[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

Spanish EuroWordNet and LCS-Based Interlingual MT

Bonnie J. Dorr
Department of Computer Science and UMIACS
University of Maryland
College Park, MD, USA 20742
bonnie@umiacs.umd.edu

M. Antonia Martí and Irene Castellón
Department of Linguistics
University of Barcelona
Barcelona, Spain 08071
amarti/castel@lingua.fil.ub.es

Abstract

We present a machine translation framework in which the interlingua—Lexical Conceptual Structure (LCS)—is coupled with a definitional component that includes bilingual (EuroWordNet) links between words in the source and target languages. While the links between individual words are language-specific, the LCS is designed to be a language-independent, compositional representation. We take the view that the two types of information—shallower, transfer-like knowledge as well as deeper, compositional knowledge—can be reconciled in interlingual machine translation, the former for overcoming the intractability of LCS-based lexical selection, and the latter for relating the underlying semantics of two words cross-linguistically. We describe the acquisition process for these two information types and present results of hand-verification of the acquired lexicon. Finally, we demonstrate the utility of the two information types in interlingual MT.

1 Introduction

We present a machine translation framework in which the interlingua—Lexical Conceptual Structure (LCS)—is coupled with a definitional component that includes bilingual (EuroWordNet) links between words in the source and target languages. While the links between individual words are language-specific, the LCS is designed to be a language-independent, compositional representation. The advantage to using the two types of knowledge is that it reduces the computational inefficiency of the lexical-selection process—paring down the number of initial target-language candidates—while providing a basis for making a final selection among the remaining possibilities (e.g., "marchar através" vs. "atravesar marchando" for the English phrase "march across"). We take the view that the two types of information—shallower, transfer-like knowledge as well as deeper, compositional knowledge—can be reconciled in interlingual machine translation, the former for overcoming the intractability of LCS-based

lexical selection, and the latter for relating the underlying semantics of two words cross-linguistically.

This paper addresses the development of an interlingual framework with respect to these two information types. We describe the acquisition process for an initial Spanish database of verbs developed at the University of Maryland using a bilingual lexicon, a set of semantically classified English verbs from (Levin, 1993), and a set of links between these verbs in the synsets in WordNet (Miller, 1986; Miller, 1990; Miller and Fellbaum, 1991). We compare this initial database with the final Spanish database after hand-verification/modification by researchers at the University of Barcelona. The final database will be incorporated into the Spanish portion of EuroWordNet (Calzolari et al., To appear). A translation example is given in which we demonstrate the utility of the two information types in interlingual MT.

2 Development of Spanish EuroWordNet

EuroWordNet is a project aimed at developing a multilingual database with basic semantic relations between words for several European languages (Dutch, English, French, German, Italian, and Spanish) (Calzolari et al., To appear). Each individual WordNet will be linked to definitions in the English WordNet for English (Miller, 1986; Miller, 1990; Miller and Fellbaum, 1991).

One component of this project involves the development of links between verb definitions in English WordNet and Spanish WordNet. Researchers at the University of Barcelona have built these links automatically (Castellón et al., 1997) by importing the English verb database of (Dorr, 1997) into Spanish, using an intermediate Spanish-English bilingual lexicon produced at the University of Maryland. The imported database was hand-checked by a native Spanish speaker; the results of verification are reported herein.

One of our objectives is to incorporate Spanish WordNet into an existing interface called "Periscope", developed by NOVELL, that allows a user to browse through definitions bilingually. A snapshot between the Spanish verb *matar* and its Dutch equivalents, *uitmoorden* and *kapotmaken,* is given in Appendix A. The links in this snapshot (marked by a very faint gray line) indicate the correspondence between the WordNet sense in each of the two languages. For example, WordNet Sense 2 of the Spanish verb *matar* (in the synonym set containing *sacrificar)* is linked to WordNet Sense 1 of the Dutch verb *uitmoorden* (in the synonym set containing *moorden, afmaken,* and *afslachten)* which corresponds to the meaning "kill a large number of people indiscriminately." Similarly, WordNet Sense 4 of the Spanish verb *matar* is linked to WordNet Sense 2 of the Dutch verb *kapotmaken* (in the synonym set containing *doodmaken* and *doden)* which corresponds to the meaning "cause to die."

Our approach to building these links is a simple transitive closure involving online resources as shown in Figure 1. The first step was to hand-tag each Levin-

classified verb with a set of WordNet senses, a process which took 1 person-month of effort at the University of Maryland using an interface for typing in human grammatically judgments of sentences in each class. The WordNet senses are presented to the user as a set of logical addresses (e.g., Sense 1, Sense 2, ...) which are then are converted internally into WordNet addresses, e.g., (00416048, 00416049, ...). The second step was to construct a bilingual lexicon for Spanish and English and to import Spanish entries into Levin-based categories (9.1-57) as an additional disambiguating feature. This process took 4 person-months of effort. The resulting database was later hand-verified over a period of 2 person-months by researchers at University of Barcelona. Finally, an automatic procedure was used to map the Spanish words into their corresponding WordNet senses, by merging the Spanish verbs in step 2 with their English WordNet sense counterparts in step 1. The entire process took 7 months, with 6 months of effort devoted entirely to Spanish. Adding a new language would presumably take on the order of 6 months since the result of Step 1 may be reused for other languages.

3 Construction of LCS's

The database resulting from Step 1 of Figure 1 is strictly based on English entries. The fully expanded form of these entries includes additional information, e.g., thematic grids—(ag)ent, (th)eme, etc.—so that the entries can be used as input to a Lexical Conceptual Structure (LCS) construction program called LEXICALL (Dorr, 1997). For example, the fully expanded form of the first entry above in Step 2 is:

```
9.1#_ag_th_goal()#arrange#Sense 2  (00416049)#
```

This entry is processed by LEXICALL to produce a LCS of the following form for the verb *arrange:*

```
(cause  (* thing 1)
  (go loc  (* thing 2)
    ((* toward 5)  loc  (thing 2)
      ([at]   loc  (thing 2)   (thing 6))))
   (arrangingly 26))
```

For presentational purposes, we provide English examples only for the remainder of this section. However, as described in (Dorr, 1997), the representations used here carries over to other languages as well. In fact, we have used the same acquisition program, without modification, for building our Spanish and Arabic lexicons, each of size comparable to our English lexicon.

We used Levin's publicly available online index (Levin, 1993) as a starting point for LEXICALL. While this index provides a unique and extensive catalog of verb classes, it does not define the underlying meaning components of each

Step 1: English Levin Verbs<-> WordNet Sense Tags: University of Maryland

Examples:	**arrange**	**class 9.1**	**Sense 2 (00416049)**
	place/position/put	**class 9.1**	**Sense 1 (00859635)**
	remove	**class 10.1**	**Sense 1 (00104355)**
	evacuate	**class 10.2**	**Sense 3 (01150129)**
	float	**class 11.2**	**Sense 1 (01069124)**
	distribute	**class 13.2**	**Sense 1 (01313552)**
	pour	**class 26.3**	**Sense 2 (01184040)**

Step 2: Spanish Verbs <->English Levin Verbs: University of Maryland, hand-verified by University of Barcelona

Examples:	**clasificar**	**arrange**	**class 9.1**
	colocar	**place/position/put**	**class 9.1**
	borrar	**remove**	**class 10.1**
	desocupar	**evacuate**	**class 10.2**
	flotar	**float**	**class 11.2**
	repartir	**distribute**	**class 13.2**
	echar	**pour**	**class 26.3**

Step 3: Spanish Levin Verbs <-> WordNet Sense Tags: University of Barcelona

Examples:	**clasificar**	**class 9.1**	**Sense 2 (00416049)**
	colocar	**class 9.1**	**Sense 1 (00859635)**
	borrar	**class 10.1**	**Sense 1 (00104355)**
	desocupar	**class 10.2**	**Sense 3 (01150129)**
	flotar	**class 11.2**	**Sense 1 (01069124)**
	repartir	**class 13.2**	**Sense 1 (01313552)**
	echar	**class 26.3**	**Sense 2 (01184040)**

Figure 1: Combining Resources at University of Maryland and University of Barcelona for Construction of Spanish WordNet

class. One of the main contributions of our work is that it provides a relation between Levin's classes and meaning components as defined in the LCS representation. We have hand-constructed a database containing 191 LCS templates, i.e., one for each verb class in (Levin, 1993). In addition, we have generated LCS templates for 26 additional classes that are not included in Levin's system. Several of these correspond to verbs that take sentential complements (e.g., coerce).

Using the template database as a "seed" to the LEXICALL program, we have built a large repository of LCS representations for English verbs. An entry in the seed database includes a semantic class number with a list of verbs associated with that class in Levin, a thematic grid, and a LCS template:

```
Class 9.1: arrange, immerse, ..., put, place, position  ...
Thematic Grid:   _ag_th_loc
LCS Template:
  (cause  (thing 1)
     (go loc  (thing 2)
        (toward loc  (thing 2)
          (Cat] loc  (thing 2) (thing 6))))
      (!!-ingly 26))
```

The semantic class label 9.1 above is taken from Levin's 1993 book (Put Verbs), i.e., the class to which the verb *arrange* has been assigned. A verb, together with its semantic class uniquely identifies the word sense, or LCS template, to which the verb refers. The thematic grid **(_ag_th_loc)** indicates that the verb has three obligatory arguments, an *agent,* a *theme* and a *location.*[1] The **!!** in the LCS Template acts as a wildcard; it will be filled by a lexeme (i.e., a root form of the verb). The resulting form is called a *constant,* i.e., the idiosyncratic part of the meaning that distinguishes among members of a verb class (in the spirit of (Grimshaw, 1993; Levin and Rappaport Hovav, To appear; Pinker, 1989; Talmy, 1985)).[2]

The inputs above (class number, verb name, and thematic grid) are all that is required for acquisition of LCS's for a verb.[3] The output of LEXICALL is a Lisp-like expression corresponding to the LCS representation, e.g., the LCS for *arrange* given above.

---

[1] An underscore (_) designates *an* obligatory role and a comma (,) designates an optional role.

[2] A reviewer points out that the 'constant' seems to act as a variable since it has a filler, namely the lexeme associated with the particular lexical item. However, the process of filling in the constant position is executed at lexicon precompilation time, not during online processing. Once this position is filled, it is never again touched by the application program; compare with a 'true' variable such as (**thing 1**), which must necessarily be filled in by the application and which can take on any number of possible values during online processing.

[3] The LCS template is implicitly a fourth input to the program, i.e., it is automatically associated with the class and thematic grid, as stored in our hand-constructed database.

4 Results of Verification of Spanish/English Database

In developing the verb database for new languages, the amount of effort required for step 2 of Figure 1 is subject to a higher degree of variability than that of the other two steps since Levin's classes were initially based on English verbs. As such, we investigated the nature of the types of modifications that were made during this porting process so that we might have a better idea of the types of mismatches that are likely to arise when we examine additional languages. There were 18353 entries (3623 verbs) in the initial Spanish-English lexicon. Of these, 3025 entries were verified to be correct and the remaining 15328 entries were modified as specified below.

4.1 Modification Type 1: Polysemy—Class Assignment Refinements

The first type of modification to the database was the elimination of incorrect assignments of Spanish verbs to semantic classes due to an association with a high number of polysemous English counterparts. There were 6082 deletions of this type out of the 15328 revised entries.

For example, the Spanish verb *escribir* had several English translations: *pen* (as in *John penned a letter to Mary), write* (as in *John wrote Mary a letter),* and *compose* (as in *John composed a letter to Mary).* These English counterparts were mapped automatically into the following semantic classes in our initial database:

- pen—9.10 (Pocket Verbs)
- compose—26.4 (Create Verbs)
- compose—26.7 (Performance Verbs)
- write—25.2 (Scribble Verbs)
- write—37.1 (Transfer of Message Verbs)

Of these, only classes 25.2, 26.7, and 37.1 survived hand-revision since 9.10 refers to *pen* in the sense of "putting into a pen" (not "writing with a pen") and 26.4 refers to *compose* in the sense of "constructing something" (not "writing something").

In addition to the elimination of incorrect class assignments, 334 entries were reclassified into alternative Levin classes. For example, the Spanish verb *acusar* was originally assigned to classes 33 (Judgment Verbs) and 10.6 (Cheat Verbs), but this verb was reassigned to class 13.4.2 (Equip Verbs).

The amount of time required for the hand-verification process would be greatly reduced if the issue of polysemy had been addressed earlier in the process. For example, hand-annotating each Spanish-English entry with a semantic class

during initial construction of the bilingual dictionary would be more efficient than blindly porting Spanish verbs into semantic classes via English translations and relying on hand-verification later.

4.2 Modification Type 2: Thematic Grid Refinements

The second type of modification to the database was thematic grid refinement, i.e., elimination of thematic grids that were not applicable to Spanish verbs or modification of prepositions or other information associated with thematic roles. Of the remaining 8912 entries in the revised Spanish-English lexicon, there were 6295 modifications of this type. There were 3648 deletions of non-applicable thematic grid entries.

For example, the Spanish verb *escribir* was correctly assigned to class 25.2 (Scribble verbs) in the initial database, but one of the two thematic grids for this assignment was removed: **_ag_th,goal** (as in *He wrote his name on the photo).* The remaining thematic grid, **_ag_th** (as in *He wrote his name)* was left in the database since it provides the most basic thematic requirements for the verb *escribir.*

The remaining 2747 thematic grid modifications involved changes to prepositions or other information associated with thematic roles. For example, the Spanish verb *leer* (read) was given the thematic grid **_exp,perc,mod-loc(de)** when it was ported into our initial database. Hand verification revealed that the preposition *de* (of) was incorrect; this was replaced with a more appropriate preposition, *sobre* (about), as in *Antonio leyó sobre el asesinato* (Antonio read about the assassination).

Another example of a thematic grid modification is one where an optional role is made obligatory, e.g., the verb *declarar* (declare) had an optional beneficiary in the initial database: **_ag_th,ben(a).** This was modified to have an obligatory beneficiary **(_ag_th_ben(a)**), as in *Mariá declaró sus intenciones a Antonio* (Maria declared her intentions to Antonio).

4.3 Other Modifications

An additional 2617 entries (955 verbs) were deleted due to the rarity of usage and/or disjointness with respect to WordNet concepts, e.g., *zapar* (sap). These deletions were (somewhat) balanced off by the addition of 1213 new entries, i.e., 1092 verbs not in the initial database—primarily reflexive forms for existing non-reflexive counterparts (e.g., *alarmarse).*[4] The total number of entries in the final Spanish-English database is 7319 (3821 verbs).

[4] These deletions and additions are not yet stabilized; we will have final figures on this in September.

5 Utility of Both Knowledge Types in Interlingual Machine Translation

Our ultimate objective is to use the two knowledge types, i.e., WordNet-based information for linking two verbs cross-linguistically and LCS-based information for relating the underlying semantics of these two verbs. The idea would be to select the appropriate target-language words for an LCS produced by the source-language sentence, through access to WordNet links, and then to make a final selection based on coverage of the meaning components in the LCS.

Consider the following example of translation between English and Spanish:

E: The soldier marched across the field.

S: El soldado marchó através el terreno (The soldier marched across the field)
El soldado atravesó el terreno (The soldier crossed the field)
?El soldado atravesó el terreno marchando (The soldier crossed the field marching)

The first of the three target-language sentences is considered to be the most acceptable by native speakers since it contains all relevant information without redundancy. The second sentence is also acceptable, but misses information about marching.[5] The third sentence contains all the relevant information, but is the most awkward.[6] In the absence of additional semantic information, e.g., about selectional restrictions, our algorithm opts for the least awkward yet most specific version of the target-language sentence, *El soldado marchó através el terreno.*

A diagram of the lexical selection process for this translation example is given in Figure 2. In this figure, we see that the LCS produced for the source-language (English) sentence contains three major pieces.[7] The first is the 'GO-BY_MARCHING' portion of the LCS marked *1. march.* The second is the 'TO ACROSS' portion of the LCS marked *2. across.* The third is the 'GO TO ACROSS' portion of the LCS marked *3. cross.*

[5] It is important to note here that the missing information might be inferred from the subject, *soldado,* which prototypically occurs as a theme of the marching action. Recent work on selectional restrictions, e.g., (Resnik, 1996; Castellon and Marti, 1997) could be used for additional guidance during this process, perhaps allowing this second sentence to be selected—for reasons of economy—when the manner of motion can be inferred from the prototypical subject or object.

[6] Again, the subject of the sentence seems to play a role in the acceptability of this sentence. If *él (*= he) were used in place of *el soldado,* this sentence would be perfectly acceptable. See related footnote *5*.

[7] The LCS given here is based on templates developed in 1996. More recently, we have refined the LCS templates to include activities (ACT), so that the entry for *march* has changed from GO to ACT. (See (Dorr and Olsen, 1997) for more details.) However, the basic mechanism for lexical selection via WordNet links is still applicable to the modified representation, which assumes ACT to be a degenerate case of GO.

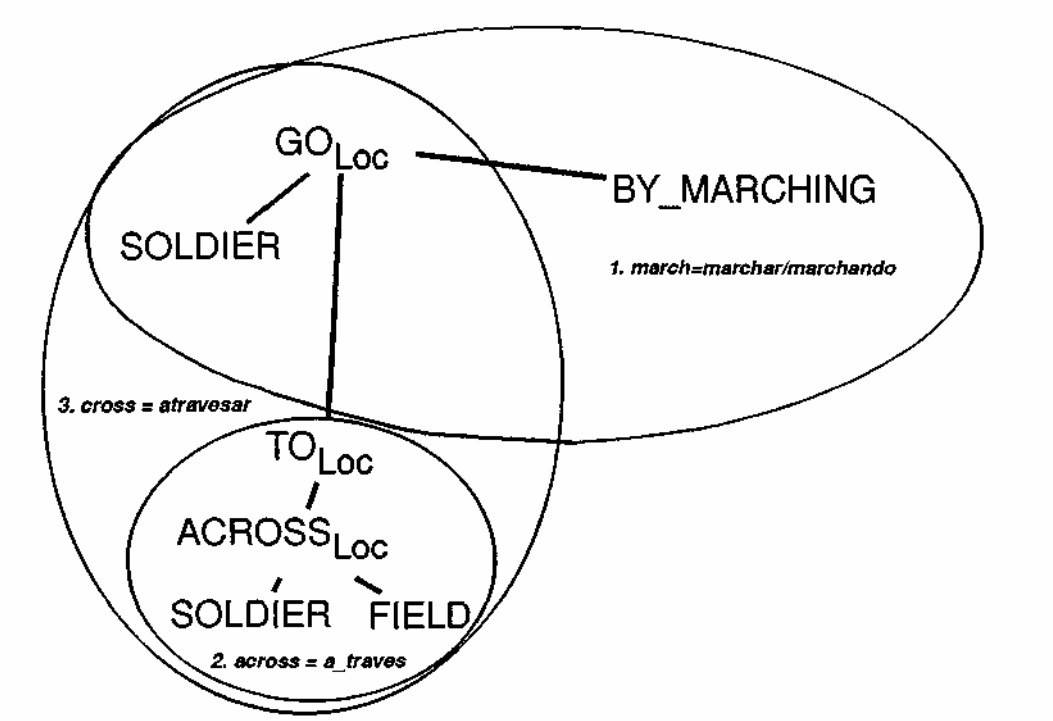


Figure 2: Mapping LCS into Target Language by means of Cross-language WordNet Entries

These three components are associated with three English words in our LGS lexicon; the English words, in turn, are linked directly to their Spanish WordNet counterparts *(march(ar/ando), a través,* and *atravesar).*[8] This direct-link method of selecting target-language candidates is more efficient than an earlier LCS-based lexical-selection algorithm (Dorr, 1993) which relied on generalized graph-matching—an impractical technique for a large-scale application.

The final selection of target-language words is based on a procedure that determines which of the three possible LCS combinations retains relevant information (i.e., full coverage of the LCS (Dorr, 1993)) while avoiding redundancy (i.e.. multiple coverage of the same LCS component) wherever possible. Ovals 1 and 3 in Figure 2 indicate an overlap in meaning between the verbs *cross* and *march*. In particular, the GO component of meaning appears in both. This, perhaps, provides a computational basis for the awkwardness of the third sentence above. However, if one of these is left out (e.g., *marcha(r/ndo)),* as in the second sentence above, then there will be a piece of the LCS that is left uncovered. Thus, the only remaining possibility is the first sentence, *El soldado marchó através el terreno.*[9]

6 Conclusions

We have provided an argument for the use of two different information types in interlingual MT, transfer-like links (cross-language WordNet links) for efficient selection of target-language candidates, and conceptual knowledge (LCS) for efficient selection among these candidates. We have described the acquisition process for Spanish verb entries in EuroWordNet and we have presented results of hand-verification of our online bilingual database. Finally, we have demonstrated the utility of combining the two different information types in interlingual MT by showing how each contributes to the lexical-selection process during generation.

One of the main innovations of this work is that it provides a technique for lexical-selection that is more efficient than an earlier LCS-based algorithm (Dorr,

[8] Although we have only discussed verbs in this paper, other parts of speech are linked via WordNet such as prepositions, adjectives, and nouns. These too would enter into the lexical selection process.

[9] Our French informants have indicated that the analogous sentence, *Le soldat a marché à travers le champ,* is not an acceptable translation for the English sentence. Rather, the most acceptable translation would be *Le soldat a traversé le champs en marchant* (analogous to the third case above). A more detailed analysis reveals that this divergence is due, in part, to the fact that the preposition *à travers* is not equivalent to the preposition 'across' used in the Spanish sentence. Rather, *à travers* includes a meaning component that corresponds to the LCS primitive VIA (through). Our algorithm would not select *à travers* if the LCS contained ACROSS since there would be no link between this preposition and ACROSS in the French lexicon (unlike the Spanish case). Note, by contrast, that *à travers* would be selected for a sentence like *He went (passed) through the wall,* i.e., *Il est passé à travers du mur.* In this case, the underlying LCS would contain VIA, which would be linked to *à travers* in the French lexicon.

1993) which relied on generalized graph-matching—an impractical technique for a large-scale application. We believe this framework coupled, perhaps, with additional semantic information about argument restrictions, to be an initial step toward providing a computational basis for the acceptability/awkwardness in human judgments for target-language sentences.

7 Acknowledgements

This research was conducted in the Computational Linguistics and Information Processing (CLIP) Laboratory at the University of Maryland (UM), the Linguistics Department at the Universidad de Barcelona (UB), and the Computer Science Department at the Universidad Politécnica (UPC). The work was supported, in part, by National Science Foundation Presidential Faculty Fellowship (PFF/PECASE) Award IRI-9629108, Department of Defense contract MDA90496C1250, DARPA/ITO Contract N66001-97-C-8540, Army Research Laboratory contract LETTER11097 through United Research Corporation, and Army Research Laboratory contract DAAL03-91-C-0034
through Battelle.

References

Calzolari, Nicoletta, Antonia Marti, Horacio Rodriguez, Felisa Verdejo, Piek Vossen, and Yorick Wilks. To appear. EuroWordNet Project (Title under Revision). *Computers and the Humanities.*

Castellón, Irene and M. Antonia Martí. 1997. Extracción de Restricciones Selectivas a partir de Corpus. In *Presentation at Fifth Workshop on Computational Lexicography,* Joanet, Spain.

Castellón, Irene, M. Antonia Martí, Roser Morante, and Gloria Vazquez. 1997. Propuesta de Alternancias de Diátesis Verbales para el Español y el Catalán (A Proposal for Verbal Diathesis Alternations for Spanish and Catalan). In *Proceedings of the Spanish Society for Natural Language Processing (SEPLN),* pages 31-48, Madrid, Spain. Volume 21.

Dorr, Bonnie J. 1993. *Machine Translation: A View from the Lexicon.* The MIT Press, Cambridge, MA.

Dorr, Bonnie J. 1997. Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP),* Washington, DC.

Dorr, Bonnie J. and Mari Broman Olsen. 1997. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th*

Annual Meeting of the Association for Computational Linguistics (ACL-97), Madrid, Spain, July 7-12.

Grimshaw, Jane. 1993. Semantic Structure and Semantic Content in Lexical Representation. unpublished ms., Rutgers University, New Brunswick, NJ.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation,* University of Chicago Press, Chicago, IL.

Levin, Beth and Malka Rappaport Hovav. To appear. Building Verb Meanings. In M. Butt and W. Gauder, editors, *The Projection of Arguments: Lexical and Syntactic Constraints.* CSLI.

Miller, George A. 1986. Dictionaries in the Mind. *Language and Cognitive Processes,* 1:171-185.

Miller, George A. 1990. WordNet: An On-Line Lexical Database. *International Journal of Lexicography,* 3:235-312.

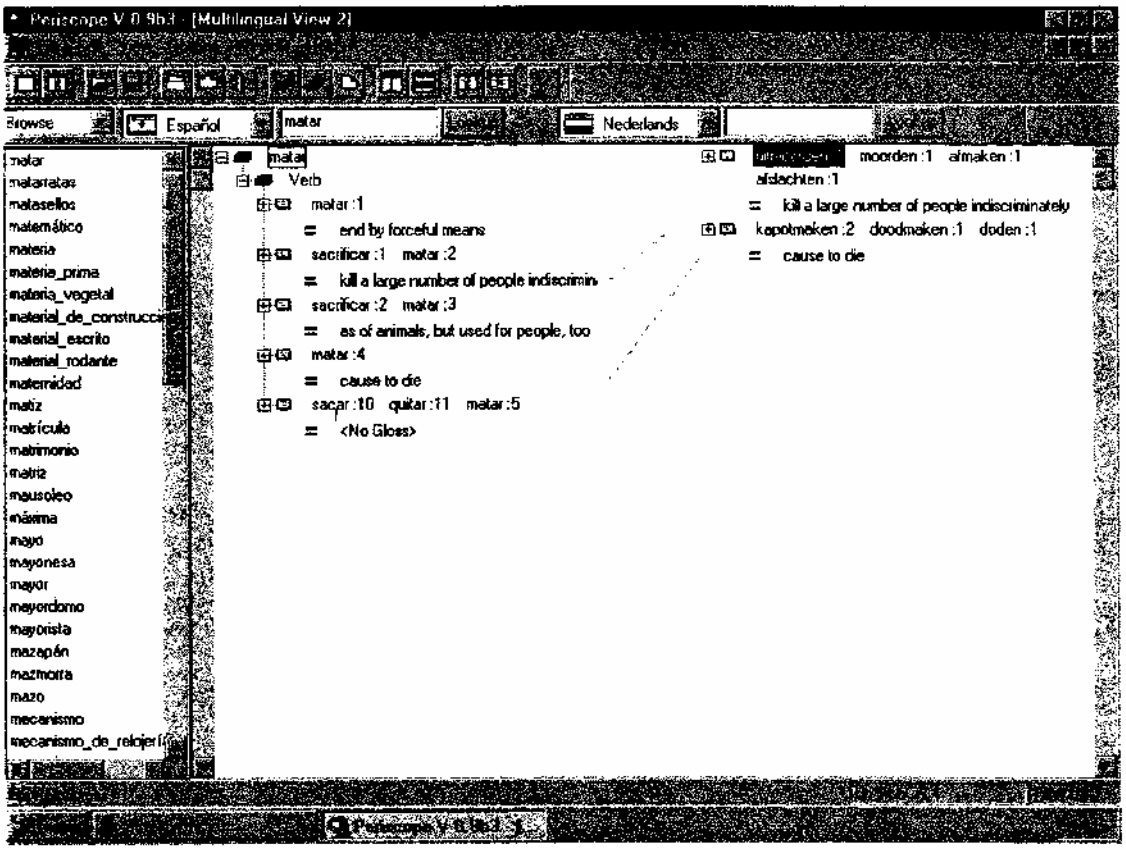
Miller, George A. and Christiane Fellbaum. 1991. Semantic Networks of English. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics, Cognition Special Issue.* Elsevier Science Publishers, B.V., Amsterdam, The Netherlands, pages 197-229.

Pinker, Steven. 1989. *Learnability and Cognition: The Acquisition of Argument Structure.* The MIT Press, Cambridge, MA.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition,* 61:127-159.

Talmy, Leonard. 1985. Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen, editor, *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon.* University Press, Cambridge, England, pages 57-149.

A Periscope: Tool for illustrating Links Between WordNet Synsets across Languages


Periscope V 0.9b3 - [Multilingual View 2]
Browse
Español
matar
Nederlands
matar
matarratas
matasellos
matemático
materia
materia_prima
materia_vegetal
material_de_construcci
material_escrito
material_rodante
maternidad
matiz
matrícula
matrimonio
matriz
mausoleo
máxima
mayo
mayonesa
mayor
mayordomo
mayorista
mazapán
mazmorra
mazo
mecanismo
mecanismo_de_relojer
mata
Verb
matar :1
= end by forceful means
sacrificar :1 matar :2
= kill a large number of people indiscrimin
sacrificar :2 matar :3
= as of animals, but used for people, too
matar :4
= cause to die
sacar :10 quitar :11 matar :5
= <No Gloss>
moorden :1 afmaken :1
afslachten :1
= kill a large number of people indiscriminately
kapotmaken :2 doodmaken :1 doden :1
= cause to die