The Latin WordNet project

Stefano Minozzi – Dipartimento di Linguistica, Letteratura e Scienze della Comunicazione
Università degli Studi di Verona
stefano.minozzi@univr.it

1. Introduction

This project of construction of a lexical knowledge-base for Latin language was born with the ambitious target to give a *specimen* of a Latin semantic network, trying to fill the gap constituted by the absence of such a resource, in order to open the possibilities of implementing new techniques of analysis derived from the studies in *Natural Language Processing*.

Through a semantic network a text can be approached not only as simple *data type* but can be *tagged* in order to process semantic-level *phenomena*, reconstructing a better model of textuality.

The implementation of a semantic network for Latin builds the possibility for experimenting new machine-driven activities. The MultiWordNet framework, which was our chosen model, offers various advantages in these fields of application:

- information retrieval: synonymy relations are used for query expansion to improve the recall of IR; cross language correspondences among synsets in the languages of MultiWordNet are used for Cross Language Information Retrieval.
- semantic tagging: MultiWordNet constitutes a large coverage sense inventory which is the basis for semantic tagging, i.e. texts can be tagged with synset identifiers.
- disambiguation: Semantic relationships are used to measure the semantic distance among words, which can be used to disambiguate the meaning of words in texts. Also semantic fields have proved to be very useful for the disambiguation task.
- ontologies: MultiWordNet can be seen as an ontology to be used for a variety of knowledge-based NLP tasks.
- terminologies: MultiWordNet constitutes a robust framework supporting the development of specific structured terminologies.

2. The MultiWordNet model

The MultiWordNet (MWN) project, as described in Pianta (2002), aims to build a number of language-specific semantic networks, maintaining the alignment with the *synsets*, groups of semantically equivalent words, available in Princeton WordNet (PWN)[1]. This task can be accomplished with the construction of new synsets aligned to PWN synsets, importing the semantic relationships that join the English synsets. The semantic connections among synsets have been considered as a constant through the languages and the words addressing a synset (a meaning) are the variables, as explained in Artale et al. (1997).

On this ground the project constitutes a multi-lingual lexical matrix (MLLM) as an extension of the bi-dimensional lexical matrix implemented in WordNet. A third dimension is added to the matrix, through which it is possible to consider different languages.

[1] Cfr. Miller et al. (1990), Fellbaum (1998)

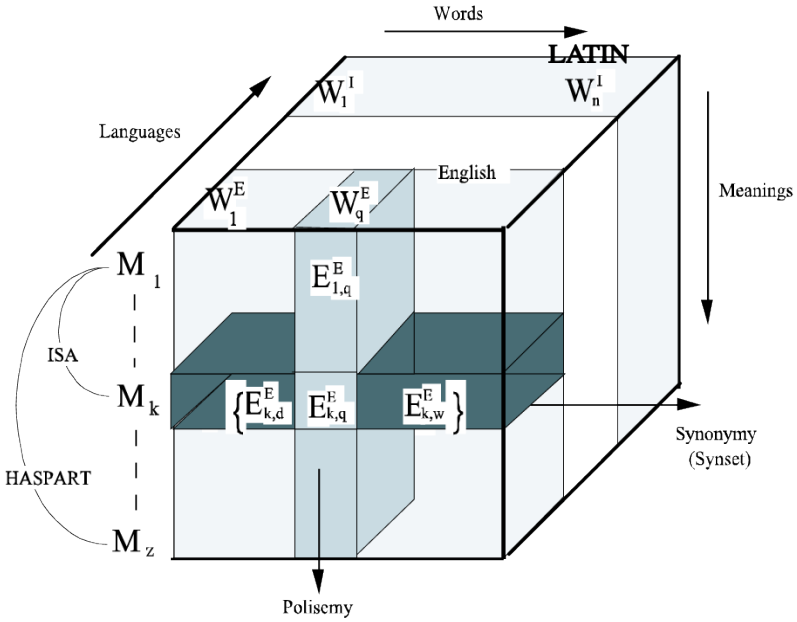

Figure 1: Multi-lingual lexical matrix

Figure 1: shows the three dimensions of the matrix: (a) words in a language, indicated by ; (b) meanings, indicated by ; (c) languages, indicated by . Moreover, the main lexical and semantic relations are visualised. From an abstract point of view, to develop the multilingual matrix it is necessary to re-map the Latin lexical forms with corresponding meanings (), building the set of synsets for Latin (making explicit the values for the intersections). The result is a complete redefinition of the lexical relations, while for the semantic relations, those originally defined for English is used as much as possible. From this point of view the dimension of meanings is considered constant in relation to the languages and words of each language. If for a certain for language *L* one obtains , with *i*=0…*t*, where *t* is the dimension of the lexicon of language *L*, this means that for language *L* there is no word that lexically realizes that meaning.

This model ensure a very great level of compatibility among different *wordnets* as stated in Vossen (1996). In fact if two wordnets are built independently for two different languages, they will exhibit differences which depend only partially on divergences among the languages. Some non trivial structural discrepancies will in fact depend on subjective decisions or different building criteria. The MWN model minimizes these discrepancies by strictly adhering to the PWN building criteria and subjective choices.

On the other hand the MWN model could drive a to an excessive dependence on the lexical structure of the English language. This is avoided allowing the creation of language specific synsets, in order to keep track of possible *semantic gaps* among the networked languages. An important advantage of the MWN model is the possibility of creating automatic procedures for the construction of synsets and for finding semantic gaps, specifically using the multilanguage character of MWN.

3. Building Latin WordNet

The construction of Latin WordNet (LWN) was based at first on an automatic assignment procedure.

Before this step the necessary lexical resources for building the controlled dictionary were gathered from a number of public available digital resources[2] and from the digitalization of written lexical resources. From those sources were obtained three large[3] machine readable dictionaries: a Latin-to-English MRD, an English-to-latin MRD and a Latin-to-Italian MRD.
Following the MWN model, our task was to build, when possible, a Latin synset which was semantically equivalent to a synset in PWN. Whenever this was not possible, a Latin-to-English or English-to-Latin lexical gap was discovered.
Similarly to what described in Pianta et al. (2002) for the construction of the Italian MultiWordNet, the Latin synsets can be built using different approaches.
The first strategy is based on English-to-Latin translating equivalents (TEs): for each PWN synset S, we look for the Latin TEs which are cross-linguistic synonyms of the English words of S. The union of such TEs is the Latin synonymous synset of S. If we cannot build any Latin synonymous synset for S, we have found an English-to-Latin lexical idiosyncrasy.
The second strategy is based on Latin-to-English TEs: for each sense σ of a Latin word L, we look for a PWN synset S including at least one English TE of L and we establish a link between L and S. When the procedure has been applied to all Latin word senses, we can build the equivalence class of all sets of Latin words which have been linked to the same PWN synset. Each set in the equivalence class is the Latin synset synonymous with some PWN synset. If, for a set of Latin synonyms there is no PWN synonymous synset, then we have found a Latin-to-English lexical idiosyncrasy.
A third approach, described in Minozzi (2008), was independently developed for the Latin WordNet project and exploits the multilingual nature of MultiWordNet. In this procedure a word is assigned to a specific synset when its translation equivalents, both Italian and English, belong to the same synsets in both the Italian and the English branches of MultiWordNet. The words assigned to a synset in this way have a greater degree of certainty to be the Latin lexicalization of the concept represented in the synset. The nature of the MRDs, which were collected for the project, drove primarily to the exploitation of the first and of the third approach.
In order to help the construction of Latin synsets we adopted a procedure that selects, for each sense of an Latin word, the PWN synsets which are most likely to have a comparable meaning, if any. Each described by the pair *<PWN synset, confidence score>*, where *confidence score* (CS) measures the degree of confidence in the link between the Latin word sense and the PWN synset
For a certain word sense listed in the Latin-to-English dictionary, this Assign-procedure considers the group of English words which are proposed as TEs for that word sense, and finds all the synsets containing at least one such TE. Such synsets constitute the set of candidates (CandSet) to be linked with the input Latin word sense. In other words the algorithm computes the CandSet of a certain Latin word meaning. The rest of the algorithm consists of ordering the CandSet by calculating the CS of each of its synsets. As in the MultiWordNet project this parameter was obtained considering a number of linking rules: *generic probability*, *gloss matching* and *synset intersection*[4].
The assignments of the data procedure were evaluated and controlled in order to improve the data-reliability. The work of correction began from the first thousand words included in the

[2] Special mention must be made to the dictionary collected by William Whitaker for the *Words 1.97* free latin dictionary (`http://users.erols.com/whitaker/words.htm`)
[3]About 40.000 entries for the Latin-to-English and the Latin-to-Italian dictionaries and about 20.000 entries in the English-to-Latin MRD
[4]cfr. Pianta et Al. (2002) and Minozzi (2008)

frequency/dispersion index of the *Frequency Dictionary of Classical Latin Words* by Gardner (1971) and was carried on to include all the words of the *Lessico Fondamentale Latino* by Riganti (1989). Evaluation and correction of the Latin WordNet were performed through the framework developed at Fondazione Bruno Kessler[5]: this framework permits easy connection of relationships involving words, management of morphological data and the constitution of a domain related hierarchy through WordNet Domains 1.6, which is described in Bentivogli et al. (2004).

4. Data structure and access

The data of Latin WordNet are structured with full compatibility with the MultiWordNet model: it is an extension of WordNet 1.6, the lexical database for English developed at the Princeton University.

Semantic level and morphological realization are separated through the database. The modular structure of the database reflects the theoretical principles of the multilingual semantic network: the semantic relations that are common among the included languages are stored in a COMMON-DATABASE and the language specific relations are stored in different modules (figure 2).

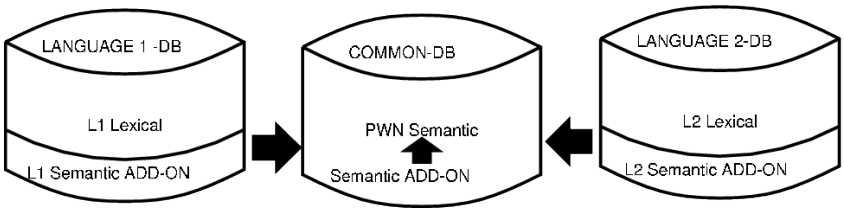

Figure 2: The data model

Latin WordNet contains information about the following aspects of the Latin lexicon:

- lexical relations among words;
- semantic relations among lexical concepts (synsets);
- correspondences among concepts with every language included in MultiWordNet;
- semantic fields (domains, via WordNet Domains).

The database is accessible online through the browsing interface[6] developed by Fondazione Bruno Kessler and is distributed through European Language Resources Association (ELRA)[7] in the form of a MySQL dump composed of six files, representing six relational tables which are compatible with MultiWordNet: common_relation.sql, latin_relation.sql, latin_synset.sql, latin_index.sql, latin_synonyms.sql, latin_morpho.sql.

The table "common_relation" lists all the semantic relations that are common to all languages. Each record contains four fields:

- type: kind of relation (see below the list of MultiWordNet relations and the corresponding symbols used to codify them);
- id_source: identifier of the source synset ("pos#offset", where pos is "n" for nouns, "v" for verbs, "a" for adjectives, and "r" for adverbs);
- id_target: identifier of the target synset ("pos#offset", where pos is "n" for nouns, "v" for verbs, "a" for adjectives and "r" for adverbs);

[5] http://www.fbk.eu/

[6] http://multiwordnet.fbk.eu/

[7] http://www.elra.info/

- status: "new" if the relation involves new synsets, i.e. synsets which are not in Princeton WordNet and have been created in MultiWordNet. When the relation involves synsets which are taken from Princeton WordNet the field is "NULL".

The table "latin_relation", that is actually undergoing construction, when finished, will contain the relations that are language dependent. These relations are instances of the standard lexical relations used in Princeton WordNet (e.g. antonymy, pertains to, etc.). Moreover, this table will contain a new type of semantic relation created within MultiWordNet, which is called "nearest". The nearest relation holds between an empty synset (a lexical gap) of a certain language and the synset with the most similar meaning in that language. There are only few instances of this relation codified so far, and are not yet available to the public.
Each record contains six fields:

- type: kind of relation (see below the list of MultiWordNet relations and the corresponding symbols used to codify them);
- id_source: identifier of the source synset ("pos#offset");
- id_target: identifier of the target synset ("pos#offset");
- w_source: the source lemma (only for lexical relations);
- w_target: the target lemma (only for lexical relation);
- status: "new" if this relation involves a new synset or "NULL" if it involves synsets taken from Princeton WordNet.

The table "latin_synset" contains the synsets (most of them are aligned with the Princeton WordNet but some are new ones). Also lexical gaps are specified in this file. Each record contains four fields:

- id: synset identifier ("pos#offset"). Either a Princeton WordNet identifier or a new synset identifier;
- word: lemmas contained in the synset, separated by a space character. The tokens of multiwords are connected by "_". The word "GAP!" is a special identifier used to describe a lexical gap;
- phrase: lemmas contained in the phraset, separated by a space character. The tokens of multiwords are connected by "_". This field is not used yet in Latin WordNet and is present for future development.
- gloss: synsets may optionally have a gloss, composed by a definition and sometimes an example.

The table "latin_index" contains the lists of the lemmas. The purpose of this table is to retrieve very quickly the synset ids and the possible searches starting from a lemma in all its PoS. Each record contains five fields:

- lemma: contains the lemma. Multiwords are connected by "_". The word "GAP!" is a special identifier used to describe a lexical gap;
- id_n: contains the list of the synset ids (separated by a space character) in which the lemma is contained as a noun;
- id_v: contains the list of the synsets ids in which the lemma is present as a verb;
- id_a: contains the list of the synset ids in which the lemma is present as an adjective;
- id_r: contains the list of the synset ids in which the lemma is present as an adverb.

The table "latin_synonyms" contains the lists of the synonym cards. Each record contains five fields:

- num: synonym card identifier;
- lemma: contains the lemma;

- pos: part of speech (could be "n" for nouns, "v" for verbs, "a" for adjectives and "r" for adverbs);
- syn: synset offset.

The table "latin_morpho" contains the list of the morphological information. Each record contains six fields:

- id: morphological card identifier (this id is used to join the morphological information to synonym card);
- lemma: contains the lemma;
- pos: part of speech (could be "n" for nouns, "v" for verbs, "a" for adjectives and "r" for adverbs);
- irregular_forms: useful for verbs;
- pronunciation (not used for Latin WordNet)
- miscellanea: other information like gender, number, ... (encoded from Whitaker's Words 1.97).

Most of the relations are the same as in Wordnet 1.6. Only the "nearest", "composed-of" and "composes" relations have been added. The complete list of the relations is showed in table 1.

POS	type	description	relation	features	coded into DB
NOUNS	!	Antonyms	antonym	:lexical	YES
	@	Hypernyms	hypernym		YES
	~	Hyponyms	hyponym		NO (see the reverse rel @)
	#m	Holonyms (* is a member of)	member-of		NO (see the reverse rel %m)
	#s	Holonyms (* is the substance of)	substance-of		NO (see the reverse rel %s)
	#p	Holonyms (* is a part of)	part-of		NO (see the reverse rel %p)
	%m	Meronymys (members of *)	has-member		YES
	%s	Meronyms (substances of *)	has-substance		YES
	%p	Meronymys (parts of *)	has-part		YES
	=	Attributes (is a value of *)	attribute		YES
	|	Synset nearest to *	nearest		YES
	+c	Composed-of (is composed of *)	composed-of	:lexical	YES
	-c	Composes (composes of *)	composes	:lexical	NO (see the reverse rel +c)
VERBS	!	Antonyms	antonym	:lexical	YES
	@	Hypernyms	hypernym		YES
	~	Hyponyms	hyponym		NO (see the reverse rel @)
	*	Entails doing	entailment		YES
	>	Causes	causes		YES
	^	Also see	also-see		YES
	$	senses of * grouped by similarity	verb-group	new 1.6	YES
	|	Synset nearest to *	nearest		YES
	+c	Composed-of (is composed of *)	composed-of	:lexical	YES
	-c	Composes (composes of *)	composes	:lexical	NO (see the reverse rel +c)
ADJ	!	Antonyms	antonym	:lexical	YES
	&	Similar to	similar-to		YES
	<	Participle of verb	participle	:lexical	YES
	\	Pertains to noun	pertains-to	:lexical	YES
	=	Value of (* is a value of)	is-value-of		YES
	^	Also see	also-see		YES
	|	Synset nearest to *	nearest		YES
	+c	Composed-of (is composed of *)	composed-of	:lexical	YES
	-c	Composes (composes of *)	composes	:lexical	NO (see the reverse rel +c)
ADV	!	Antonyms	antonym	:lexical	YES
	\	Derived from adjective	derived-from		YES

	|	Synset nearest to *	nearest		YES
	+c	Composed-of (is composed of *)	composed-of	:lexical	YES
	-c	Composes (composes of *)	composes	:lexical	NO (see the reverse rel +c)
	!	Antonyms	antonym	:lexical	YES

Table 1: Relations in MultiWordNet and in Latin WordNet

An offline version of Latin WordNet is undergoing development to make use of a morphological database in order to recall lemmas from inflected forms. This system was created for semantic tagging of full texts, in order to speed up the word recognition process.

5. Actual extension of the database

The actual database contains 9.378 words aligned in 8973 synsets which connects 143.701 arcs of relations. In table 2 is showed the amount of each part of speech represented in Latin WordNet.

	Nouns	**Verbs**	**Adj.**	**Adverbs**
SYNSETS	5621	2283	775	294
LEMMAS	4777	2609	1259	479
WORD SENSES	13060	10062	2054	732

Table 2: Extension of the database

The process of evaluation and correction of Latin WordNet is 35% complete. The actual project is evolving into the development of a tool for semantic tagging and information retrieval that is undergoing testing on documents from the A.L.I.M. database[8] (*Archivio della Latinità Italiana del Medioevo*).

6. Possible improvements

The data alignment and the correction of the automatic Assign-process are being constantly revised. We are developing a method for a better representation of words diacronicity in order to give reason of semantic shift in the different periods of the Latin Language. In order to extend the possibility of the integration with other online projects we are considering the creation of a web service for browsing and querying the database from third-party interfaces.

References

ARTALE, A., B. MAGNINI & C. STRAPPARAVA 1997. „Lexical discrimination with the Italian version of WordNet“. In: P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo & Y. Wilks (eds), *Proceedings of the ACL/EACL Workshop on Automatic Extraction and Building of Lexical Semantic Resources for Natural Language Applications*. New Brunswick, 32-39.

BENTIVOGLI, L., P. FORNER, B. MAGNINI & E. PIANTA 2004. „Revising the wordnet domains hierarchy: semantics, coverage and balancing“. In: G. Sérasset (ed), *COLING 2004 Multilingual Linguistic Resources*. Geneva, 94-101.

FELLBAUM, C. (ed) 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge.

[8] http://www.uan.it/Alim/

GARDNER, D. 1971. *Frequency dictionary of Classical Latin Words*. Stanford.

MILLER, G. A., R. BECKWITH, C. FELLBAUM, D. GROSS & K. J. MILLER 1990. „Introduction to wordnet: an on-line lexical database“. *International Journal of Lexicography* 3(4), 235-244.

MINOZZI, S. 2008. „La costruzione di una base di conoscenza lessicale per la lingua latina: Latinwordnet“. In: G. Sandrini (ed), *Studi in onore di Gilberto Lonardi*, Verona, 243-258.

PIANTA, E., L. BENTIVOGLI & C. GIRARDI. 2002. „Multiwordnet: Developing an aligned multingual database“. In: P. Vossen & C. Fellbaum (eds), *Proceedings of the first International WordNet conference*, Mysore, 293-302.

RIGANTI, E. 1989. *Lessico fondamentale latino*. Bologna.

VOSSEN, P. 1996. „Right or wrong: combining lexical resources in the eurowordnet project“. In: M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, & C.R. Papmehl (eds), *Proceedings of Euralex-96*, Goetheborg, 715–728.