun categories. The cognitive mapping between these 149 classifiers and their associated nouns are straightforward. Hence it is relatively easy for users to quickly have a big picture of how a classifier is associated with certain type(s) of nouns. For the rest of 19 classifiers listed in Table 1, each occurs in at least 3 noun categories. At the current stage my work focuses on individual noun classifiers; the other types of classifiers will be added in the future when more people are involved in the project. In the e-learning part of the dictionary, I temporarily exclude the general classifier *ge* because cognitively it is not assumed to be a difficult one to learn.

Through semantic decomposition, the cognitive mapping between a classifier and its associated nouns is revealed. Take the classifier *tiao* for example. It is associated with nouns such as *rainbow*, *leg*, *snake*, *cucumber*, *road*, *scarf*, *potato chip*, *boat* and *necklace*, which are from 9 of the 11 noun categories listed above. Despite of the different categories they belong to, the 9 nouns share one same cognitive property – the shape of the noun referents that is defined as "longitudinal". This shows that the classifier *tiao* is inhabited with this semantic feature as a cognitive basis and links itself to the nouns accordingly.

Similarly, the classifier *gen* is connected to the nouns such as *stick*, *bone*, *banana*, *pillar*, *sausage*, *needle,* and *ribbon* that belong to 7 noun categories respectively. These nouns possess the same "longitudinal" feature as *tiao*. This shows that extracting one same feature from *gen and tiao* is not helpful enough for learners to understand the difference between the two classifiers, though classifying nouns into categories can constrain the interference to learners to some extent. What needs to be carried

out is to define each noun with a unique feature of its own, no matter whether they are from its lexical semantic meanings, pragmatic functions, or human perceptions. For instance, besides "longitudinal", "for supporting walking" is added as a feature to *stick*, "a piece of human skeleton" to *bone*, "turns from green to yellow when ripe" to *banana*, "one end stuck to the ground" to *pillar*, etc. More are needed until finally each noun is distinguished from other nouns that are associated with the same classifier. These feature extractions and definitions are the core part of the database for the e-learning tool linked to the e-dictionary.

4. Methodology

4.1. Application of cognitive strategies in noun classifier acquisition

In this section we describe an approach that can enhance the practical use of the classifier dictionary. Developed in the software environment of FileMaker Pro 8.5 (see Figure 2), the dictionary is established within a database system. Categorical records created as data files are used to store the associated nouns. The records created so far include 11 categories of nouns described in Section 3. Such a categorization appears explicit, but its top-down approach fails to reveal the feature-based mapping between a classifier and its associated nouns. However, the e-learning part of the dictionary can guide learners to search for correct classifier and noun pairs by looking for the defined features of the noun referents in a different database, firstly from those broadly defined as "animacy", "shape", "size", "thickness", "length", "function", etc., to those specific ones extracted from each particular noun referent.

With such a bottom-up approach, the e-dictionary allows users to learn the particular interrelated features of a classifier and its associated noun referents in a case-by-case fashion. In this way, learners can better understand the point that a classifier reflects the cognitive classification of its associated noun referents. Each individual record thus contains both general and specific information of a classifier and its associated nouns as data entries. The features decomposed from the noun referents are defined and recorded as independent data entries linked to the e-learning tool. For instance, if a learner wants to know which classifier is the correct one for *boat*, he can enter the word *boat*, finds its category as "vehicles", choose its shape as "longitudinal". Then, *tiao* should automatically pop up in this case because *boat* is the only noun referent from the "vehicles" category (see Table 2). In other cases where there are two or more items that are featured as "longitudinal", the learner will be guided to look for a more specific or unique feature with a few more clicks on the users' interface.

The e-learning environment in the dictionary also provides learners the noun-classifier phrases that are commonly used but they may not be easy for learners to acquire. Take the noun classifier *zhi* for example. It is associated with noun referents that belong to "animals and body-parts", and "man-made objects", such as *bird*, *hand*, *pen*, etc. The unique perceptual features of these noun referents are identified and built into the e-learning system so that users can click different categories in the interface to make particular associations as long as they have some general knowledge of the entities, such as their functions and perceptual features, etc.

4.2 Implementation of Agent-based Model in Classifier E-learning

The e-learning tool in our classifier e-dictionary is targeted for automatic classifier-noun associations. By adopting an agent-based model (Holland, 1995), we[2] have developed a classifier-noun network for learners to learn step by step classifier phrases. Included in the prototype model will be nouns and classifiers, divided into two groups of agents. To design a semantic interface between the two types of agents with a computational approach, a tag is attached to each agent. The tags are of opposite polarity, one to a noun, and the other to a classifier. Each tag is a pseudo-binary-bit string of {0, 1, #}, where "#" is the "doesn't care" symbol. The position a symbol occupies in the string corresponds to a particular semantic feature of the agent, with "#"

[2] Acknowledgements to Ni Wei my research assistant for his contributions to the technical experiment and grants from Nanyang Technological University that supported preparation of this paper.

indicating that the corresponding feature is not critical for the formulation of the classifier phrase, even though the noun referent owns such a feature. When a noun agent meets a classifier agent, we line up the two tags and match the digits in one string with those in the other position by position. To report a match score at the end of this comparison, there are three match rules to follow: (i) it scores 1 given there is a match between two "1"s or between two "0"s; (ii) it scores 0 given there is a match between a "1" and a "#" or a "0" and a "#" or between two "#"s; (iii) it scores -1 given there is a match between a "1" and a "0". The aggregate match score indicates the likeliness of a correct classifier phrase with the involved classifier and the noun.

More specifically, in this model each tag consists of 4 pseudo-binary bits. Out of the noun's many semantic features, let's selectively represent two of them: the first feature with the first two symbols, and the second with the last two. For example, a tag "1100" is assigned to the agent (noun) *leg* to represent the noun's features defined as "longitudinal" and "body-part" respectively. In this case, "longitudinal" might be considered as the most salient feature of *leg* with regards to the selection of a classifier. Hence, it is represented by "11". On the other hand, if "longitudinal" is by no means an external or internal feature of the associated noun referent, the symbols at the corresponding positions would be "00". Other possible combinations of symbols such as "01" and "10" are reserved for fuzzy states, which are associated with marginally accepted classifier phrases.

Besides, the noun referent *leg* also has a "body-part" property listed, but it is not of primary importance for finding its classifier match. Therefore, it is represented by "##" at the last two string positions, rather than explicitly indicated by any of the four combinations mentioned above.

We assign the tags to classifier agents in a similar way. For instance, "11##" may be assigned to the classifier *tiao*, due to the fact that *tiao* often occurs in a classifier phrase with nouns defined as having "longitudinal" features. On the other hand, "##11" may be assigned to the noun classifier *zhi*, which is commonly applied to noun referents of body-part.

Regarding the agent's interaction with those agents of classifiers, when the tag "1100" of *leg* is compared with the tag "11##" of the agent *tiao*, the match score is 1+1+0+0 = 2. In contrast, its match score with the tag "##11" of the agent *zhi* is reported as 0+0+0+0 = -2. The match score 2 indicates *tiao* is more likely to be linked to *leg*, and the match score -2 implies an undesirable match between *leg* and *zhi*. It is noteworthy, however, that if a user assigns "1111" to *leg*, they will obtain a match score of 2 (0+0+1+1) with *zhi*. They will hence conclude that, beside *tiao*, *zhi* is another correct classifier for *leg*.

In addition, we include the defined features of nouns and classifiers as a group of interactive agents. This group is designed to facilitate the learning process from learner's perspective. Take L2 learners for example. First they may learn that *tiao* is the correct classifier for *leg* because the noun referent of *leg* has the longitudinal attribute. Next, they tend to look for other nouns with the longitudinal feature, such as *necklace* and *snake*, and to verify whether *tiao* is also the correct classifier for these nouns. By establishing the mapping between the defined features of nouns and classifiers, the agent-based model explicitly shows learners the possible connections between these groups of agents.

Among the semantic features, some are defined as unique features which distinguish their corresponding nouns from the rest of the nouns' group. For instance, we may define "chained jewel" as the unique feature of a necklace, and "limbless reptile, some of which produce venom" as that of a snake (see Figure 1). We assign two kinds of tags respectively, one for non-unique feature agents and the other for unique feature agents. Each non-unique feature agent is attached with an adhesion tag (Holland, 1995). This adhesion tag provides the possibility of forming multi-feature agent aggregates with individual unique feature agents. On the other hand, each unique semantic feature is attached with a two-segment tag. The first segment plays the same role as the classifier/noun tag, which controls the agent's interaction with agents of other groups, i.e. nouns and noun classifiers. The second segment functions simply as an adhesion tag. To decide whether to form a multi-feature agent aggregate, we can match a non-unique feature agent's adhesion tag and the second

segment of a unique feature agent's tag. The match score is calculated in a similar way with that between noun's agents and classifier's agents. To simplify the discussion, we assume that adhesion only occurs between one unique feature agent and one or more non-unique feature agents. In other words, adhesion does not occur between either two unique feature agents or two non-unique feature agents.

To explicitly show the cognitive mapping between classifiers/nouns and their features, we use a collection of condition/action if-then rules (Holland, 1995). In our model, both the condition and the action are linguistic variables, which are in turn represented by pseudo-binary-bit strings. The rules represent the interconnection among the agent group of classifiers, the agent group of nouns, and the group of defined features. For instance, the same noun classifier *tiao* occurs in the classifier phrase *yi tiao xianglian* 'a necklace'. Let ①, ② and ③ respectively denote the features of "chained jewel", "man-made", and "longitudinal", where ① is the unique feature to identify the noun referent of *necklace*. As discussed previously, the individual features ①, ②, and ③ can form a multi-feature agent aggregate, which we denote as ①②③. The if-then Rule 1 can be implemented as:

Rule 1: {If (①②③) Then (necklace)}.

Following the tag interaction approach discussed previously in this section, Rule 2 can be implemented to reflect the inter-agent communication between the noun and its classifier:

Rule 2: {If (necklace) Then (tiao)}.

Based on these two rules, Rule 3 can be implemented as

Rule 3: {If (①②③) Then (tiao)}.

Note that Rule 3 has the same input (condition) with Rule 1 and the same output (action) with Rule 2. Rule 1 outputs its action as a message, which is subsequently received by Rule 2 as its condition. This is an example of transitivity, a property of the rule-based network. The condition and action part in each of the three rules could also be exchanged to implement three inverse rules.

Now let's take a look at the noun *snake* and its classifier *tiao*. Given that ④ represents "animate" and ⑤ represents "limbless reptile, some of which produce venom", we can retrieve ③, ④, ⑤ from the features' group and form them as another multi-feature agent aggregate as ③④⑤. Here ⑤ is the unique feature of *snake*. We add another three if-then rules concerned with *snake* and *tiao* as follows:

Rule 4: {If (③④⑤) Then (snake)};
Rule 5: {If (snake) Then (tiao)};
Rule 6: {If (③④⑤) Then (tiao)}.

So far only multi-feature agent aggregate, rather than single feature agents are used as conditions. It is also noteworthy that non-unique feature agents are incapable of interacting directly with noun agents or classifier agents, since their adhesion tags cannot be matched with the classifier/noun tags. The property of transitivity implies, however, that we can establish the mapping between nouns and non-unique feature agents indirectly. For example, we represent the relation between the noun *necklace* and the unique feature agent ① "chained jewel" by Rule 7 as follows:

Rule 7: {If (necklace) Then (①)}

We also represent the relation between the noun *snake* and the unique feature agent ⑤ "limbless reptile, some of which produce venom" by Rule 8 as follows:

Rule 8: {If (snake) Then (⑤)}

Either ① or ⑤ is related with the non-unique feature agent ③ "longitudinal", which could be represented by Rule 9 & 10.

Rule 9: {If (①) Then (③)}
Rule 10: {If (⑤) Then (③)}

The mapping between *necklace*/*snake* and the non-unique feature agent ③ "longitudinal" could then be implemented by Rule 11 & 12.

Rule 11: {If (necklace) Then (③)}
Rule 12: {If (snake) Then (③)}

In Rule 11 and 12, the noun is taken as the input and the non-unique semantic feature as the output. By swapping the two kinds of agents' roles in the message-processing rules, we may inversely implement Rule 13 by taking the non-unique feature as the input and the noun as the output. If a learner chooses ③ as the single input agent, two possible outputs pop up for his/her selection.

Rule 13: {If (③) Then (necklace or snake)}

More rules could be added in the classifier network by selecting different agents from the three groups in a similar way as we implement Rule 1-13. In this way the if-then rule-based network explicitly shows the cognitive mapping between the classifiers and their associated nouns. Learners will find out the association between the target words and their features, which are essential for their classifier acquisition.

So far we have tested some commonly used classifiers and their associated nouns selected from the e-dictionary and tried within the agent-based model. The automatic matching is successful, though more pairs need to be tested.

5. Conclusion

This paper presents a feature-based approach in designing a classifier e-dictionary with an e-learning environment created for learners to use cognitive strategies to explore and learn the classifier phrases.

The current dictionary is based on a database with classes of nouns (11 classes at present) and classifiers (168 added) that are stored as individual records. The records are not organized according to the lexical meanings of the words. Instead, the classification scheme is based on the noun referents' semantic and salient external or functional features. The objective of the design is to use such features to set up a classifier network that can automatically associate all possible nouns. A computer-based model with such a design is expected to show learners of Chinese the cognitive base of linguistic combinations. The proposed agent-based model uses the match between pseudo-binary-bit strings to indicate the probability of interactions between agents. It hence predicts how likely a classifier and a noun occurs in a classifier phrase. The relations among the agent groups are shown within the framework of the if-then rule-based network. Learners can explore case by case, when using the dictionary's e-learning function, the classifier and noun associations and the defined features that the associations are based on. The future task is to include the rest of the classifiers and all possible associated nouns. Linguistically, a challenge to carry out the task would be the definitions of the unique features of the noun referents and their classifiers that have fuzzy boundaries. Technically, the challenge would be the solution to making perfect matches of those cases where one classifier agent as input is expected to link automatically a number of noun agents as output, which should follow a step-by-step procedure that is interesting and effective from learners' perspective.

References

Adams, Karen L., and Nancy Faires Conklin. 1973. Toward A Theory Of Natural Classification. *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, Univeristy of Chicago, 1-10.

Allan, Keith. 1977. Classifiers. *Language*, *53*(2), 285–311.

Chang, Yu-Chia, Jason S. Chang, Hao-Jan Chen, and Hsien-Chin Liou. 2008. An Automatic Collocation Writing Assistant for Taiwanese EFL Learners: A Case of Corpus-based NLP Technology. *Computer Assisted Language Learning*, *21*(3), 283-299

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. University of California Press.

Donaldson, Randall P., and Margaret A. Haggstrom. 2006. *Changing Language Education Through CALL*. Routledge.

Downing, Pamela. 1993. Pragmatic and Semantic Constraints on Numeral Quantifier Position in Japanese. *Linguistics*, *29*, 65–93.

Gao, H. H. (2010 to appear). A Study of the Swedish Speakers' Learning of Chinese

Classifiers. *Nordic Journal of Linguistics.* Special Issue, Vol. 33.

Goddard, Cliff. 1998. Semantic Analysis: A Practical Introduction. Oxford: Oxford University Press.

Guo, Hui., and Huayan Zhong. 2005. Chinese Classifier Assignment Using SVMs. Paper presented at the 4th SIGHAN Workshop on Chinese Language Processing, Jeju Island, pp. 25–31.

Holland, John. H. 1995. *Hidden Order: How Adaption Builds Complexity*. Addison-Wesley.

Huang, Chu-Ren, and Ahrens, Katherine. 2003. Individuals, Kinds and Events: Classifier Coercion of Nouns. *Language Sciences*, *25*, 353–373

Hwang, Soonhee, Ae-sun Yoon, and Hyuk-Chul Kwon. 2008. Semantic Representation of Korean Numeral Classifier and Its Ontology Building for HLT Applications. *Language Resources and Evaluation*, *42*, 151–172.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

Mallon, Adrian. 2006. ELingua Latina: Designing a Classical-Language E-Learning Resource. *Computer Assisted Language Learning, 19*(4), 373-387.

Matsumoto, Yo. 1993. Japanese Numeral Classifiers: A Study Of Semantic Categories and Lexical Organization. *Linguistics*, *31*(4), 667–713.

Nichols, Eric, Francis Bond, and Daniel Flickinger. 2005. Robust Ontology Acquisition from Machine-Readable Dictionaries. Paper Presented at the 19th International Joint Conference on Artificial Intelligence, Edinburgh, pp. 1111–1116.

Nirenburg, Sergei, and Victor Raskin. 2004. *Ontological Semantics*. Cambridge: MIT Press.

Philpot, Andrew G., Michael Fleischman, and Eduard H. Hovy. 2003. Semi-automatic Construction of A General Purposeontology. Paper Presented at the International Lisp Conference, New York, pp. 1–8.

Quek, See Ling. 2010. A Diachronic Semantic Study of the Two Collocations: “Tiao + Ming” and “Tiao + Xinwen”. The 10th Chinese Lexical Semantics workshop (CLSW2010), 21-23 May, 2010. Soochow University, China.

Sowa, John. F. 2000. *Knowledge Representation*. Pacific Grove, CA: Brooks Cole Publishing Co.

Tai, James H-Y., and Lianqing Wang. 1990. A Semantic Study of the Classifier Tiao. *Journal of the Chinese Language Teachers Association,* 25.1: 35-56.

Classifier Dictionaries Consulted

陈保存等《汉语量词词典》，福州：福建人民出版社，1988。

郭先诊《现代汉语量词手册》，北京：中国和平出版社，1987。

郭先诊《现代汉语量词用法手册》，北京：语文出版社，2002。

何杰《量词一点通》，北京; 北京语言文化大学出版社，2003。

焦凡《看图学量词》，北京：华语教学出版社，1993。

焦凡《汉英量词词典》，北京：华语教学出版社，2001。

刘学武、邓崇谟《现代汉语名词量词搭配词典》，杭州：浙江教育出版社，1989。

吕叔湘《现代汉语八百词》（增订本），北京：商务印书馆，1999。

殷焕光、何平《现代汉语常用量词词典》，济南：山东大学出版社，1991。

俞士文等《现代汉语语法信息词典详解》，北京：清华大学出版社，2003。

褚佩如、金乃莉《汉语量词学习手册》，北京：北京大学出版社，2002。

Classifier in Chinese	Classifier	No. of categories the classifier occurs with	Examples of nouns the classifier occurs with
条	tiao	9 (nature, humans & body parts, animals, vegetables & fruits, buildings, clothing, food, vehicles, other man-made objects)	rainbow, leg, snake, cucumber, road, scarf, potato chip, boat, necklace
根	gen	7 (nature, humans & body parts, vegetables & fruits, buildings, food, tools, other man-made objects)	stick, bone, banana, pillar, sausage, needle, ribbon
块	kuai	6 (nature, humans & body parts, clothing, food, tools, other man-made objects)	stone, scar, handkerchief, candy, eraser, soap
层	ceng	5 (nature, humans & body parts, building, clothing, other man-made objects)	wave/fog, skin, building storey, curtain, paper
张	zhang	5 (humans & body parts, food, furniture, tool, other man-made objects)	mouth, pancake, bed, bow, map
只	zhi	5 (humans & body parts, animal, clothing, vehicle, other man-made objects)	ear, tiger, sock, sailing boat, watch
粒	li	4 (nature, vegetables & fruits, food, other man-made objects)	sand, cherry, rice, sleeping tablet
段	duan	4 (nature, vegetables & fruits, building, other man-made objects)	wood, lotus root, city wall, iron wire
口	kou	4 (humans & body parts, animal, tools, other man-made objects)	person(people), pig, sword, well
面	mian	4 (buildings, tools, furniture, other man-made objects)	wall, drum, mirror, flag
节	jie	4 (building, food, tool, vehicle)	chimney, sugarcane, battery, railway carriage
道	dao	3 (nature, humans & body parts, building)	lightening, eyebow, dam
滴	di	3 (nature, humans & body parts, other man-made objects)	water/rain, blood, ink
件	jian	3 (clothing, tools, other man-made objects)	shirt, (music) instrument, toy
把	ba	3 (furniture, tools, other man-made objects)	chair, knife, cello
截	jie	3 (nature, tools, other man-made objects)	rope, pencil, pipe
颗	ke	3 (nature, humans & body parts, other man-made objects)	star, tooth, artillery shell
片	pian	3 (nature, food, other man-made objects)	leaf, loaf, tablet
枝	zhi	3 (nature, tools, other man-made objects)	rose, pen, arrow/rifle

Table 1. A Selection of classifiers sorted by how many noun categories they are associated with

English equivalent of Chinese classifier phrase	Classifier phrase in Chinese			Properties	
	numeral	classifier	noun	cognitive	intrinsic
a rainbow	*yi*	*tiao*	*caihong*	longitudinal	nature
a leg	*yi*	*tiao*	*tui*	longitudinal	human
a snake	*yi*	*tiao*	*she*	longitudinal	animal
a cucumber	*yi*	*tiao*	*huanggua*	longitudinal	vegetable
a road	*yi*	*tiao*	*lu*	longitudinal	building
a scarf	*yi*	*tiao*	*weijin*	longitudinal	clothing
a potato chip	*yi*	*tiao*	*shutiao*	longitudinal	food
a boat	*yi*	*tiao*	*Chuan*	longitudinal	vehicle
a scarf	*yi*	*tiao*	*weijin*	longitudinal	man-made

Table 2. A Selection of noun-classifier phrases of *tiao.*

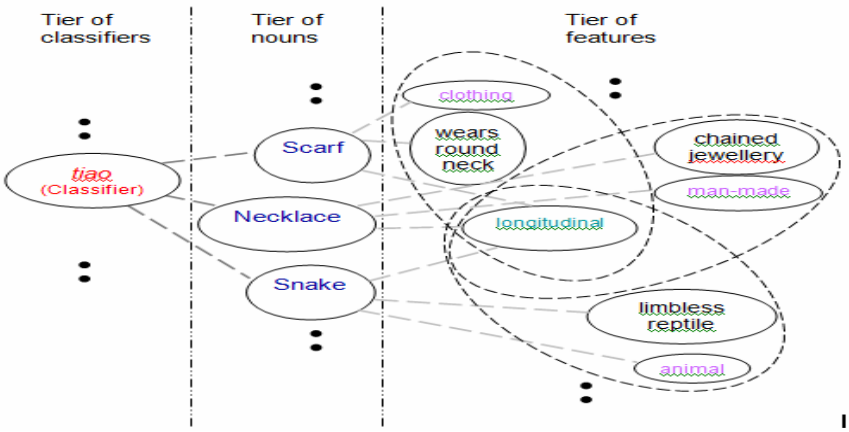


Figure 1. Mapping among the tiers of classifiers, nouns, and defined features.

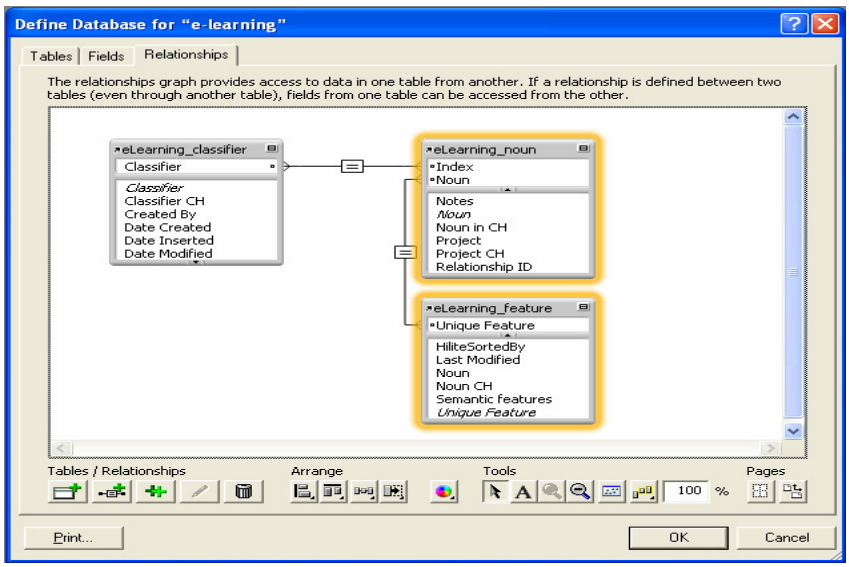


Figure 2. A display of the database in the e-learning environment

In search of the "right" word

Stella Markantonatou
Institute for Language and Speech Processing / R.C. "Athena"
marks@ilsp.gr

Aggeliki Fotopoulou
Institute for Language and Speech Processing / R.C. "Athena"
afotop@ilsp.gr

Maria Alexopoulou
Institute for Language and Speech Processing / R.C. "Athena"
maralex@ilsp.gr

Marianna Mini
Institute for Language and Speech Processing / R.C. "Athena"
minimar@ilsp.gr

Abstract

We report on a user needs investigation carried out in the framework of the project EKFRASIS[1] that developed a platform for supporting authoring work in Modern Greek. The platform had as a backbone a conceptually organised dictionary enhanced with rich lexicographic and morphosyntactic information. Organisation of information and encoding drew on Semantic Web technologies (ontologies). Users were all professional authors (of literature, editors, translators, journalists) working for well-established firms. They were all familiar with printed conceptually organized dictionaries while most of them used a computer. They were asked to specify how the platform would be helpful to them when they searched for a word for which they had only vague or few clues, a situation that was familiar to all of them. Users preferred to have, in a first step, easy access to limited but to-the-point lexical information while access to rich semantic information should be provided at a second step. They were interested in rich lexical material although they were not really able to identify the relations that would help them retrieve it. They strongly preferred an organization of material by concept and PoS and appreciated easy access to normative information.

1 Introduction

We present an investigation of user needs that was conducted in the process of developing a platform for supporting authoring work in Modern Greek. One main component of this platform is the dictionary "EKFRASIS" (literally '*expression*'). EKFRASIS exploits technology and ideas developed for the Semantic Web (Guarino and Welty, 2002) to encode a conceptually organized dictionary enriched with translations as well as a wealth of lexicographic and morphosyntactic information. The overall organisation of the dictionary and, partially of the platform, aims at helping the user who needs a word but has few clues or just guesses about it and about its way of use.

The interviewed users were all professionals: journalists, translators, editors and authors of literature.

In this survey, we aimed to map user expectations concerning interaction with the dictionary rather than to look for appropriate ways for populating it. Of course, some of the conclusions reached here may be useful for collecting linguistic material as well.

2 The "recollection problem"

Authors (of any type of text) often find themselves in the uncomfortable position whereby they remember or feel that the "right" word exists but they can not recall it. Here, we will call this situation the "recollection problem". Alphabetically organised dictionaries are of little help

[1] EKFRASIS http://www.ilsp.gr/ekfrasi_eng.html was funded by the General Secretariat of Research and Technology / Greece.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 66–74,
Beijing, August 2010

in such situations. Of course, the recollection problem is a well known one (for a discussion, see Zock and Bilac, 2004; Zock and Schwab, 2008). Works with an international reputation have tried to face it and, of course, Roget's Thesaurus, a printed dictionary, and WordNet and EuroWordNet, both digital ones, spring easily to mind. An interesting difference between the two dictionaries is exactly about the interaction with the user: while in Roget's lexical relations are left implicit and material is explicit, in WordNet one has to go via lexical relations such as "hyponymy" and "sister term" to find the material. In short, in Roget's one is given the material directly and no previous familiarity with taxonomies is required while in WordNet one has to guess what taxonomic labels can offer in each case.

Furthermore, just finding a word is often not enough to guarantee confidence in its usage, so information provided by an assortment of relevant sources is indispensable. Of course, printed dictionaries would expand to unmanageable volume if they accommodated all necessary information, but digital dictionaries are not subject to such limitations. For instance, Roget's does not provide syntactic information but WordNet does. Digital lexica that have been compiled for NLP purposes such as Acquilex (Briscoe et al., 1993) and Simple have tried to accommodate semantic, syntactic, morphological and pragmatic information structured in a principled way and make it available to machines. In the Semantic Web era, efforts to better axiomatise the established resources have been made (Gangemi et al., 200; Old, 2004) as well as to combine detailed linguistic with ontological information for the purposes of Machine Translation (Nirenburg and Raskin, 2004). Such efforts were oriented to NLP mainly.

As regards human users, Roget's success suggests that while wealth of material and good organisation of it matters, it is not necessary to present explicitly the relations among words in a 'concept'[2] in order to provide a good solution to the recollection problem. On the other hand, humans do instinctively look for words on the basis of domain relations several of which can be argued to mirror (aspects of) human cognition (Gaume et al., 2003; Kremer et al., 2008). Furthermore, the data we will present here indicate that searching for a word is also a matter of habit and training, having a lot to do with one's profession and familiarity with certain types of dictionary.

As already said, we report on a user requirement survey that was conducted in the framework of the project EKFRASIS. Our aim was primarily to see how the lexicographic material should be presented to the user; however, we were also interested in fine-tuning our ideas about the nature of the material required. We start with a brief presentation of the main ideas behind the dictionary EKFRASIS. Next, we talk about how we organized the survey. For each group of questions addressed to the users, we present and comment their responses. Finally, we present our conclusions and decisions as regards the architecture of the dictionary EKFRASIS.

3 About EKFRASIS

A few words about the dictionary EKFRASIS are in order to set the context of the survey discussed in this paper.

EKFRASIS is a digital lexicon of Modern Greek (MG).[3] MG market makes available mainly alphabetically ordered printed dictionaries and, more recently, dictionaries of synonyms and antonyms. There is no dictionary of MG that combines lexicographic with semantic and morphosyntactic information.

There is one excellent thesaurus of Modern Greek, the "Onomasticon" by Theofilos Vostantzoglou, published back in 1962. "Onomasticon" enjoys great reputation among Greek intellectuals. T. Vostantzoglou drew a lot on Roget's Thesaurus. His material is organized in concepts and each concept is further structured into smaller groups of words. Notably, he introduced certain innovations of his own: contrastive presentation of concepts, eg "Joking" and "Speaking seriously" are presented contrastively on the same page using a special format, short definition of groups of words that form a concept and information on style. In addition, spe-

[2] We use the term *concept* to define a set of words that are conceptually related. Such sets are defined in Roget's Thesaurus and the "Onomasticon".

[3] It comprises 6000 entries at the moment.

cial attention is given to expressions and proverbs. Links between concepts exist at lexical level introducing, in this way, relations among words in different concepts. There is no actual typology of these relations which remain implicit to the user. Very much like Roget's, all the words in the conceptually organized part of the Onomasticon are also listed in alphabetical order in the second half of the book. Each word is indexed for the set of concepts it belongs.

In EKFRASIS we understand "conceptual" organization as the result of the interplay of a set of domain relations. In addition, EKFRASIS is planned to help with the usage of words by providing definition of concepts, glosses and translation of words, usage examples and full morphosyntactic information. In our survey, we exposed the users to large amounts of information organized in more than one ways and asked them whether such a tool would be useful to them in their professional lives. Throughout the present discussion, we ask the reader to keep in mind that we aim at developing a resource that would be of use to a wide audience.

4 The Survey

Although the project EKFRASIS is addressed to anyone who writes in Modern Greek, we thought that professionals would have a clearer view of the authoring procedure and its needs.

Authors of literature	5
Editors	2
Translators	8
Journalists	6
Total	**21**

Table 1. Composition of the user group

Table 1 shows the composition of the user group. Each group of professionals works under different conditions and has different requirements. In general, journalists and translators work under time pressure while authors of literature and editors are more concerned with linguistic and aesthetic quality. In the user group, authors, literature translators and editors are well known intellectuals of the country publishing with the publishing house "Kastaniotis" (http://www. kastaniotis.com/), industrial translators work for a medium size firm (http://www.orco.gr/ loc/frameset-gr.html?navbar-gr.html&0) and journalists work for the prestigious daily newspaper "I Kathimerini" (http://www.kathimerini.gr/).

Users were interviewed by teams of researchers and

- each of the literature authors and editors were interviewed personally
- industrial translators and journalists were interviewed in groups according to their own requirement.

To each person or group a presentation of the aims of the project was given and then, a specially developed mock-up of a couple of usage cases of the authoring tool was presented. We used a different mock-up (.ppt file) for each author and editor and for each of the journalist and translator groups because EKFRASIS is a novel application by Modern Greek standards and, since no user had some relevant experience, verbal description would not help at all. The presentations were followed by discussion and an interview that was taped and, finally, the users were asked to fill a questionnaire. Each user (author, editor, translator, journalist) filled in a separate questionnaire.

For the first part of the mock-up, pieces of text produced by each one of the authors, a piece of Greek translation from the site of Nokia and a piece from "I Kathimerini" were selected to develop the presentation that relied on the following scenario: hypothetically, when they developed the particular text, the authors experienced the recollection problem and used EKFRASIS to find a couple of words. They started their search with an input word of somehow related meaning to the intended one but not necessarily of the same PoS. We present as an example extracts from the mock-up developed for one of the authors of literature.

In Figure 1 the author uses the noun *mania* as an input word. EKFRASIS returns the names and definitions of all the concepts where the input word occurs (gray area on the left). The concepts are not necessarily related to each other. Here, the first one corresponds to '*proclivity, propensity'*, the second to '*urge'*, the third to '*mania'* as a disorder, the fourth to '*mania, passion'*, the fifth to '*fury'*, the sixth to *'wrath'* and the seventh to '*love'*. On the right, for each concept a set of semantically relevant words are

given, all of the same PoS with the input one (here, a noun). For instance, concept 6 is labeled *οργή 'wrath'* and the words given are *οργή 'wrath', θυμός 'anger', νεύρα 'nerves', αγανάκτηση (being indignant), μανία 'mania'.*

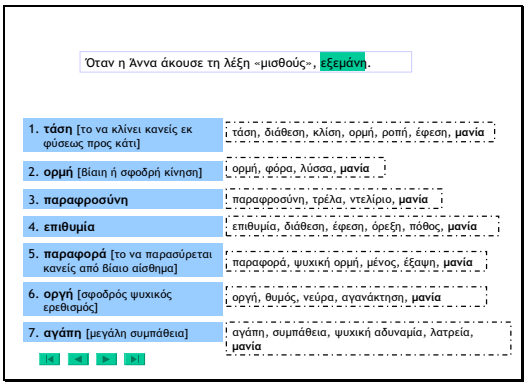


Figure1. Unrelated numbered concepts indicated with the labels on the left

In this hypothetical scenario, the user selects the 6th option (*wrath*). The screen shown in Figure 2 pops up (on the upper left corner the label is *οργή* '*wrath*').

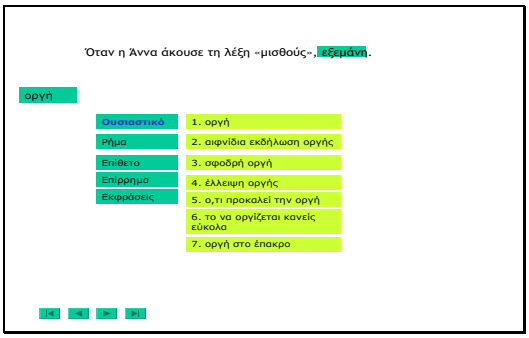


Figure 2. The right hand column consists of the set of related concepts within a PoS (noun here)

In this screen, the concepts are related to each other as they all are about *wrath.* The concepts in successive order are *'wrath', 'sudden expression of wrath', 'strong wrath', 'luck of wrath', 'what causes wrath', 'to become furious easily', 'extreme wrath'* and are listed on the right hand column. The list of PoS (noun, verb, adjective, adverb, expressions) is on the left hand column. In Figure 2, the activated PoS is 'noun', however, the user could activate any other PoS, an option he picks in Figure 4, where he activates the PoS 'verb'. In our hypothetical scenario, the user selects the 7th concept (*extreme wrath*) and receives the set of synonyms and antonyms, enclosed in the boxes in Figure 3. Antonyms are given in the box at the bottom of the screen. The user stores these findings and carries on with the PoS 'verb' (Figure 4).

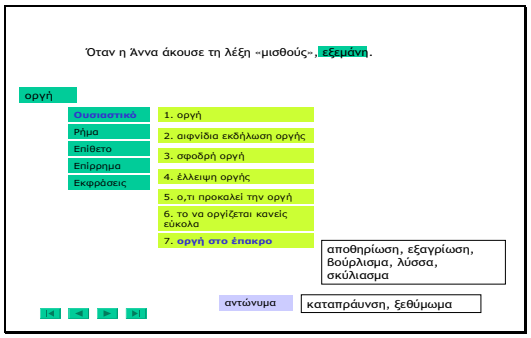


Figure 3. Synonyms and antonyms in a concept and PoS ('*extreme wrath*' and 'noun'). Words are organized in related numbered concepts indicated with the labels on the right

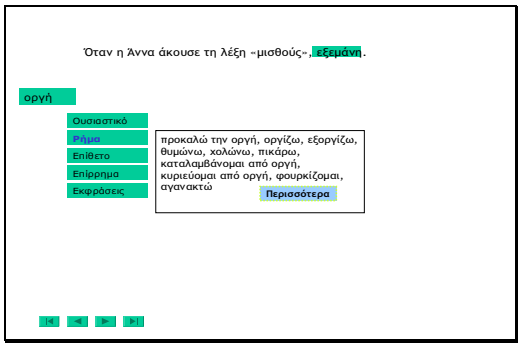


Figure 4. The PoS 'verb' for the concept '*extreme wrath'*

This PoS comprises only one concept, so the system shows the box with the lexical material. There is plenty of lexical material so the indication 'more' guides the user who would require access to more information.

Two series of user responses were elicited -- one with the questionnaire and one with the interviews. The results of the interviews were not directly measurable but they greatly helped to clarify the picture. In what follows we will present the measurable results but in the conclusions, we will draw on the interviews as well.

The questionnaire included different types of question (multiple choice questions, questions with graded answers, open questions). In this report we are interested in the parts of the ques-

tionnaire on (i) the contribution of dictionaries and spelling checkers to authoring work (ii) how authors experience the recollection problem and (iii) efficient searching for words.

5 Questioning the users

5.1 Contribution of technology and dictionaries to authoring work

We first asked if professionals used computers and authoring aids.

Nearly 100% of users worked on a computer. All users with the exception of journalists used general language dictionaries or specialized ones including dictionaries of synonyms (Table 2). Journalists, on the other hand, seemed to pay attention mainly in word spelling.

	Authors	Editors	Translators	Journalists	Total
Dictionaries (printed, digital)	5	2	6	1	14
Special dictionaries	5	2	4	0	11
Spelling checkers	0	0	2	4	6

Table 2. Usage of authoring aids

Why do users use dictionaries and spelling checkers? We asked the users to give a yes/no answer to two contradictory statements. So, to the statement "I can find easily what I look for in dictionaries and spelling checkers" all other users and 2 journalists answered "yes". However, to the opposite statement "I can not find easily what I look for in dictionaries and spelling checkers" no journalists replied while one editor and 3 authors replied "yes". So, the picture is somehow messed up but in the overall it shows that people have formed expectations about dictionaries and look for certain types of information that they more or less find. In addition, 1 editor, 4 translators, 3 authors and 4 journalists stated that dictionaries and spelling checkers speed up their work.

So, there is some kind of information in the dictionaries and the spelling checkers that helps and speeds up authoring work. Certainly, these resources provide little help with the recollection problem. So, in the next parts of the questionnaire we set out to make clear what kind of information users considered useful in the dictionaries and what else they would like to find and how. Mock-up presentations were of crucial help in this task.

5.2 Experiencing the recollection problem

We asked how often and in which way users experienced the recollection problem. Users turned out to be quite familiar with experiences such as the following: (i) looking for a synonym (ii) entertaining the belief that a word exists but not being able to identify it (iii) having a general notion in mind but not being able to find words that ground it properly (Table 3).

	Journalists	Editors	Translators	Authors	Total
Synonym	1	2	6	5	14
Heard in the past	0	2	6	5	13
General notion	4	1	6	5	16
Other	0	0	0	3	3

Table 3. The recollection problem

It is clear that professional demands influence authoring behavior at work: authors of literature, editors and translators are more into searching for the appropriate word than journalists. According to the users, two are the main motives for searching words: (i) precision (ii) aesthetics (the example put forward by the users was the repetition of a word).

It is also interesting that the answers already point to some important lexical relations, namely synonymy and hypernymy / hyponymy, which can be used for both constructing and navigating a resource (and have been used in many dictionaries, for instance in WordNet).

5.3 Searching for lexical information

Despite the fact that answers such as "heard in the past" and "general notion" (Table 3) indicate that a variety of semantic relations can be used for searching for a word, it was not easy for the users to imagine which those relations would be. This became obvious because

- users were able to fill in the part of the questionnaire on searching only after they

had seen the mock-up that presented them with searching strategies other than alphabetical listing

- although they had twice in the questionnaire the opportunity to describe some way of searching for words other than the ones suggested in the mock-up (see below) they did not do it –despite the fact that interviews lasted for about 2.00 hours each on the average.

Although they did not identify alternative searching strategies, they in the overall welcomed the possibility of using morphologically (Table 4) and pragmatically (Table 5) related words for dealing with the recollection problem.

The example given to the users for illustrating morphologically related words was the triple *adopt->adoption->adopted*. The users were asked whether they would like the dictionary to ensure that such sequences would be automatically offered.

	Journalists	Editors	Translators	Authors	Total
YES	2	2	8	2	14
NO	4	0	0	2	6

Table 4. Morphological relations

A similar question was asked about domain relations with the following example: "if you type in *sell* or *seller* or *offer* would you like to be given automatic access to *buyer, client, consumer*."

	Journalists	Editors	Translators	Authors	Total
YES	5	2	6	4	17
NO	1	0	0	0	1

Table 5. Domain relations

Of the relations named explicitly in the questionnaire, synonymy turned out to be the most likely one to be used for helping with the recollection problem (Table 6).

Hypernymy/hyponymy (Table 7) and antonymy (Table 8) were in the overall considered less useful for the recollection problem than synonymy, but still a high percentage voted for them.

	Journalists	Editors	Translators	Authors	Total
Often	2	2	8	3	15
Occasionally	4	0	0	2	6
Rarely	0	0	0	0	0
Never	0	0	0	0	0

Table 6. Synonymy

	Journalists	Editors	Translators	Authors	Total
Often	1	2	4	3	10
Occasionally	4	0	4	1	9
Rarely	1	0	0	1	2
Never	0	0	0	0	0

Table 7. Hypernymy/Hyponymy

	Journalists	Editors	Translators	Authors	Total
Often	1	2	4	1	8
Occasionally	1	0	2	2	5
Rarely	3	0	2	2	7
Never	1	0	0	0	1

Table 8. Antonymy

Next, users were asked how they would like to see information presented (Table 9). They were offered mock-ups of the following four choices:

	Journalists	Editors	Translators	Authors	Total
A	4	0	0	1	5
B	0	1	2	1	4
C	2	1	4	1	8
D	0	0	2	2	4

Table 9. Presentation of information (a)

- **A**. All linguistic material pops up in strict alphabetical order only (no PoS or semantic classification)
- **B.** Only the linguistic material in the same PoS as the input word pops up in alphabetical order

- **C.** All linguistic material pops up in the form of a semantic tree[4] together with PoS information—the tree reflects the conceptual organization of the dictionary
- **D.** Only the part of the tree that belongs to the PoS of the input word pops up

In fact, options A and C are opposed to options B and D in that the first pair offers immediate access to the whole lot of the related lexical material while the second pair only to the semantically related words in the same PoS with the input word---practically, its synonyms. In Table 9.a. we sum the answers according to this division. While the A&C option was preferred by 62% of the users, the fact that a good 38% has chosen the B&D option had to be taken into account.

	Journalists	Editors	Translators	Authors	Total
A&C	6	1	4	2	13
B&D	0	1	4	3	8

Table 9.a. Presentation of information (b)

A compromising solution to this split of requirements is to present first the minimal necessary information, practically the synonyms of the input word, and then allow users navigate through the whole lot of it.

Organisation of information in the form of a 'semantic tree' seems more popular than simple alphabetical presentation of material. Still, some users prefer information to be presented as it always had, in alphabetical order, without any other complication.

5.4 More information of interest

	Journalists	Editors	Translators	Authors	Total
Often	1	2	2	4	9
Occ/ly	1	0	4	0	5
Rarely	4	0	2	1	7
Never	0	0	0	0	0

Table 10. Expressions

[4] 'Semantic tree' is a description of the situation in Figure 2 where words are organized in concepts indicated with the labels on the left hand column).

The recollection problem concerns both one word units and expressions or collocates as is clearly indicated in Table 10 and Table 11. The question asked here was "How often do you look for an expression / a collocate?".

We must note here that, ahead of its time, Onomasticon puts special emphasis on both types of information. Usage examples (Table 12) were requested mainly by translators and authors while glosses (Table 13) turned out to be of medium interest.

	Journalists	Editors	Translators	Authors	Total
Often	1	1	6	3	11
Occ/ly	5	1	0	0	6
Rarely	0	0	2	2	4
Never	0	0	0	0	0

Table 11. Collocates

	Journalists	Editors	Translators	Authors	Total
Often	0	1	4	2	7
Occasionally	1	0	4	1	6
Rarely	1	1	0	2	4
Never	4	0	0	0	4

Table 12. Usage examples

	Journalists	Edi3tors	Translators	Authors	Total
Often	0	1	2	1	4
Occasionally	3	0	4	2	9
Rarely	2	1	2	2	7
Never	1	0	0	0	1

Table 13. Glosses

On the other hand, users were interested in inflection (Table 14) and spelling (Table 15) information, as well as "the right context of usage" (Table 16) of words or expressions.

	Journalists	Editors	Translators	Authors	Total
Often	4	1	4	4	13
Occ/ly	0	0	2	1	3
Rarely	2	1	2	0	5
Never	0	0	0	0	0

Table 14. Inflection

This comes as no surprise because Modern Greek reflects the long history of the language in several, often confusing ways:

- by providing more than one spellings for a word, eg *μείγμα / μίγμα* (*mixture*)
- by being heavily inflected with many types for the same set of PoS and grammatical features, eg

  έτρωγαν, τρώγαν, τρώγανε
  eat – 3rd, plural, past continuous

Furthermore, the different forms are related with different styles, for instance in the example above the first form is considered the standard and the last the colloquial one.

	Journalists	**Editors %**	**Translators**	**Authors**	**Total**
Often	3	0	6	1	10
Occ/ly	1	0	0	1	2
Rarely	1	2	2	2	7
Never	0	0	0	0	0

Table 15. Spelling

	Journalists	**Editors**	**Translators**	**Authors**	**Total**
Often	4	0	6	3	13
Occ/ly	0	0	2	1	3
Rarely	1	1	0	1	3
Never	1	1	0	0	2

Table 16. Right context of usage

Naturally, editors need the particular facilities less than the other groups given their specialty.

6 Conclusions and decisions

The professional occupation of the users seemed to determine the kind of available authoring aids they preferred. This may be due to the fact that they work under serious time pressure. However, they too experienced the recollection problem and would take advantage of domain relations to solve it. Editors require the least normative and usage information. Authors and translators seem to appreciate all types of information. Translators also stressed that efficient presentation of information is important, probably because they use translation aids regularly.

On the basis of the questionnaire and the interviews, we conclude that the features of the dictionary that would be attractive to the majority of users are:

- **Easy access to limited but to-the-point lexical information**: the most usable synonyms and the most useful derivatives
- **Access to rich semantic information must be provided** to those interested, although at a second step. **Users are more interested in the lexical material than the labeling of semantic relations**
- **Categorisation of lexical material by PoS**. The PoS of the input word pops up first but all the other PoS in the same concept are available
- **Easy access to normative information** given the particularities of Modern Greek in morphology and syntax

On the basis of the above general conclusions, EKFRASIS interface design would satisfy the following minimum requirements:

- All concepts to which an input word belongs are shown at the first step of the search, together with the synonyms of the word in each concept
- Once a concept is selected, all existing sub-concepts are presented
- If no sub-concepts exist, all PoS in a concept are made available to the user who sees the concept definition and the set of synonyms for each PoS (if they exist)
- Once a word is selected all material about it pops up: gloss, example of usage, lexical relations, inflection, syntactic properties, collocations, translation

Figure 5 is an extract from a mock-up where the hypothetical user has retrieved the multiword expression *φέρνω στο φως* '*bring to light'*. Information given includes: gloss, synonyms *('present', 'give away', 'make something visible', 'make something obvious'*), inflection information (concerning tense formation) and hints on its the usage. Users agreed that for each word an exhaustive summary of its properties should be provided.

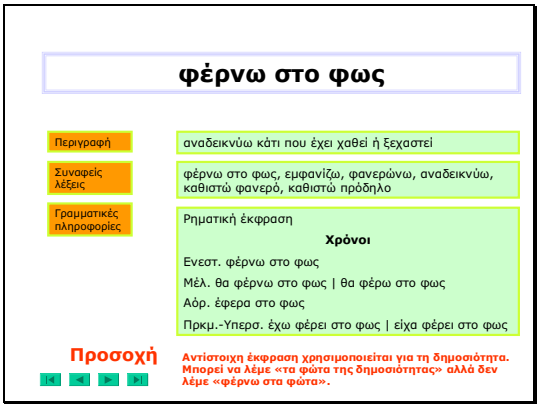


Figure 5. Presenting words

References

ACQUILEX http://www.cl.cam.ac.uk/research/nl/acquilex/

Briscoe, Ted, Ann Copestake and Valeria de Paiva. 1993. (eds). *Inheritance, Defaults and the Lexicon.* Cambridge University Press, Cambridge, UK.

Fellbaum, Christiane. 1998. *WordNet: An electronic Lexical Database*, MIT Press, Cambridge.

Gangemi, Aldo, Roberto Navigli and Paola Velardi. 2003. *The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet*. In R. Meersman et al. (eds): CoopIS/DOA/ODBASE 2003, LNCS 2888, Springer-Verlag Berlin Heidelberg: 820-838.

Gaume, Bruno, Karine Duvignau, Laurent Prevot and Yann Desalle. 2008. *Toward a cognitive organisation for electronic dictionaries, the case for semantic proxemy.* In COLING 2008: Proceedings of the Workshop on cognitive Aspects of the Lexicon (COGALEX 2008):86-93.

Guarino, Nicola and Chris Welty. 2002. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM.* 45(2):61-65. New York:ACM Press

Kremer, Gerhard, Andrea Abel and Marco Baroni. 2008. *Cognitively Salient Relations for Multilingual Lexicography.* In COLING 2008: Proceedings of the Workshop on cognitive Aspects of the Lexicon (COGALEX 2008): 94-101.

Nirenburg, Sergei and Victor Raskin. 2004. *Ontological Semantics.* The MIT Press.

Roget's II: The New Thesaurus, Third Edition. 1995. (URL: http://www.bartleby. com/62/11.html)

Old, L. John. 2003. *The Semantic Structure of Roget's: A whole language thesaurus.* (URL: http://www.dcs.napier.ac.uk/~cs171/LJOld/papers/Dissertation.pdf)

Old, L. John. 2004. *Roget's Thesaurus of English Words and Phrases, by Roget ed. 1911* (URL: http://www.gutenberg.org/etext/10681).

SIMPLE http://www.ub.es/gilcub/SIMPLE/simple.html

Vossen, Piek. 1999. (ed.) *EuroWordNet, General Document*, (URL: http://www.illc.uva.nl/EuroWordNet /docs.html)

Zock, Michael and Bilac, Slaven. 2004. *Word lookup on the basis of associations: from an idea to a roadmap.* In COLING 2004: Proceedings of the Workshop on enhancing and using electronic dictionaries: 29-35.

Zock, Michael and Didier Schwab. 2008. *Lexical Access Based on Underspecified Input.* In COLING 2008: Proceedings of the Workshop on cognitive Aspects of the Lexicon (COGALEX 2008): 9-18.

Βοσταντζόγλου, Θεολόγος. 1962. *Αντιλεξικόν ή Ονομαστικόν της νεοελληνικής*. Δομή, Αθήνα.

Lexical Access, a Search-Problem

Michael Zock (1), Didier Schwab (2) and Nirina Rakotonanahary (2)
(1) LIF-CNRS, TALEP, 163, Avenue de Luminy, 13288 Marseille, France
(2) LIG-GETALP, University of Grenoble, 38041 Grenoble, France
zock@free.fr, didier.schwab@imag.fr, damanidaddy@msn.com

Abstract

Our work is confined to *word access*, that is, we present here our ideas of how to improve electronic dictionaries in order to help language producers (speaker/writer) to find the word they are looking for. Our approach is based on *psychological findings* (representation, storage and access of information in the human mind), observed *search strategies* and typical *navigational behavior*.

If one agrees with the idea that lexical access (word finding) is basically a search problem, then one may still want to find out *where* and *how* to search. While the space, i.e. the *semantic map* in which search takes place is a *resource problem*,— any of the following could be used: dictionary, corpus, thesauraus, etc. or a mix of them,— its exploration is typically a *search problem*. Important as it may be, the building of a high quality resource is not the focus of this work, we rely on an existing one, and while we are concerned with its quality, we will be mostly concerned here with search methods, in order to determine the best.

1 Problem: find the needle in a haystack

One of the most vexing problems in speaking or writing is that one knows a given word, yet one fails to access it when needed. This kind of search failure, often referred to as *dysnomia* or *Tip of the Tongue-problem*, occurs not only in communication, but also in other activities of everyday life. Being basically a search problem it is likely to occur whenever we look for something that exists in real world (objects) or our mind: dates, phone numbers, past events, peoples' names, or you just name it.

As one can see, we are concerned here with the problem of words, or rather, how to find them in the place where they are stored: the human brain, or an external resource, a dictionary. Our work being confined to lexical access, we would like to develp a *semantic map* and a *compass* to help language producers to find the word they are looking for. More precisely, we try to build an index and a navigational tool allowing people to access words no matter how incomplete their conceptual input may be. Our approach is based on psychological findings concerning the *mental lexicon* (Aitchison, 2003; Levelt et al., 1999), i.e. *storage* and *access* of information in the human mind, observed *search strategies* and typical *navigational behavior*.

2 Consider the following elements before attempting an engineering solution

Before conceiving a roadmap leading to an engineering solution it may be useful to consider certain points. The list here below is by no means complete, neither is the following discussion. Nevertheless we believe that the following points are worth consideration: features of the mental lexicon, how to build and use the resource, searching, ranking and weights, interface problems. For reasons of space constraints we will touch briefly only upon some of these points.

Our main goal is the enhancement of electronic dictionaries to help speakers or writers to find

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 75–84,
Beijing, August 2010

quickly and intuitively the word they are looking. To achieve this target we take inspiration in the findings concerning the human brain (structure, process) when it tries access words in the mental lexicon.

2.1 The *mental lexicon*, a *small-world network*?

While *forms* (lemma) and *meanings* (lexical concepts, definitions) are stored side by side in paper dictionaries (holistic presentation), the human brain stores them differently. The information concerning meaning, forms and sound is distributed across various layers. Lexical fragmentation or information distribution is supported by many empirical findings,[1] and while this fact is arguably the reason accounting for word access problems, it is probably also the explanation of the power and the flexibility of the human mind which generally manages to find in no time the right term after having searched for it in a huge store of words.

While it is still not entirely clear what is stored, or whether anything is stored at all [2] coming close to the kind of information generally found in dictionaries, it does seem clear though that the structure of mental lexicon is a multidimensional network in which the user navigates. "Entries in the lexicon are not islands; the lexicon has an internal structure. Items are connected or related in various ways...There are item relations *within* and *between* entries." (Levelt, 1989). While the former relate *meanings* and *forms*: syntactic (part of speech), morphological, phonological information, the latter connect lexical entries.[3] In sum, lexical networks store or encode the information people typically have with regard to words, and finding the needed information, amounts to enter the graph at some point,– in the case of writing or speaking, usually a node dedicated to meaning,– and to follow the links until one has reached the goal (target word). While computer scientists call this kind of search 'navigation', psychologists prefer the term 'activation spreading. While not being exactly the same, functionally speaking they are equivalent though.

As every day language experience shows, things may go wrong, we lack information, hence we get blocked. Yet when trying to complete the puzzle we do not start from scratch, we rely on existing information, which, in terms of the network metaphor means that we start from (information underlying) a word being close to the target word.[4]

It is interesting to note, that our lexical graphs seem to have similar characteristics as *small-world networks*. These latter are a type of graph in which most nodes, eventhough not being direct neighbors, can be reached via a small number of clicks, about 6, regardless of the starting point. This property of networks, where objects, or the nodes standing for them, are highly connected has first been described by Frigyes Karinthy (1929) a Hungarian writer, to be tested then many years later by a social psychologist (Milgram, 1961). Nodes can be anything, people, words, etc. If they represent people, than edges specify their relationship, i.e. the very fact that they know each other, that they are friends, etc. Given this high connectivity, anything seems to be at the distance of a few mouse clicks. Hence, it is easy to connect people or to find out who entertains with whom what kind of relationship. Obviously, there is a striking similarity to our lexical graphs, and the small-world feature has been tested by mathematicians, who concluded that the distance for words is even smaller than in the original Milgram experiments, namely 4 rather than 6. Indeed, (Motter et al., 2002) and colleagues could show that more than

[1] *Speech errors* (Fromkin, 1980), studies on *aphasia* (Dell et al., 1997; Blanken et al., 2004) or *response times* i.e. *chronometric studies* (Levelt et al., 1999), *neuroimaging* (Shafto et al., 2007; Kikyo et al., 2001), *eye movements*, (Roelofs, 2004), experiments on *priming* (Schvaneveldt et al., 1976) or the *tip of the tongue problem* (TOT) (Brown and McNeill, 1996).

[2] An important feature of the mental lexicon lies in the fact that the entries are not *accessed* but *activated* (Marslen-Wilson, 1990; Altmann, 1997). Of course, such a detail can have far reaching consequences concerning knowledge representation and use, i.e. structure and process.

[3] These are typically the kind of relations we can find in WordNet (Fellbaum, 1998), which happens to be quite rich in this respect, but relatively poor with regard to intrinsic, i.e. intralexical information.

[4] As TOT experiments have shown (Brown and McNeill, 1996), people always know something concerning the target word (meaning, form, relation to other words), hence finding a word in such a situation amounts to puzzle-completion.

99 percent of the word pairs of their corpus could be connected in 4 steps at the most.

2.2 Building the resource

There are two elements we need to get a clearer picture of: the nature of the *resource* (semantic map), and the *search method* i.e. the way to explore it. Concerning the resource, there are many possible sources (dictionary, thesaurus, corpora, or a mix of all this) and many ways of building it. Since our main goal is the building of an index based on the notion of word relations (triples composed of two terms and a link), the two prime candidates are of course *corpora* and *association lists* like the ones collected by psychologists. While the former are raw data, containing the information in a more or less hidden form, the latter (often) contain the data explicitely, but they are scarce, subject to change, and some of the links are questionable.[5]

Corpora: Concerning the resource the following points deserve consideration: *size*, *representativity* and *topic sensitivity*.

- *Size or coverage*: While size or coverage are critical variables, they should not be overemphasized though, trading quantity against quality. We need to define the meaning of quality here, and whether, when or how lack of quality can be (partially) compensated by quantity. In other words, we need to define thresholds. In the absence of clear guidelines it is probably wise to strive for a good balance between the two, which again assumes that we know what quality means.

- *Representativity*: Obviously, the system we have in mind is only as good as the data we use, i.e. the purity/accuracy and representativity of the word/feature-association lists. No single set of data (dictionary, corpus, thesaurus) will ever suffice to capture the knowledge people have. While it would be unrealistic to try to model the semantic map of everyone, it is not unreasonable to try to reach an average user, say someone who has been to school and is a computer literate. If we want to capture the world-knowledge of this kind of user (target), than we must beware that it is contained in the material we use, since our resource will be based on this data. Hence, taking as corpus only the newspapers read by an elite (say, *Le Monde*, in France), will surely not suffice to capture the information we need, as it will not relate information ordinary citizens, say sport fans, are familiar with or interested in. In sum, we need to take a wide variety of sources to extract then the needed information. While there is shortage of some document types needed, there are nevertheless quite a few sources one may consider to begin with: Wikipedia, domain taxonomies, topic signatures, (Lin and Hovy, 2000), a database like (http://openrdf.org), etc.

- *Topic sensitivity*

  Weights are important, but they tend to change dynamically with time and the topic. Think of the word 'piano' uttered in the contexts of a 'concert' or 'household moving'. It is only in this latter case that this term evokes ideas like *size* or *weight*. The dynamic recompution of weights as a function of topic changes requires that the system be able to recognize the topic changes, as otherwise it might mislead the user by providing of inadequate weights. For some initial work see (Ferret and Zock, 2006).

Association lists: Psychologists have built such lists already decades ago (Deese, 1965; Schvaneveldt, 1989). Similar lists are nowadays freely available on the web. For example, for English there is the Edinburgh Associative Thesaurus [6] and the compilation done by Nelson and his colleagues in Florida [7]. There are also some re-

[5] This flaw is due to the experimental protocol. Subjects are asked to give the first word coming to their mind right after a stimulas. Not having been asked to specify the link it is the experimenter who does so. Yet, many word pairs,– say, cat and dog,– allow for various links (love, tease, chase, etc.), and it is not obvious at all which is the one intended by the user. This problem could have been avoided to a large extent if the instruction had been, "build a *sentence* containing the following word". Another potential problem may be due to the distance between the source and the target word: the link may be mediated.

[6] http://www.eat.rl.ac.uk/

[7] http://cyber.acomp.usf.edu/FreeAssociation/

sources for German (see [8] or [9]), for Japanese,[10] and probably many other languages.

While association lists are generally built manually, one can also try to do so automatically or with the help of people (see section 5 in (Zock and Bilac, 2004)). JeuxdeMot (JdM), a collectively built resource focusing on French being an example in case.[11]

2.3 Searching

The goal of searching is more complex than one might think. Of course, ultimately one should find the object one is looking for,[12] but the very process should also be carried out quickly and naturally. In addition we want to allow for recovery in case of having taken the wrong turn, and we want to avoid looping, that is, walking in circles, without ever getting closer to the goal. Last, but not least we want to make sure that stored information can also be accessed.

That this is less obvious than it might seem at first sight has been shown by (Zock and Schwab, 2008). Taking two resources (WN and Wikipedia) that contain both a given target word, we wanted to see whether we could access it or not. The target word was ‹*vintage*›. In order to find it we provided two access keys, i.e. trigger words: ‹*wine*› and ‹*harvest*›. Combining the two produced a list of 6 items in the case of WN and 45 in the case of Wikipedia, yet, while the latter displayed the target word, it was absent from the list produced by WN. This example illustrates the fact that our claim concerning storage and acess is well founded. Having stored something does by no means guarantee its access.

In the next sections we will present a small experiment concerning search.

3 System architecture

To allow for word access, we need at least two components: an index, i.e. a resource, representing or encoding the way words are connected (database or semantic network encoding associative relations between words) and an efficient search algorithm to find the needed information, in our case, words.

In other words, since search requires a map or a resource in which to search and a good algorithm to perform the search, we are keen in finding out how different resources (for example, Wikipedia, WordNet or JeuxdeMots) and various search algorithms might affect efficiency of word access. While there is a link between (the quality of) the resource and the searching, we will separate the two, focusing here mainly on the search algorithms and possible ways to evaluate them.

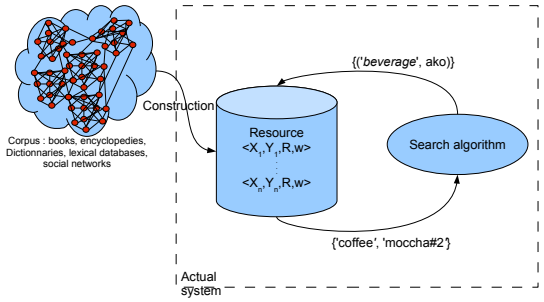


Figure 1: *Overall architecture of the system: the resource (association matrix) and a set of search algorithms*

4 Corpora and resources

Resources are built from corpora which can be of various kinds: news, books, social media, encyclopedias, lexical databases,... They can be general or specific, representing a particular domain. 'Text genre' may of course have an impact on what we can expect to retrieve. Obviously, one will not take a database of stock exchange news if one is looking for words referring to tennis- or fishing equipment.

To build our resource, we relied on WordNet (WN).[13] The resource can be seen in various ways: as a semantic network, an association

[8] http://www.schulteimwalde.de/resource.html

[9] http://www.coli.uni-saarland.de/projects/nag/

[10] http://www.valdes.titech.ac.jp/ terry/jwad.html

[11] http://www.lirmm.fr/jeuxdemots/rezo.php

[12] This poses special requirements concerning the organization, indexing and ranking of the data, i.e. words. We will not get into these issues here.

[13] Note, that one may consider WN (Fellbaum, 1998) not only as a dictionary, but also as a corpus. Actually we used precisely this kind of corpus for building our resource.

matrix, or as a list or database of 4-tuples composed of terms, links and weights. These elements can be represented in the following way, $<$ X,Y,R_Z,w$>$, where X and Y are terms or arguments of a given link, whose name expresses the type of relationship holding between them [R_Z (synonyme, antonyme, hyperonyme, collocation,...)]. Links and terms have a certain weight w which can be crucial for navigation and information display (interface). Words may be grouped into clusters,[14] and the clusters as well as their elements may be presented in the descending order of the weight: frequent terms being shown on top of the list.

Weights can be calculated in various ways: mutual information, co-occurences, etc. For example, for corpora, they can be seen as the number of times two items co-occur in a given window (sentence, phrase, paragraph, n words before or after, ...). Unfortunately, this kind of information is not always available in current resources. For example, WN relates terms (or senses), but does not assign them any particular weight. Yet, this information is very important and might be added via some learning method.

5 Search Algorithms

5.1 Definitions

Informally, a search algorithm is a method allowing to retrieve a list of terms (candidate target words presented in a given order) from a list of pairs containing the cue- or trigger words and their relations. For example figure 1 shows that the pair (‹*beverage*›, `ako`) allows the retrieval of {‹*coffee*› and ‹*moccha#2*›}, two potential target words. More formally, let's define

$$f(\{(t_1, R_1), (t_2, R_2), \ldots, (t_m, R_m)\}) = \{(T_1, w_1), (T_2, w_2), \ldots, (T_m, w_m)\} \quad (1)$$

where t_1,t_2,...,t_m are keys, cue- or *trigger-words*, R_1,R_2,...,R_m the *type of relation*, T_1,T_2,...,T_m candidate *target-words* and w_1, w_2,...,w_m, the associated *weights*. The curley brackets {} represent the fact that we have an ordered set.

[14] All words having the same link will be stored and presented together. For example, 'cat' and 'dog' are likely to fall in the category 'animal', while 'hammer' and 'screwdriver' will fall in the category 'tools'.

Indeed, the 'trigger word-relation' pairs are ordered, that is, they parallel the order in which these terms were given as input at query time. The candidate target words are ordered in terms of confidence, ranking which may vary with respect to a given search algorithm. In this paper, confidence is based on the weights in the resource. Of course, one could imagine other ways to define or compute it.

Note that we can have $R_1 = R_2$, provided that we do not have at the same time $t_1 = t_2$. For instance, while it is possible to have {(‹*island*›, instance), (‹*island*›, `ako`)}, one cannot have at the same time {(‹*island*›, instance), (‹*island*›, instance)}

We present in the next sections various ways to use the resource and various search algorithms. In these experiments, we tried to use *direct* and *indirect links* (mediated associations) contained in the tuples and to establish linearly the weight as a function of the position of the word in the list of the trigger words.

5.2 General algorithm

In the general algorithm, we consider that our candidate *target words* are at the intersection of the sets corresponding to the pairs of *trigger words* and their *relations*.

$$f(\{(t_1, R_a), (t_2, R_b)\}) = f(\{(t_1, R_a)\} \cap f(\{t_2, R_b\})) \quad (2)$$

We will now show, step by step, how $f(\{(t, R)\})$ is affected by various uses (direct vs. indirect use) and orderings. This will yield 4 kinds of search algorithms.

5.3 The use of the tuples

To illustrate our algorithms, let us consider the following resource:

$<$ ‹*mouse*›,‹*rodent*›,`ako`,3$>$;$<$ ‹*rodent*›,‹*animal*›,`ako`,4$>$;
$<$ ‹*rat*›,‹*rodent*›,`ako`,1$>$;$<$ ‹*rat*›,‹*animal*›,`ako`,2$>$;

5.3.1 Direct use

In this case, we rely only on the direct links $<$ t,T,R,$w$$>$ of the resource Res:

$$f(\{(t, R)\}) = \{(T, W) | < t,T,R,\text{w} > \in Res\} \quad (3)$$

that is, in the case of direct use, the search algorithm fed with the trigger word t and the relationship R found all target words T contained in the tuple $< t,T,R,\text{w}>$ of the resource Res. The computation of the weight W is defined in 5.4. For example, ‹*mouse*›, would yield ‹*rodent*›, while ‹*rat*›, would trigger ‹*animal*› and ‹*rodent*›.

5.3.2 Indirect use

In order to boost recall this algorithm takes also indirect links into account.

$$f(\{(t,R)\}) = \{(T,W)|< t,T,R,\text{w}> \in Res\} \cup \{(T,W)|< t,X,R,w_1> \in Res \text{ and } < X,T,R,w_2> \in Res\} \quad (4)$$

Hence, we consider neighbor words of the neighbors of the trigger words.[15] Again, for ‹*mouse*›, we get ‹*rodent*› and ‹*animal*›, while for ‹*rat*›, we continue to get ‹*animal*› and ‹*rodent*›.

5.4 Weighting

5.4.1 Basic Weighting

$$\text{For } f(\{(t_1,R_1)\},\{(t_2,R_2)\},\ldots,\{(t_n,R_n)\}), \quad W = \Sigma_{t_i,T,R_i,w_j} w_j \quad (5)$$

In our example and for *direct use*, if our trigger word list is {‹*mouse*›, ‹*rat*›} then the weight (W) will be 4 (3 + 1) for ‹*rodent*› and 6 (4 + 2) for ‹*animal*›. In the case of *indirect use*, the weight will be 4 (3 + 1) for ‹*rodent*› and 10 (3 + 4 + 1 + 2) for ‹*animal*›.

5.4.2 Weighting based on the cue-word's *position* and its *relation* to other words

Let us suppose that the user gave in this order the following cue words ‹*A*›, ‹*B*›, ‹*C*›. In this case we assume that ‹*A*› is more important than ‹*B*›, which is more important than ‹*C*› in order to find the target word. Following this line of reasoning, we may consider the following cases:

$$\text{For } f(\{(t_1,R_1)\},\{(t_2,R_2)\},\ldots,\{(t_n,R_n)\}), \quad W = \Sigma_{t_i,T,R_i,w_j}(n-i+1)\times w_j \quad (6)$$

[15] Please note, we consider here only one intermediate word (two links), as this is, computationally speaking, already quite expensive.

In our example of direct use, if our trigger word list is {‹*mouse*›, ‹*rat*›}, W will be 7 ($2\times 3+1\times 1$) for ‹*rodent*› while W = 10 ($2\times 4+1\times 2$) for ‹*animal*›. For indirect use, W is 4 ($2+1\times 1$) for ‹*rodent*› and 17 ($2\times(3+4)+1\times(1+2)$) for ‹*animal*›.

5.5 Proposed Search Algorithms

Crossing the characteristics of *weight* (direct vs. indirect) and *use* (direct vs. indirect), we get 4 possible *search methods* : direct use with *basic* weighting (A1) or *linear* weighting (A2); and *indirect* use with *basic* weighting (A3) or *linear* weighting (A4).

6 Evaluation

6.1 The problem of evaluation.

The classical *in vivo / vitro* approaches do not seem to fit here. While the former tests the system for a given application, the latter tests the system independently. Given the fact that our system has several components, we can evaluate each one of them separately. More precisely, we can evaluate the quality of the *resource* and/or the quality of the *search algorithm*. We will focus here on the search method.

6.2 Procedure

The basic idea is to provide each algorithm with an ordered set of *trigger words* and to see how many of them are generally needed in order to reveal the *target word*.

Another way to evaluate the quality of the search mechanism is to check at each step the *position* of the target word in the list generated by the algorithms (output).

6.3 Building the test corpus

Psychologists have studied the differences of monolingual and bilingual speakers experiencing the 'tip-of-the-tongue' problem (Pyers et al., 2009). Their experiments were based on 52 pictures corresponding to low-frequency names. Starting from this list we were looking for associated words. In order to build this list we used as resource the results produced by 'Jeux de Mots' (Lafourcade, 2007).[16]

[16] As mentionned by one of the reviewers, we could and probably should have used the Edinburgh Associative The-

6.3.1 JeuxdeMots (JdM)

JeuxdeMots, meaning in English 'word games' or 'playing with words', is an attempt to build collaboratively, i.e. via a game, a lexical resource. The goal is to find out which words people typically associate with each other and to build then the corresponding resource, that is, a lexical-semantic network. What counts as a *typical association* is established empirically. Given some input from the system –(term and link, let us say 'Americans' and 'elect as president')– the user produces the associated word –(second term, let us say 'Obama'),– answering this way the question, what term x is related to y in this specific way. Once the network is built, terms should be accessible by entering the network at any point (via some input) and by following the links until one reaches the target word. This is at least the theory. Unfortunately, in practice things do not always work so well.

Actually, JdM has several flaws, especially with respect to acces or search. The shortcomings are probably rooted in too heavy reliance on the notion of weight and in excessive filtering of the output, i.e. premature elimination of the candidates presented to the user, list of elements among which the user is meant to choose. Indeed, JdM presents only the highest ranked candidate. Hence, words may never make it to the critical level to be included in the set from which the user will choose the target word. Also, weights do not necessarily correlate with users' interests. This problem can be solved in interactive search, provided that the output contains a critical mass of candidates (possibly organized according to some point of view), but the problem is most likely remain if one presents only one candidate (the highest ranked term), as this latter is not necessarily the target, neither is it always a term from which one would like to continue search.

There is also a problem with the *link names*, i.e. (metalanguage),[17] though, to be fair, one must admit that identifying and naming links is a very difficult problem. Last, but not least, though more related to the quality of the resource than to the problem of search, there is a chance of user-bias. Indeed, it is not entirely clear whether people really give the first association coming to their mind, or the one fitting them best to continue the game and win more points.

Despite these shortcomings, we will use JdM as a resource as it exists not only for various languages, but is also quite rich, at least for French. Unfortunately, the English version is very poor compared to the French part.[18] This is why we've decided to use the French version for the test corpus.

6.3.2 Building the test corpus

Starting from Pyer's list, we translated each term into French and inspected then JDM in order to find the 10 most frequently connected words according to this resource. Next, we translated these terms into English, producing the list shown in the appendix.

As we will see later, in our experiment we do not have *typed relations* between the words. Actually we took from JDM what they call "associated ideas".

Nevertheless, when building the list we did have some problems. Some words do not have any, only one, or simply very few associated ideas. This is particularly true for low frequency words. This being so, we deleted them (in our case 7) from the list.

6.4 Description of the tool

Our tool is implemented in Java. To allow for online access [19] we use Google's Web Toolkits[20]. The interface is very simple, akin to Google's search engine. At the top of the page the user is invited to provide the input, i.e. the *trigger words*,

saurus (`http://www.eat.rl.ac.uk/`) as it contains authentic word associations collected from people. This point is well taken and we will consider this resource in the future as its coverage is better than our current one and it also avoids possible problems due to the translation. Though being generic, JeuxdeMots has mainly data on French, yet our tests were run on English.

[17] The term *typical association* is underspecified to say the least.

[18] For example, while for English JdM has by today (july 9, 2010) only 654 relations, the French part contained 1.011.632 the very same day, and 994.889 a month ago.

[19] `http://getalp.imag.fr/homepages/schwab`

[20] http://code.google.com/intl/fr/webtoolkit/

a checkbox allows to choose *relations* and at the bottom are shown the candidate *target words*.

6.5 Description of the resource for the experiment

In this experiment, we use the English version of Wikipedia to build our resource. Due to corpus characteristics, only one relation is used: *neighbor* (ngh). We consider "words occuring in the same paragraph" as neighbours. After having deleted 'stop words' (articles, conjunction, ...) we lemmatize 'plain words' by using DELA? [21]

For example, a corpus containing the following two sentences "The cat eats the mouse \ The mouse eats some cheese" would yield the following resource :

< ‹cat›,‹mouse›,ngh,1>; < ‹cat›,‹eat›,ngh,1>; < ‹eat›,‹cat›,ngh,1>; < ‹eat›,‹mouse›,ngh,2>; < ‹mouse›,‹cat›,ngh,1>; < ‹mouse›,‹eat›,ngh,2>; < ‹mouse›,‹cheese›,ngh,1>; < ‹eat›,‹cheese›,ngh,1>; < ‹cheese›,‹mouse›,ngh,1>; < ‹cheese›,‹eat›,ngh,1>

It should be noted that in this experiment, links are symmetrical.

$$< \text{X,Y,ngh,w} > \rightarrow < \text{Y,X,ngh,w} > \qquad (7)$$

6.6 Comparison and evaluation of results

Due to time contraints, we decided to use only a small sample of words, 10 to be precise. Concerning search we have tested two parameters: the *scope* (direct vs. indirect links, i.e. associations, A1 vs. A3) and the *weight* (presence or absence, A2 vs. A4).

The results are shown in table 1, where $\emptyset$ means that the algorithm did not find any solution, while ∞ implies that the trigger word list has been fully exhausted without being able to produce the target word among the top ten candidates. Indeed, in order to be considered as a hit, the found target word has to be among the top ten.

As one can see our algorithms with indirect use (A3 and A4) never manages to find the target word. Actually, it does not fail totally. It is just that the candidate term appears very late in the list, too late to be considered. The algorithms with direct use (A1 and A2) do find the elusive word or produce a 'list' of zero candidates.

[21] http://infolingu.univ-mlv.fr/DonneesLinguistiques/ Dictionnaires/telechargement.html

TARGET	A1	A2	A3	A4
hive	1	1	∞	∞
peackock	4	4	∞	∞
comet	3	2	∞	∞
microscope	$\emptyset$	$\emptyset$	∞	∞
snorkel	5	5	∞	∞
pitcher	4	3	∞	∞
axe	$\emptyset$	$\emptyset$	∞	∞
gazebo	$\emptyset$	$\emptyset$	∞	∞
hoe	$\emptyset$	$\emptyset$	∞	∞
castle	3	3	∞	∞

Table 1: *Comparison of the number of steps needed by each search algorithm to find the target word ie. to put it in the list of the top ten.* $\emptyset$ *signals the fact that the algorithm does not find any solution, while* ∞ *implies that eventhough the trigger word list has been totally used, it did not manage to come up with the target word among the top ten candidates.*

target	1	2	3	4	5	6
A1	115	112	325	$\emptyset$	$\emptyset$	$\emptyset$
A2	118	98	273	$\emptyset$	$\emptyset$	$\emptyset$
A3	256	288	254	189	114	59
A4	234	267	262	156	115	54

Table 2: *Comparison of the position of the target word at each step of the algorithm for* ‹microscope›

Table 2 illustrates this last point by showing the position of the *target word* with respect to one of the six *trigger words*.

While the *target word* always appears in A3-A4, A1 and A2 never produce any results beyond the 4th *trigger word*. The two experiments also show that *linear weighting* has hardly any effect on the results.

7 Conclusions and Future Work

We have started to caracterize lexical access as a search problem. Since search requires a resource in which to search and a good algorithm to perform the search, we were interested in establishing how different *resources* –(Wikipedia (WiP), WordNet (WN), JeuxdeMots (JdM))– and various *search algorithms* might affect efficiency of word access. The focus here has been on the latter.

Next to search algorithms, we presented some methods for evaluating them. While our results are clearly preliminary and on a very small scale, we believe that the questions we have raised are of the right sort. Of course, a lot more work is needed in order to answer our questions with more authority.

8 Appendix

hive (t_w): bee; honey; queen; cell; royal jelly; pollen; wax; group; frame; nest; (a_{ws})
peacock: bird; feather; animal; spread; shout; blue; tail; disdainful; arrogant; despise;
comet: star; space; shooting star; astronomy; sky; galaxy; night; universe; apparition;
microscope: small; observe; enlarging; microscopic; observation; ocular; optical; twin; eyeglass;glasses; sea; diving; breath; ocean; mask; flipper;
pitcher: jug; jar; carafe; dishes; vase; ewer; container;
axe: cut; kill; split; fell; murder; saw; agriculture; arboriculture;
gazebo: pavilion; platform; viewpoint; terrace;view; architecture; house; pavillon; esplanade
hoe (tool): farming; tool; shovel; pick; technique;spade;
castle: tower; king; dungeon; fort; queen; drawbridge; prince; princess; embrasure;
eclipse: moon; sun; astronomie; disappearance; darkness;
bolted joint: door; lock; padlock; key; close; button; metal; box; house; portal; bolt;
megaphone: sound; loudspeaker;
manta ray: fish; sea; wing;
wheelbarrow: wheel; carry; garden; shovel; gardener; fill; push;
dynamite: bomb; explosion; explosive; weapon;wick; chemistry; plastic;
compass: direction; navigation; navy; windrose;
chisels: cut; scissor; prune; pair; hairdresser; paper; school; chisel;
ostrich: egg; bird; Australia; cassowary; emu;
grater: woodwork; tool; poverty; polished; dishes;
braille: blind; alphabet; writing;
water well: dig; drill; pierce;
guillotine: scaffold; widow; death penalty; head; reaper; decapitation; execution; torture;
weathervane: rooster; direction; wind; rooftop; rotation; east; south; north; west;
churning (butter): container; oil; milk; bottle; container; cuve; jerrycan; tank;
carousel: fun fair; amusement park; children; entertainment;
canteen (place): meal; school; eat; dessert; dish; tableland; entrance; restaurant; food; refectory; supervisor; glass; wood; tail; tooth; trapper; trunk;
goggles: sight; glasses; myopia; eyes; rim; twin-lens; optician; sun; vision; see; rectification; improvement; astigmatic; ophthalmologist; farsighted; longsighted; blind; binoculars; nose; optical; pair; telescope; protection;
boomerang: Australia; object; flying; throw; come back;
easel: painting; drawing; support; tripod;
propeller: boat; ship; plane; propulsion; curve; rotation; rolling;
walnut: fruit; hazelnut; almond; tree; cashew; nutcracker; oil; salad; woodwork;
catapult: ejection; old weapon; throw; stone; aircraft carrier; crossbow; ballista; sling;
udder: milk; cow; chest; nipple tit;
gyroscope: direction; rotation; instruments; gyrostat;
mummy: pharaoh; Egypt; pyramid; strip; dead; embalm; sarcophagus; fruits; funeral;
hinge: junction; woodwork; middle; locksmiths; assemblage;
harmonica: music; instrument; breath; flute; musical instruments;
metronome: musical, tempo, musical instruments;
noose: hang; boat; attach; bind; cord;
harp: musical instruments; zither; lute; lyre; psaltery;
slingshot: weapon; attack; projectile weapon; catapult;
eiffel tower: Paris; steel; monument; metal;
syringe: drug; injection; sting; nurse; sick; ill; bodycare; drug addiction;

Words containing too little information to be usable for tests baster, unicycle, thermos,

antlers, plunger, cleftchin, handcuffs.

References

Aitchison, J. 2003. *Words in the Mind: an Introduction to the Mental Lexicon (3d edition)*. Blackwell, Oxford.

Altmann, G. T. M., 1997. *The ascent of Babel: An exploration of language, mind, and understanding*, chapter Accessing the Mental Lexicon: Words, and how we (eventually) find them. Oxford University Press.

Blanken, G., F. Kulke, T. Bormann, B. Biedermann, J. Dittmann, and C.W. Wallesch. 2004. The dissolution of word production in aphasia: Implications for normal functions. In Pechmann, T. and C. Habel, editors, *Multidisciplinary Approaches to Language Production*, pages 303–338. Mouton de Gruyter, Berlin.

Brown, R. and D. McNeill. 1996. The tip of the tounge phenomenon. *Journal of Verbal Learning and Verbal Behaviour*, 5:325–337.

Deese, J. 1965. *The structure of associations in language and thought*. Johns Hopkins Press.

Dell, G.S., M.F. Schwartz, N. Martin, E.M. Saffran, and D.A Gagnon. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4):801–838.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database and some of its Applications*. MIT Press.

Ferret, O. and M. Zock. 2006. Enhancing electronic dictionaries with an index based on associations. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 281–288.

Fromkin, V. 1980. Errors in linguistic performance: Slips of the tongue, ear, pen and hand.

Kikyo, H., K. Ohki, and K. Sekihara. 2001. Temporal characterization of memory retrieval processes: an fMRI study of the tip of the tongue phenomenon. *European Journal of Neuroscience*, 14(5):887–92.

Lafourcade, M. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand.

Levelt, W., A. Roelofs, and A. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–75.

Levelt, W. 1989. *Speaking : From Intention to Articulation*. MIT Press, Cambridge, MA.

Lin, C-Y and E. H. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING*, pages 495–501. M. Kaufmann.

Marslen-Wilson, W.D. 1990. Activation, competition, and frequency in lexical access. In Altmann, G.T.M., editor, *Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives*, pages 148–172. MIT Press, Cambridge, MA.

Milgram, S. 1961. The small world problem. *Psychology Today*, 1(1):61–67.

Motter, A. E., A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta. 2002. Topology of the conceptual network of language. *Physical Review E*, 65(6).

Pyers, J. E., T. H. Gollan, and K. Emmorey. 2009. Bimodal bilinguals reveal the source of tip-of-the-tongue states. *Cognition*, 112(2):323 – 329.

Roelofs, A. 2004. The seduced speaker: Modeling of cognitive control. In Belz, A., R. Evans, and P. Piwek, editors, *INLG*, volume 3123 of *Lecture Notes in Computer Science*, pages 1–10. Springer.

Schvaneveldt, R., D. Meyer, and C. Becker. 1976. Lexical ambiguity, semantic context and visual word recognition. *Journal of Experimental Psychology/ Human Perception and Performance*, 2(2):243–256.

Schvaneveldt, R., editor. 1989. *Pathfinder Associative Networks: studies in knowledge organization*. Ablex, Norwood, New Jersey, US.

Shafto, M. A., D. M. Burke, E. A. Stamatakis, P. P. Tam, and L. K. Tyler. 2007. On the tip-of-the-tongue: Neural correlates of increased word-finding failures in normal aging. *J. Cognitive Neuroscience*, 19(12):2060–2070.

Zock, M. and S. Bilac. 2004. Word lookup on the basis of associations : from an idea to a roadmap. In *Workshop on 'Enhancing and using electronic dictionaries'*, pages 29–35, Geneva. COLING.

Zock, M. and D. Schwab. 2008. Lexical access based on underspecified input. In *COGALEX, Coling workshop*, page 8, Manchester.

Author Index