Evelyne Viegas & Sergei Nirenburg, Computing Research Laboratory, New Mexico State University

The Ecology of Lexical Acquisition: Computational Lexicon Making Process

Abstract

In this paper, we present the process of dictionary making, as we defined it to **build Spanlex,** a Spanish lexicon, for **Mikrokosmos,** a machine translation (MT) system based on semantics. Our aim in this one-year project has been to acquire a lexicon of about 40,000 word meanings. We adopted a computational semantics approach where the semantics has to be entered semi-automatically by the acquirers. This approach pointed toward automating the task of acquisition as much as possible. Therefore, we had to develop and implement tools to help and guide acquisition. We present here in a diagram, representing the process of acquisition, some of the resources needed. Finally, we focus on the way to acquire a large-scale high quality lexicon by using derivational morpho-semantic rules.

1. Introduction

The dictionary-making process we will illustrate has been tested and attested for **Mikrokosmos**, a Machine Translation (MT) system between Spanish and Japanese. Mikrokosmos is a knowledge based system based on semantics (Nirenburg et al., 1994) and adopts an interlingual approach to MT. The interlingual representation is called a Text Meaning Representation (TMR). We cannot develop the entire process of translation, it is enough for present purposes to say that the final TMR is essentially built with the unsaturated or underspecified TMRs coded in the lexicon. For instance, *eat* encodes in its TMR the following selectional restrictions: ANIMAL for agent and EDIBLE for its theme; it is only at the text level that we will have a saturated TMR.

In *John likes eating candies,* ANIMAL will be constrained to HUMAN. We adopt a computational linguistic perspective, where the notion of lexical organisation is central to the theory. We take advantage of techniques coming from (computational) linguistics and artificial intelligence.

Our aim in this one-year project, has been to acquire a lexicon of about 40,000 word meanings. This implied automating the task of acquisition as much as possible to facilitate the most unpleasant tasks, such as spell checking, so that the acquirers could concentrate on more productive and interesting tasks, such as acquiring the semantics, pragmatics, etc.

This implies access to on-line dictionaries, on-line corpora, and software allowing lexicographers to access all the on-line information in an easy way. Our interfaces have been elaborated with respect to users' needs, and continue to evolve on a as needed basis.

The acquirer (lexicographer, terminologist, etc.) is presented a series of predefined semantico-syntactic templates, which guide him/her in the phase of acquisition.

In the following sections, we present the type of information which is required inside computational lexicons, and the way this information should be structured. We then present some aspects of the acquisition task, with a diagram, thus exemplifying the dictionary-making process. We also show how it is possible to acquire a lexical semantic large-scale and high quality lexicon, by using morphosemantic rules.

2. Organisation and Use of Computational Lexicons

The type of information which should be included in the lexicon highly depends on the domain of application for which it has been built. For instance, for multilingual translation, transfer dictionaries could be enough, *manger/comer/eat,* in French, Spanish and English respectively. In generation, information on word order, *(a hot coffee* vs *a coffee hot),* and collocations inside a generation lexicon, *a heavy smoker* vs *un grand fumeur* in French, are necessary.

Acquiring a large-scale lexicon is very expensive work, this is why it is recommended to build lexicons that are reusable for other domains or applications.

We thus turn now toward the type of **organisation and structure** of the lexicon we want. It is well known in computational lexical semantics that a sense enumeration approach only based on subcategorisation differences is computationally expensive and unrealistic from a theoretical viewpoint, where we fail to capture the core meaning of words (Boguraev and Pustejovsky, 1990, Viegas and Nirenburg, 1995).

Our lexicons are composed of superentries (Meyer et al., 1990), where each entry consists of a list of words, stored independently of their part of speech (the verb and noun form of *walk* are under the same superentry). Each word meaning is identified by a unique identificator, or lexeme (Mel'cuk et al. 1984, Onyshkevych and Nirenburg 1994).

2.1 The Different Zones inside a Lexeme

The information contained inside a lexeme is divided into **zones** corresponding to various levels of lexical information (Meyer et al. 1990).

CATegory: *Noun, Verb, Pronoun,...;* **MORPH**ology: for irregular forms and stem changes *mouse* vs *mice;* **COMMENTS**: providing administrative information, definition, examples...; **ORTH**ography: for abbreviations, *United States of America* vs *USA*; **PHON**ology; **SYN**tactic-**STRUC**ture: giving essentially subcategorisations; **SEM**antic-**STRU**Cture: giving the semantics, with selectional restrictions, in terms of its unsaturated TMR; **LEX**ical-**REL**ations: encoding colloca-tional information; **LEX**ical-**RULES**: give the rules that apply to this lexeme; **STYL**istics: give information on stylistic factors, such as familiarity, etc..., and include sub-zones containing triggers for analysis and generation.

In the following, we focus on the **SYN**tactic-**STRU**Cture and **SEM**antic-**STRU**Cture zones. Each lexical entry contains a representation of its semantics, represented by using terms from the ontology as primitives (in addition to other non-ontological primitives, e.g., to reflect speaker attitudes and modality). These representations of lexical meaning may be defined using any number of ontological primitives, which we call *concepts.* Below is an example of the syntax and semantics for the Spanish entry *beber* (drink) (Figure 1), using the typed feature structures (tfs) as described in Pollard and Sag (1987).

$$\begin{bmatrix} \text{beber-V1} & \\ \text{syn:} & \begin{bmatrix} \text{root:} & \boxed{0} & \\ \text{subj:} & \boxed{1} & \begin{bmatrix}\text{cat:} & \textbf{NP}\\ \text{sem:} & \boxed{11}\end{bmatrix} \\ \text{obj:} & \boxed{2} & \begin{bmatrix}\text{cat:} & \textbf{NP}\\ \text{opt:} & +\\ \text{sem:} & \boxed{21}\end{bmatrix}\end{bmatrix} \\ \text{sem:} & \begin{bmatrix}\textbf{ingest} & & \\ \text{agent:} & \boxed{11} & \textbf{animal}\\ \text{theme:} & \boxed{21} & \textbf{liquid}\end{bmatrix}\end{bmatrix}$$

Figure 1: Partial Lexicon Entry for the Spanish lexical item *beber*

Notice that the concept *beber* maps into INGEST, which has selectional restrictions included in the ontology, such as ANIMAL and INGESTIBLE for its agent and theme respectively. These selectional restrictions work fine for *eat* but not *drink,* there we constrained the theme of INGEST to LIQUID as shown in the entry above for Spanish. When the meaning is represented using multiple concepts, they are tightly interconnected and constrained as appropriate. Any of the concepts in the ontology (currently numbering about 5,000 in Mikrokosmos) may be used (singly or in combination) in a lexical meaning representation.

The Ontology. The set of symbols and possible relationships between them are grounded in a language-independent knowledge source called the *ontology.* The symbols are defined as *concepts* in the ontology. As described, e.g., in Mahesh and Nirenburg (1995) and Mahesh (1996), the ontology is a large collection of information about EVENTs, OBJECTs and PROPERTYs in the world. In addition to the taxonomic multi-hierarchical organization, each concept has a number (currently averaging 14) of other local or inherited links to other concepts in the ontology, via relations (themselves defined in the PROPERTY sublattice). These links include case-role-like relations linking EVENTs to semantic constraints on the allowable fillers of those case-roles (i.e., selectional restrictions), properties (such as MANUFACTURER-OF) of things like COMPANYs, etc.

The Spanish word *beber (to drink)* was mapped into the concept LIQUID, whose selectional restrictions are ANIMAL and ingestible for its agent and theme, respectively, except that the selectional restriction specified in the theme of the lexicon entry of *beber* (Figure 1) constrained it to be of type LIQUID.

In a multilingual situation, however, it is not easy to determine this boundary. As a result, ontology and lexicon acquisition involves a process of daily negotiations between the two teams of acquirers, as is described and illustrated in Figure 2.

It is important to note that there need not be any correlation between syntactic category and semantic or ontological class. For example, although many verbs are EVENTs and a number of nouns are represented by concepts from the OBJECT subtree (such as the class of artifacts), frequently this is not the case. This is particularly the case with words derived via Lexical Rules (LRs). Many LRs change the syntactic category of the input form; in our model the semantic category is often preserved in many of these LRs. For example, the verb *destroy* may be represented by an EVENT, as will the noun *destruction* (with a different linking in the syntax-semantics interface, of course). Similarly, *destroyer* (as a person) would be represented using the same event with the addition of a HUMAN as a filler of the agent case-role. This built-in transcategoriality is a very natural aspect of the interlingual approach to MT, avoiding many of the category mismatches and misalignments that plague other paradigms in MT.

2.2 Towards a Multi-purposes Knowledge Base

From previous sections, we can generalise that we need lexicons that are multilingual, multi-media, multi-purposes:

a. **multilingual:** French, English, Japanese, Russian, Spanish, etc.

b. **multi-media:** containing linguistic and ontological information for natural language processing as well as phonological information, essentially for speech recognition and production

c. **multi-process:** applicable for analysis, generation (both mono- and multilingual), MT, summarization, information extraction, or speech processing.

3. The Acquisition Process: the Different Tasks

Below we present the general picture of the process of acquisition as we developed it for **Spanlex.** Figure 2 illustrates the acquisition process of static knowledge, such as the lexicon and the ontology, which are being used dynamically by the semantic analyser: lexicon acquirers have access to various on-line resources, such as **corpus search**, **look-up dictionary**, **ontology browser** tools.

So far, the tool suite has been used to support the input of lexicon entries for Spanish, Japanese, English, and Russian. This set of tools is being shared across geographical, disciplinary, and project group boundaries on a daily basis.

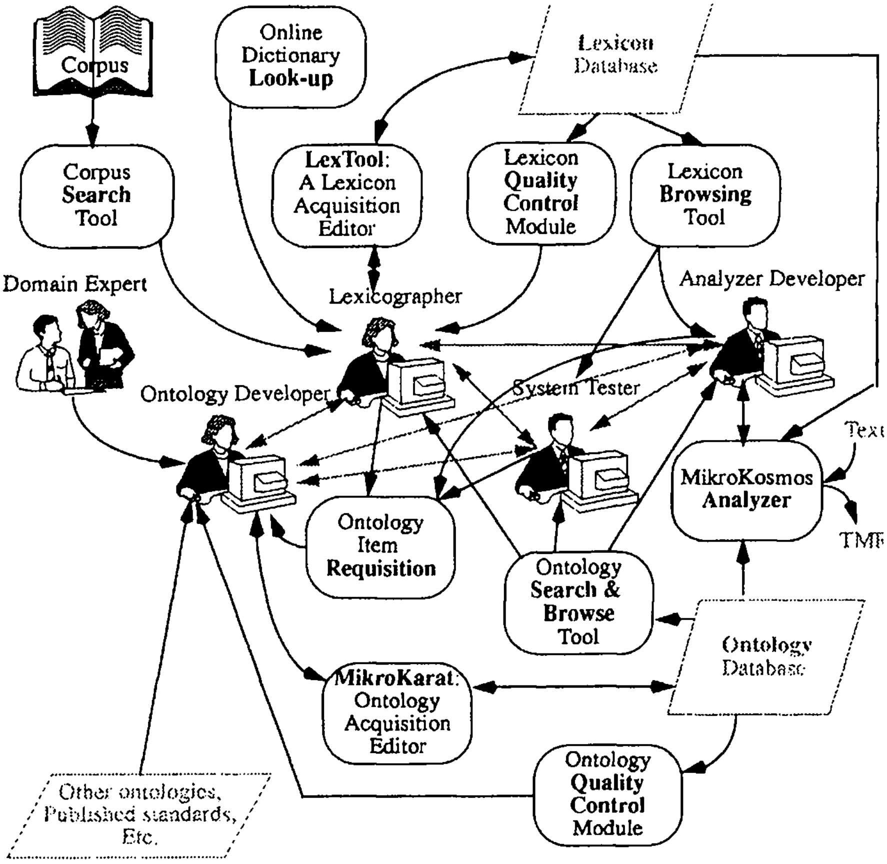
Corpus
Online Dictionary Look-up
Lexicon Database
Corpus Search Tool
LexTool: A Lexicon Acquisition Editor
Lexicon Quality Control Module
Lexicon Browsing Tool
Domain Expert
Lexicographer
Analyzer Developer
Ontology Developer
System Tester
Text
MikroKosmos Analyzer
TMR
Ontology Item Requisition
Ontology Search & Browse Tool
Ontology Database
MikroKarat: Ontology Acquisition Editor
Ontology Quality Control Module
Other ontologies, Published standards, Etc.

Figure 2. Acquisition Process of Static Knowledges

Developing a large-scale lexicon including this much of information (see section 2.1), cannot be done totally manually. Therefore, apart from the tools to help acquiring the data, we have also developed programs to check the semi-automatically acquired data. Using this approach, we have acquired about one-fifth of our lexicon, and have developed a morphosemantic acquisition program, which has allowed us to acquire the remaining four-fifths entirely automatically. We briefly show here a partial sample of the derivational morphology output for *comprar,* with the associated lexical-rules which are later used to actually generate the word entries[1] (see Viegas et al. 1996a, 1996b):

comprar, v, LR1event
comprador, n, LR2socialrole_relation1a
compra, n, LR2event1O
compra, n, LR2theme_of_event1O
comprable, adj, LR3feasibility_attribute1
comprado, adj, LR3event_telic
compradizo, adj, LR3feasibility_attribute5a
comprador, adj, LR3social_role_relation1a
malcomprar, v, LRneg_affect1, LR1event
...

For instance, *comprable, adj, LR3feasibility_attribute1,* is morphologically derived from *comprar,* and adds to the semantics of *comprar* the characteristics of being possible or not.

4. Conclusion

In this paper we illustrated the necessary resources to acquire semi-automatically largescale high quality lexicons. We also insisted upon a human interaction in the process of acquisition, and therefore the needs to build tools to help them acquire and check the word entries.

We also focused on the necessity to use knowledge bases, which are applicable to different types of applications. This latter point on portability and reusability goes in the same direction as the European initiatives on sharing the content of large lexicons.

Notes

[1] The results of the derivational morphology program output are checked against existing corpora and dictionaries, automatically.

Bibliography

Boguraev, B. and J. Pustejovsky (1990): *Knowledge Representation and Acquisition from Dictionary.* Coling Tutorial, August 16–18, 1990, Helsinki, Finland.

Mahesh, K., Nirenburg, S. (1995): A situated ontology for practical NLP. Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, *WCAI-95,* Montreal, Canada, August 1995.

Mahesh, K. (1996): *Ontology Development: Ideology and Methodology.* Technical Report MCCS-96-292, Computing Research Laboratory, NMSU.

Mel'cuk, I., N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja et A. Lessard (1984): *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexicosémantiques* I. Montréal: Presses de l'Université de Montréal.

Meyer, I., Onyshkevych, B. and L. Carlson (1990): *Lexicographic Principles and Design for Knowledge-based Machine Translation.* Technical Report CMU-CMT-90-118, Camegie Mellon University.

Nirenburg, S., V. Raskin, and B. Onyshkevych (1994): *Apologiae ontologiae.* Memoranda in Computer and Cognitive Science MCCS-95-281, NMSU.

Onyshkevych, B. and S. Nirenburg (1994): *The Lexicon in the Scheme of KBMT Things.* Technical Report MCCS-94-277, NMSU.

Pollard, C. and I. Sag (1987): *An Information-based Approach to Syntax and Semantics: Volume I. Fundamentals.* CSLI Lecture Notes 13, Stanford CA.

Viegas, E. and S. Nirenburg (1995): The Semantic Recovery of Event Ellipsis: its Computational Treatment. Proceedings of the Workshop on Context and Natural Language, *IJCAI 95,* Montréal, 1995.

Viegas, E., B. Onyshkevysh, V. Raskin and S. Nirenburg (1996a) *Submit* to *Submitted* via *Submission:* On Lexical Rules in Large-Scale Lexicon Acquisition. To Appear in Proceedings of *ACL'96.*

Viegas, E., Gonzalez, M., Longwell, J. (1996b) *Morpho-semantics and Constructive Derivational Morphology: a Transcategorial Approach to Lezical Rules.* Technical Report MCCS-96-295, CRL, NMSU.