Automatically extracting Translation Links using a wide coverage semantic taxonomy

German Rigau[1], Horacio Rodríguez and Jordi Turmo.

Departament de Llenguatges i Sistemes Informàtics.
Universitat Politècnica de Catalunya.
Pau Gargalló 5, 08028 Barcelona, Spain.
telf.: 34 3 4017293 fax: 34 3 4017039
g.rigau@lsi.upc.es
horacio@lsi.upc.es
turmo@goliat.upc.es

Abstract

TGE (Tlink Generator Environment) is a system for semi-automatically extracting translation links. The system was developed within the ACQUILEX II[2] project as a tool for supporting the construction of a multi-lingual lexical knowledge base containing detailed syntactic and semantic information from MRD resources. A drawback of the original system was the need of human intervention for selecting the more appropriate translation links in the case where more than one were extracted and proposed by the system. This paper deals with the task of overcoming this drawback. What is presented is an heuristic method based on conceptual distance that uses information from an external wide-coverage semantic taxonomy (WordNet). Our aim is to overcome the problem in an automatic way or to provide the user with complementary information in order to make his/her choice easier.

1 Motivation and introduction

The research reported here is part of the ACQUILEX II project which has as one aim the automatic or semi-automatic construction of fragments of a multi-lingual lexical knowledge base containing detailed syntactic and semantic information from MRD resources.

Lexical information acquisition is generally considered as a major 'bottleneck' in NLP. It is clear that techniques which either partially or totally automate this process should be investigated. The use of monolingual machine-readable versions of conventional dictionaries (MRDs) in the acquisition of lexical knowledge has been widely extended because they provide substantial quantities of lexical information that can be extracted with limited difficulty and limited human intervention [Dolan et al. 93]. But less attention have been paid to bilingual MRDs. [Tanaka & Umemura 94] use a third language to construct a bilingual dictionary and [Knight & Luk 94] explains a simple approach to connect Spanish nouns extracted from a bilingual dictionary to a large-scale knowledge base.

During the early steps of ACQUILEX[3] project the aims were to build semi-automatically large scale monolingual lexicons represented as Lexical Knowledge Bases (LKB) from lexical databases (LDB) automatically extracted from MRDs (see [Ageno et al. 92], [Copestake 92]). In such a way, lexicons covering restricted domains (food, drink, ...) were built for four Languages (Dutch, English, Italian and Spanish).

Further on we concentrated on using the LKB to represent multilingual information in the form of links between monolingual lexical entries, which we refer to as tlinks (translation links). and on developing a methodology for constructing tlinks from the available MRD resources. The fundamentals of the system, as well as the representational issues, can be seen in [Copestake et al. 94] and [Copestake et

[1]*This researcher has been suported by a grant of the Ministerio de Educación y Ciencia, 92-BOE-16392.*
[2]*AcquilexII EC Esprit project BRA 7315.*
[3]*Acquilex EC Esprit project BRA 3030*

al. 92]. The description of a software environment, Tlink Generation Environment (TGE), supporting the extraction methodology can be seen in [Ageno et al. 94].

Several experiments have been carried out following this methodology and supported by TGE. The results have been summarised, reported and commented in [Copestake et al. 94] and [Ageno et al. 94]. The core of our methodology is the use of bilingual dictionaries as a main Knowledge Source for extracting tlinks. The fitness of the extracted tlinks depends heavily on the quality and coverage of such dictionaries. One of the drawbacks of the system was the need of huge specialised human intervention for selecting the more appropriate tlink in the case where more than on was allowable. The present paper makes a proposal for avoiding this intervention or for providing the user with complementary information in order to make easier his/her choice. The proposal is based on the use of a conceptual distance between the alternatives. The base for computing this distance is the use of an external wide-coverage semantic taxonomy for English (WordNet).

The organisation of this document is the following: After this introduction, in section 2, a short description of the underlying concepts and methodology is presented as well as a summary of the results obtained from their application. Section 3 will introduce our proposal of conceptual distance. On section 4 the way this concept can be applied for overcoming the drawbacks detected in section 2 is presented. The results of an experiment applying these ideas are presented too. Finally, some conclusions are stated in section 5.

2 Background

In our approach, the basic units for defining lexical translation equivalence are the lexical entries in the monolingual LKBs, which should, in general, correspond to word senses in the dictionary. Although in the simplest cases we can consider the lexical entries themselves as translation equivalent, in general, more complex cases occur corresponding to lexical gaps, differences in morphologic or lexical features, specificity, etc.

We represent the relationships between words in terms of tlinks. The tlink mechanism is general enough to allow the monolingual information to be augmented with translation specific information, in a variety of ways.

LKB formalism uses a typed feature structure (FS) system for representing lexical knowledge. We can, so, define tlinks in terms of relations between FSs. Lexical (or phrasal) transformations in both source and target languages are a desirable capability so that we can state that a tlink is essentially a relationship between two rules (of the sort already defined in the LKB) where the rule inputs have been instantiated by the representations of the word senses to be linked.

Figure 1 presents an example of tlink between the English entry *furniture* and the Spanish entry *muebles*, resulting from applying the lexical rule *plural* to the lexical entry *mueble*.

As any other LKB object, a tlink can be represented as a feature structure. The type system mechanism, in LKB, allows further refinement and differentiation of tlink classes in several ways. A **simple-tlink** is applicable whenever two lexical entries which denote single place predicates (nouns, etc.) are straightforwardly translation equivalent, without any previous transformation. The example presented in figure 1 belongs to this class. A **partial tlink** is applicable when we want to transfer the qualia structure from one sense to another. An example of this class is the Spanish entry *rioja*. There is no direct correspondence between this word and any English one because of the absence of such entry in the bilingual dictionary. We can however link the genus term, *vino*, to the corresponding English term *wine*, transferring to the later all the qualia structure from the former (and specially the *origin_area = Rioja*). In this way *rioja* can be roughly translated to English as a *wine with origin_area = Rioja*. Finally, the **phrasal tlink** is necessary when we need to describe a single translation equivalence with a phrase. *Ahumado*, for instance, must be linked to *smoked food*.

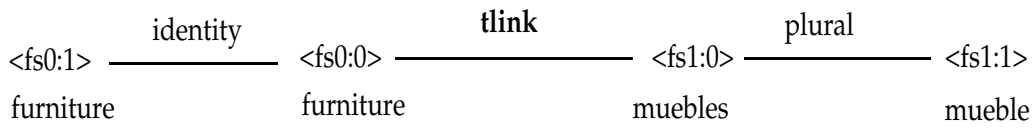

Figure 1: A tlink between furniture and muebles.

The establishment of tlinks can be performed, of course, manually, but the multiplicity of possible cases and the existence of several Knowledge Sources, KSs, (such as bilingual dictionaries, monolingual LDBs, or a multilingual LKB) allows and motivates the (partial) automatization of the process. To help in performing such a task we have developed an interactive environment: **TGE**. This environment is described in [Ageno et al. 94].

As we said before, TGE is a tool designed for supporting a tlink extraction methodology. The core of the methodology is the use of a bilingual dictionary as a main KS. Depending on the characteristics of the dictionary entry (or on its absence) different kinds of tlinks with different degree of fitness can be extracted. An important consideration is that in spite of using a bilingual dictionary as KS what we are linking are not words but lexical entries, placed in the LKB and owning not only orthographic information but also lexical information, basically the qualia structure, both local and inherited.

The way of organising the extraction process is by means of the performance of a set of extraction modules, each one corresponding to a different kind of tlink, implemented as rulesets in a Production Rules Environment (PRE, see [Ageno et al. 94]).

By now up to seven modules have been implemented for dealing with different situations. These modules are the following: *Simple Tlink Module*, when there exists a direct translation of the source entry in the bilingual dictionary, *orthographic Tlink Module*, when in both languages the same word with exactly the same spelling is used, *Compound Tlink Module*, when the corresponding entry in the target lexicon is a composed one, being the target lexical entry made up of the concatenation of the two English words that appear in the bilingual entry, *Phrasal Noun Tlink Module*, when the translation is the concatenation of two other nouns, *Parent Tlink Module*, when the entry does not appear in the bilingual but its hyperonym in the taxonomy does, generating a partial tlink, *Grandparent Tlink Module*, performing in a similar way and, finally, *General Tlink Module*, when the translation appearing in the bilingual is composed of more than one word. from which the genus term must be extracted in order to be linked to the source entry.

Several experiments were carried out on narrow semantic domains (food, drinks,...) using as KSs the taxonomies corresponding to these domains, extracted from the MRDs Spanish VOX monolingual [Biblograf 87] and English LDOCE and a LDB extracted from the bilingual Spanish/English VOX-Harraps [Biblograf 92]. The main results are reported on [Copestake 92] and [Copestake et al. 94] and are summarised below.

The Spanish taxonomy of drink-nouns, extracted from VOX dictionary, consists of 235 noun senses, and has 5 levels. The English taxonomy of drink-nouns, extracted from LDOCE, consists of 192 noun senses. Going from Spanish to English 223 out of 235 drink nouns were correctly linked by means of different, often more than one, tlinks (95%). A total of 377 tlinks were extracted.

The main drawbacks of the system, as discussed in [Ageno et al. 94] were 1) the poor coverage of English entries (only 27%) partially explained by the limited coverage of the bilingual dictionary used and 2) the need of human intervention for selecting the appropriate tlinks. This second point will be addressed in the following section.

3 Using Conceptual Distance

In our previous approach all the tlinks extracted by means of the corresponding KSs (basically the bilingual dictionary and the LKB) were offered to the user in order to allow the selection of the appropriate ones. This process was relatively high time consuming and needed a knowledge of both source and target Languages by the user. Our proposal is to measure the conceptual distance between the lexical entry corresponding to the source language and the different lexical entries corresponding to the target language. Three modes of performance are then allowed to the user: 1) select automatically the most feasible, 2) select automatically all the tlinks over a determined threshold and 3) rank the tlinks and allow the user to make the selection manually.

Several measures of relatedness among words based in the coocurrence of them in a text have been described; mutual information, t-test, association ratio, etc. [Church et al. 91], the cosine function in Context Space [Schütze 92], conditional probability [Wilks et al. 93], etc. We think, however, that in our case, taking into account the characteristics of source material an approach based on pre-existing class-based conceptual knowledge could be better. [Resnik 93] combines a knowledge based approach involving semantic classes taken from WordNet with coocurrence data extracted from corpora. Less attention has been paid lately to measures of relatedness based on semantic structured hierarchical nets.

We have selected a measure based on the relation of concepts in a structured hierarchical net (see [Rigau 94] for details). We use WordNet as support for the measure because of 1) the lack of similar conceptual base for Spanish and 2) the availability, huge coverage and quality of WordNet.

WordNet is an on-line lexicon based on psycholinguistic theories [Miller 90] that attempts to organise lexical information in terms of word meanings, rather word forms. In that respect, it resembles a thesaurus more than a dictionary. It comprises nouns, verbs, adjectives and adverbs, highly organised in terms of their meanings around semantic relations, which include among others, synonymy and antonymy, hypernymy and hyponymy, meronymy and holonymy. Lexicalised concepts, represented as sets of synonyms called synsets, are the basic elements of WordNet.

A measure of the relatedness among concepts can be a valuable prediction knowledge source to several decisions in Natural Language Processing. Relatedness can be measured by a fine-grained conceptual distance (as claimed [Miller & Teibel 91]) among concepts in a hierarchical semantic net such as WordNet. This measure would allow to discover the most reliable lexical cohesion of a given set of words in English.

Conceptual distance tries to provide a basis for determining closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is defined in [Rada et al. 89] as the length of the shortest path that connects the concepts in a hierarchical semantic net. Besides applying conceptual distance in a medical bibliographic retrieval system and merging several semantic nets, they demonstrate that their measure of conceptual distance is a metric. In a similar approach, [Sussna 93] employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. Following these ideas, [Agirre et al. 94] describes a new conceptual distance formula for the automatic spelling correction problem and [Rigau 94], using this conceptual distance formula, presents a methodology to enrich dictionary senses with semantic tags extracted from WordNet. The conceptual distance proposed by [Agirre et al. 94] is:

$$CD(c_1,c_2) = \sum_{i \in shortestpath(c_1,c_2)} \frac{1}{depth(c_i)} \quad \textbf{(1)}$$

The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.

But also:

- the measure should be independent of the number of concepts we are measuring.

The conceptual distance described in (1) only holds the first two conditions. This formula expresses that the Conceptual Distance between two concepts depends on the length of the shortest path that connects them and the specificity of the concepts in the path. That is to say, the lower the concepts are in a hierarchy, the closer they seem to be. For the purposes of the work presented here, this simple formula (is not sensitive to the local density in WordNet and only can be applied to pairs of concepts) discriminates between pairs of possible synset candidates to tlinks formation, selecting the most feasible ones.

4 Using Conceptual Distance for Disambiguating Tlinks

The nominal part of WordNet 1.5 has 60557 synsets and 87642 English nouns (76127 monosemous) in 107424 connections. That is, the polysemous ratio is 1.23 synsets per English noun and the synonymy degree is 1.77 English noun per synset. The Spanish/English bilingual dictionary contains 12370 Spanish nouns and 11467 English nouns in 19443 connections among them. On the other hand, the English/Spanish bilingual dictionary is less informative than the other one containing only 10739 English nouns, 10549 Spanish nouns in 16324 connections. Merging both dictionaries a list of equivalence pairs of nouns have been obtained. The combined dictionary contains 15848 English nouns, 14879 Spanish nouns and 28129 connections. For instance, for the word "masa" in Spanish the following list of equivalence pairs can be obtained:

---------- English/Spanish
bulk masa
dough masa
mass masa
---------- Spanish/English
cake masa
crowd_of_people masa
dough masa
ground masa
mass masa
mortar masa
volume masa

From the combined dictionary, there are only 12665 English nouns placed in WordNet 1.5 which represents 19383 synsets. That is, the maximum coverage we can expect WordNet1.5 using both bilingual Spanish/English dictionaries is 32%, taking into account that can be different sources of errors (e.g. there are no correct translation in the bilinguals, there are not the correct sense in WordNet, etc.).

The TGE environment using Conceptual Distance proceeds the creation of tlinks among lexical entries placed into the LKB and synsets in WordNet in a top-down fashion. Starting from the top lexical entry of the Spanish taxonomy the specialist selects the most feasible synsets of WordNet from those proposed by the rulesets using the bilingual dictionaries. Once the specialist has selected the equivalent synsets of the Spanish lexical entry in WordNet no further selection by the user is required. Then the program recurs applying the TGE rulesets for the hyponym lexical entries of the Spanish taxonomy. The Conceptual Distance among the equivalence translations proposed by the TGE environment and those selected previously (normally hypernym synsets) is computed, selecting those more closer (a Conceptual Distance threshold can be used for selecting a set of feasible synsets), and so on. Applying the Conceptual Distance measure the tlinks proposed by one ruleset can also be rejected. In this situation the TGE control mechanism decides what other ruleset must be launched. The tlinks generation process is illustrated with the following example:

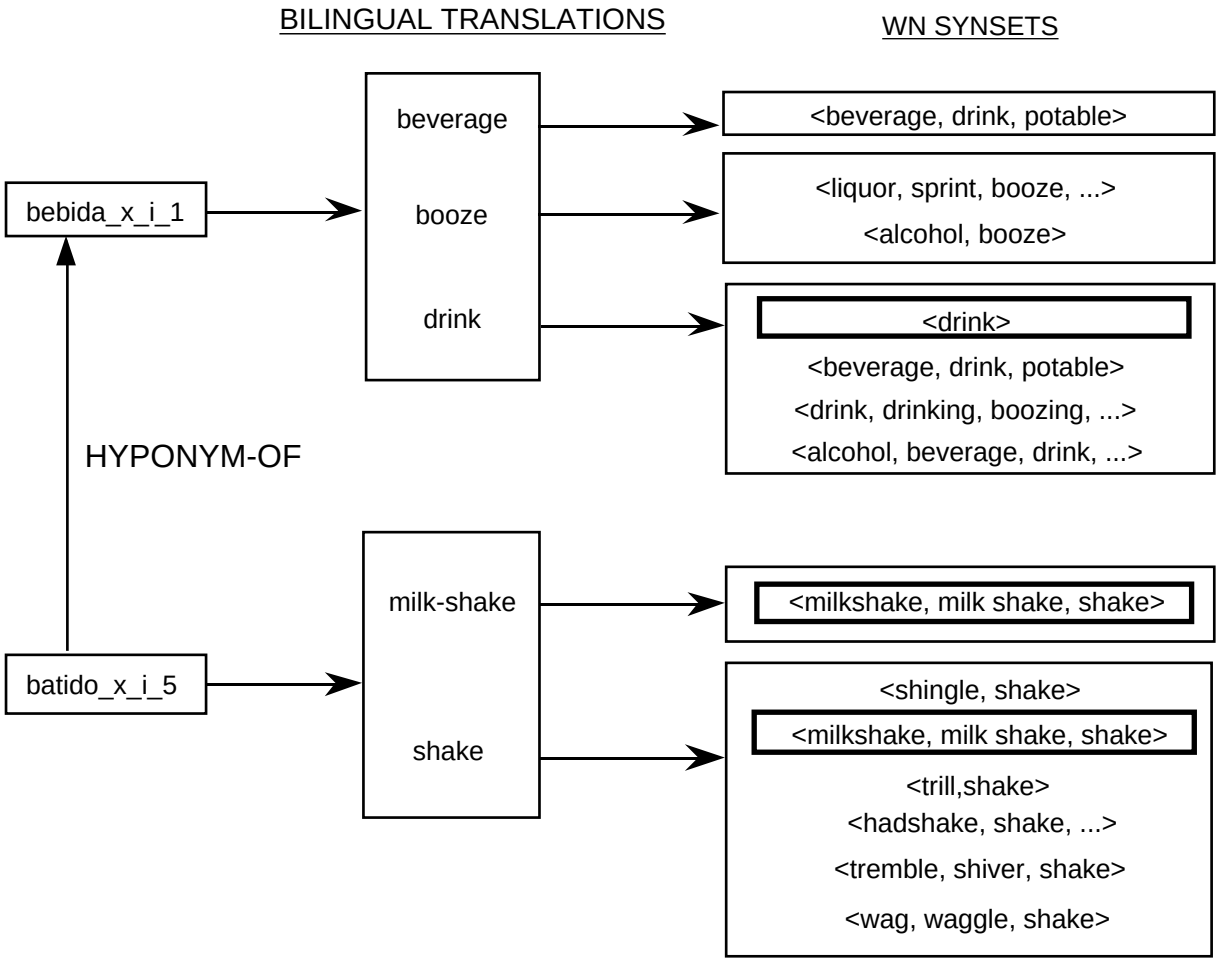

Figure 2, translation equivalence selection.

Once the translation links for *bebida_x_i_1* have been selected, all the possible translations of *batido* are looked up from the bilingual dictionary (if no translations are found in the bilingual dictionary other rulesets are launched in order to overcome this lexical gap, such as *parent-tlink-ruleset*, etc.). Applying the disambiguation module using the Conceptual Distance among those synsets proposed for *batido_x_i_5* and those previously attached for *bebida_x_i_1*, the most closer ones are selected (in bold squares). These selected synsets act as constraints for further disambiguation processes with the hyponyms of *batido* .

Several experiments have been undertaken on the same domains of precedent ones. In the food domain from 140 source lexical entries, up to 54 lexical entries (only 39%) has direct (by means of bilingual dictionaries) and correct (a correct sense for the translation is placed in WordNet) equivalent synsets in WordNet. This result is good taken into account the different sources of error: 1) inexistence of translation in the bilingual dictionary (50 cases), 2) there is a translation but not the correct one (30 cases), 3) there is no correct sense into WordNet (6 cases), 4) the translation does not appears in WordNet (no errors detected in this taxonomy).

Although the lexical gap among the three lexical knowledge sources used in his experiment all the lexical entries that belongs to the taxonomy of *comida* have been linked to WordNet synsets using the rulesets presented in [Ageno et al. 94] in a fully automatic way. The results have been the following:

simple-tlinks	**57**
simple-tlink-ruleset	52
compound-tlink-ruleset	2
orthographic-tlink-ruleset	3
phrasal-tlinks	**1**
phrasal-noun-tlink-ruleset	1
partial-tlinks	**84**
parent-tlink-ruleset	78
grandparent-tlink-ruleset	6

5 Conclusions

A method for automatically selecting the most likely tlink among a set of candidates has been presented. The proposal tries to overcome the main problem found on semiautomatically extracting translation links between multilingual lexical entries using as main KSs bilingual dictionaries.

The system mechanism is based on calculating the conceptual distance between the competing lexical entries in the target language and a central concept that corresponds to a previously linked lexical entry that appears higher in the taxonomy and further on selecting the higher ranked option.

The measure of Conceptual Distance we have used tries to discover the minimal distance between the corresponding synsets in WordNet. An extended experiment following this method has been carried out and reported here. We are planing to improve the Conceptual Distance formula in order to support different densities in the heterogeneous topology of WordNet and to extent the performance of the formula to a set of concepts.

References

[Ageno et al. 92] Ageno, A., I. Castellon, G. Rigau, H. Rodriguez, M.F.Verdejo, M.A.Marti and M.Taule *SEISD: An Environment for Extraction of Semantic Information from On-Line Dictionaries*, 3rd Conference on Applied Natural Language Processing (ANLP-92) Trento, Italy 1992

[Ageno et al. 94] Ageno, A., F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou, *TGE: Tlinks Generation Environment* 15th International Congress on Computational Linguistics (COLING 94) Kyoto, Japan 1994.

[Agirre et al. 94] Agirre E., Arregi X., Artola X., Díaz de Ilarraza A. and Sarasola K., *Conceptual Distance and Automatic Spelling Correction*, Workshop on Computational Linguistics for Speech and Handwriting Recognition, Leeds, 1994.

[Biblograf 87] Diccionario General Ilustrado de la Lengua Española VOX. Ed. Biblograf S.A. Barcelona, 1987.

[Biblograf 92] VOX Harrap´s Diccionario esencial Inglés-Español, Español-Inglés. Segunda Edición. Biblograf S.A. Barcelona, 1992.

[Church et al. 91]Church K., Gale W., Hanks P. and Hindle D., *Using Stadistics in Lexical Analisys*, in Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Zernik U. Ed. Lawrence Erlbaum Associates, publishers. Hillsdale, New Jersey. 1991.

[Copestake 92] Copestake, A. *The ACQUILEX LKB: representation issues in semi-automatic acquisition of large lexicons*,3rd Conference on Applied Natural Language Processing (ANLP-92) Trento, Italy 1992

[Copestake et al. 94] Copestake, A., Briscoe T., Vossen P., Ageno A., Ribas F., Rigau, G., Rodriguez H., Samiotou A. *Acquisition of Lexical Translation Relations from MRDs*, in Machine Translation 1995 (forthcoming). Also Esprit BRA-7315 Acquilex-II Working Paper n.040. 1994.

[Copestake et al. 92] Copestake, A., B. Jones, A. Sanfilippo, H. Rodriguez, P. Vossen, S. Montemagni and E. Marinai, *Multilingual lexical representatio*," in "The (other) Cambridge ACQUILEX papers", A. Sanfilippo (ed.) pages 117--129, University of Cambridge Computer Laboratory. Technical Report No. 253. 1992.

[Dolan et al. 93] Dolan W., Vanderwende L. and Richard son S., *Automatically deriving structured knowledge bases from on-line dictionaries.* in proceedings of the first Conference of the Pacific Association for Computational Linguistics (Pacling'93), April 21-24, Simon Fraser University, Vancouver, Canada. 1993.

[Knight & Luk 94] Knight K. and Luk S., *Building a Large-Scale Knowledge Base for Machine Translation*, in proceedings of the American Association for Artificial Inteligence. 1994.

[Miller 90] Miller G., *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

[Miller & Teibel 91] Miller G. and Teibel D., *A proposal for Lexical Disambiguation,* in Proceedings of DARPA Workshop on Speech and Natural Language, 395-399, Pacific Grave, California, February, 1991

[Rada et al. 89] Rada R., Mili H., Bicknell E. and Blettner M., *Development an Applicationof a Metric on Semantic Nets,* in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30. 1989.

[Ribas 94] Ribas F., *An Experiment on Learning Appropriate Selectional Restrictions from Parsed Corpus.* In Proceedings of the 16th International Conference on Computational Linguistics (Coling'94). Kyoto, Japan. 1994.

[Rigau 94] Rigau G., *An Experiment on Automatic Semantic Tagging of Dictionary Senses,* in Proceedings of the International Workshop The Future of the Dictionary, Uriage-les-Bains, Grenoble, France, 1994, also published as .Research Report LSI-95-??. Computer Science Department. UPC. Barcelona. 1995.

[Schütze 92] Schütze H., *Context Space*, in Workshop Notes of Fall Session of Statistically-Based Natural Language Processing Techniques, AAAI'92.

[Sussna 93] Sussna M., *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network,* in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia USA. 1993.

[Tanaka & Umemura 94] Tanaka K. and Umemura K., *Construction of a Bilingual Dictionary Intermediated by a Third Language*, in proceedings of the 16th International Conference on Computational Linguistics (Coling'94). Kyoto, Japan. 1994.

[Wilks et al. 93] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B., *Providing Machine Tractablle Dictionary Tools,* in Semantics and the Lexicon (Pustejowsky J. ed.), 341-401, 1993.