Simple-Clips ongoing research: more information with less data by implementing inheritance

Riccardo Del Gratta, Nilda Ruimy, Antonio Toral

Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Via Moruzzi,1 56124
Pisa Italy
{riccardo.delgratta, nilda.ruimy, antonio.toral}@ilc.cnr.it

Abstract

This paper presents the application of inheritance to the formal taxonomy (*is-a*) of a semantically rich Lexical Resource (LR) based on the Generative Lexicon theory, SIMPLE-CLIPS. The aim is to lighten the representation of its semantic layer by reducing the number of encoded relations. A prediction calculation on the impact of introducing inheritance as regards space occupancy is carried out, which yields a significant space reduction of 22%. This is corroborated by its actual application that reduces the number of explicitly encoded relations in this lexicon by 18.4%. Later on, we study the issues that inheritance poses to the Lexical Resources and discuss sensitive solutions, illustrated by examples, to tackle each of them. Finally, we present a discussion on the application of inheritance, from which two advantages arise: consistency enhancement and inference capabilities.

1. Introduction

In order to encode semantic information in the four-layered Italian computational lexicon PAROLE-SIMPLE-CLIPS (PSC), a template-driven approach was adopted (Lenci et al., 2000). Such an approach, designed in the framework of the SIMPLE model, allowed providing a highly refined description of word meanings while ensuring a uniform, harmonized, consistent and model-conformant structuring of semantic information throughout the description of different languages. Coherence in the representation of meaning was all the more necessary because of the sensitive issue to be tackled, namely the semantic typing of heads and arguments. The template-driven methodology aimed at guiding the lexicographer by presenting her/him, after the semantic type assignment, a lexically underspecified schema of type-defining properties to be instantiated in each entry. This clearly contributed to speed up the encoding process to a certain extent. But still, since the database management tool did not allow for the computation of inheritance, every single property of each semantic unit (*SemU*) expressed either as features or semantic relations had to be explicitly specified in the corresponding entry.

Since its completion, the PSC lexicon has been constantly maintained and updated and has undergone periodic checks with a view to monitoring space occupancy and internal coherence. In addition, the recent formalization of the Generative Lexicon-based semantic layer, SIMPLE-CLIPS, into OWL has allowed assessing its internal structure and the correct application of Generative Lexicon principles (Toral et al., 2007). Besides, the use of the lexical resource in a number of applications has provided interesting critical feedback to the developers and has allowed to evaluate coherence and performance from both a linguistic and a database administration perspective.

2. Current research

2.1. The semantic database

The SIMPLE-CLIPS database consists of $57,000$ entries, $28,500$ of which are fully encoded with both mandatory and optional semantic features and relations foreseen by the SIMPLE model. For the sake of clarity, let us briefly remind that the core of SIMPLE relations is the *Extended Qualia Structure* which consists of four roles (*formal*, *constitutive*, *agentive* and *telic*). Each role subsumes a set of semantic relations[1] that encode the componential aspect of word meaning and capture the nature of the relationships holding among word senses.

Two main issues have emerged from the analysis of feedbacks and checks of the SIMPLE-CLIPS database:

1. the inexistence of some conceptual links, inexpressible through lack of appropriate representational vocabulary in the SIMPLE model (section 2.2.);
2. the high degree of information redundancy in the lexicon, especially as far as semantic relations are concerned (section 2.3.).

2.2. Missing links

To express the relationships holding between word senses, the SIMPLE model offers a large set of semantic relations. Among these are 60 *Extended Qualia* relations, the relevance and expressive power of which is largely acknowledged. These relations allow for the expression of very fine-grained distinctions, both for structuring the information regarding the componential aspect of word meanings and for capturing the nature of the relationships holding among word senses.

This relation set, however, does not provide the means to account for some conceptual links holding between events

[1] SIMPLE semantic relations relate two word senses or *SemUs*: $< sourceSemU >$ R $< targetSemU >$

and their participants and among co-participants in events, links which provide crucial information for Natural Language Processing tasks such as word sense disambiguation, text understanding, information retrieval, summarization and question answering.

2.3. Redundant semantic relations

A redundant representation of information in a lexicon is not necessarily to be seen as a drawback; it may be useful and even necessary in some cases. Information that could be perceived as merely redundant might actually permit to capture knowledge from different perspectives. However, although not all redundancy should be stripped off from a lexicon, no needless duplication of information should either make it uselessly heavy and cumbersome. This is unfortunately what happens in the SIMPLE-CLIPS semantic lexicon, due to the lack of inheritance computation. As a matter of fact, in terms of figures, the fully encoded set of entries has entailed the instantiation of $63,700$ semantic relations. For each single semantic unit, all properties expressed as semantic relations were in fact explicitly defined, although many of them, shared by a high number of *SemUs*, could have been inherited from their ancestors' entries. To give but a simple example, in the entry of the word *poodle*, a semantic relation links the word sense to the verb *to bark*, although the same relation is already encoded in the entry of *dog*, to which *poodle* is related by a hyperonymic link.

2.4. Actions

Research work has therefore been undertaken on both the above issues. On the one hand, the SIMPLE-CLIPS' relation network is now being enriched by some conceptual links holding between events and their participants and among co-participants in events (Ruimy, 2007); the expressive means to capture such relationships have been borrowed from the EuroWordNet/ItalWordnet model. On the other hand, efforts are being devoted to the implementation of the inheritance principle with a view to reducing redundancy and optimizing the lexicon format (Briscoe et al., 1993).

3. Evaluating the impact of inheritance

The ongoing implementation of the inheritance principle aims at remedying the situation illustrated in section 2.3. by defining the concept of *inherited relation*, i.e. a relation that is automatically assigned to *SemUs* subsumed by the word sense for which the relation is defined. Recording only those links which are typical and specific to each entry obviously hopefully brings about a dramatic reduction of the number of explicitly encoded relations.

The strategy followed consists in distinguishing between the *formal (is-a)* relation and the *constitutive*, *agentive* and *telic* orthogonal ones. In this way, the formal relation becomes the standard taxonomic hierarchy.

Since this taxonomy can be considered as a tree, we expect to have:

$$N_{edge} = N - 1$$

is-a relations for N *SemUs*, with $N \geq 1$.

In the SIMPLE model, semantic relations are assigned a weight value, viz. *prototypical* or *essential*. This weight value is not inherent to a qualia relation, though: it differs according to the relevance of such relation in the definition of a semantic type (ontology node). Type-defining relations are weighted as *prototypical*, while relations providing additional information (and, therefore, optional ones) are weighted as *essential*. During the encoding process, lexicographers must instantiate the type-defining relations characterizing the semantic type the encoded *SemU* belongs to, but may avoid the instantiation of additional, optional ones. If we exclude the *is-a* relation, which is, of course, mandatory, we find out that only 52% of the instantiated relations are *prototypical*. This is the reason why we have deemed necessary to go beyond the inheritance of the sole mandatory relations.

3.1. Space occupancy

This section shows a statistical prediction of space occupancy, followed by the actual result in the SIMPLE-CLIPS database.

Let $\langle k \rangle$ be the average number of *prototypical* orthogonal relations defined for the root *SemUs*[2], the total number of relations automatically gained by applying the inheritance principle is:

$$N_{rel} = N_{edge} \cdot \langle k \rangle$$

The N_{rel} number of relations represents *inherited relations*, i.e. relations which are not recorded each time in the database but they have been recorded only once.

Let us suppose to follow a path in the taxonomy with a total number of N_{path} *SemUs*. Adding $\langle M \rangle$ specific (*prototypical* orthogonal) relations for a *SemU*, at tree level i, the total number of relations at level $i+1$ is:

$$N_{increl}(i+1) = N_{path} \cdot \langle k \rangle + N_{path} \cdot \langle M \rangle$$

Regarding the level $i+1$, only the $\langle M \rangle$ specific relations were recorded in the database and N_{increl} relations were gained through inheritance. The expansion of the table which records the orthogonal relations is only of $\langle M \rangle$ whereas, at the moment, the same table records N_{rel} relations. The gain is therefore:

$$G = (N_{path} - 1) \cdot (\langle k \rangle + \langle M \rangle)$$

If we look at the current database, the table that records relations for *SemUs* contains $80,000$ rows and the mean space occupancy is $0.1Kb$ per row.

We have the following values for N_{path}, $\langle k \rangle + \langle M \rangle$[3], G and occupancy gain (θ):

$$N_{path} \sim 25,000$$

$$\langle k \rangle + \langle M \rangle \sim 0.7$$

$$G \sim 18,000$$

[2] root *SemUs* are semantic units belonging to the four top nodes of the ontology, i.e. ENTITY, CONSTITUTIVE, AGENTIVE and TELIC. Currently, the number of prototypical relations for root SemUs is 490.

[3] The value of 0.7 has been calculated from prototypical relations only.

$$\theta \sim 1.7Mb\,(22\%)$$

Hence, the expected space occupancy gain obtained by applying inheritance reaches 22%.
We applied this strategy to the current database. We calculated that, so far, $7,663$ instantiated links belonging to 51 types of semantic relations defined in the SIMPLE model (both *prototypical* and *essential*) can be physically removed from the descendant entries and inherited from their ancestor entries. In other words, more than 18% of the relations explicitly encoded in the database until now ($7,663$ / $42,349$[4]) could be implicitly instantiated by applying inference based on inheritance.
This result is coherent with the value obtained in the prediction calculus.

4. Inheritance issues

In the previous section we calculated the extent to which the implementation of inheritance principle can lighten the semantic data of the SIMPLE database; in this section, we point out challenges that this principle may cause and tackle them. We foresee the following issues:

1. relation $\tilde{\text{R}}$ has to be added, but with a different target;
2. the target of relation $\tilde{\text{R}}$ has to be replaced;
3. relation $\tilde{\text{R}}$ has to be removed.

In the first scenario, conflicts may emerge between the targets of relation $\tilde{\text{R}}$, while in the second and third situations two factors enter into play, namely the cardinality of relations and the acceptability of a given relation with respect to the other ones defined for a *SemU*. To limit semantic inconsistencies in relations defined at the descendant level, we add an *acceptance* flag. This flag is managed by lexicographers and set to false when the inherited relation does not apply to the current *SemU*. In such a case, the inherited relation is saved into the database also for the descendant *SemU*, but since its *acceptance* flag is equal to false, this relation is filtered out by the software. This means that only "accepted" relations are shown to the users.
To address the three issues above we have to answer the following questions:

1. Can we add the same relation $\tilde{\text{R}}$ with a different target?
2. Can we override the relation $\tilde{\text{R}}$?
3. How can we manage relation scope ?

The rationale of distinguishing between the *is-a* relation and the others is that, contrary to the formal relation, the cardinality of the orthogonal ones is not limited to 1. Therefore, if $\tilde{\text{R}}$ is an orthogonal relation defined for a *SemU* at level j and instantiated C times, every hyponym inherits C relations of type $\tilde{\text{R}}$. Then, to provide specific information about the hyponym, lexicographers have no need to overwrite the target of these inherited relations: they simply add any number of instances of the same relation they deem necessary, provided no semantic conflict exists between the target *SemUs* of inherited and new relations[5]. In this way, the relation(s) $\tilde{\text{R}}$ defined for one of the hyponym are the only relations of type $\tilde{\text{R}}$ explicitly instantiated for this hyponym in the database.
From the database management tool point of view, in the hyponym's entry, the new relation(s) of type $\tilde{\text{R}}$ are (explicitly) editable whereas the C inherited relations $\tilde{\text{R}}$ are *not*. Moreover, as explained above, lexicographers may edit inherited relations for special reasons.
Table 1 below, shows the type of cardinality available for relations. The first question can be answered by using cardinality, i.e. the allowed number of instances of each semantic relation entering in the definition of a semantic type. If a relation is defined as *RecYes* or *RecNo*, then, it can be instantiated more than once. Two *SemUs* linked by an *is-a* relation either share a semantic type membership or belong to semantic types related by a subsumption relation. So, when applying the inheritance principle to *prototypical* relations only, we do not care of (possible) conflicts among the $\tilde{\text{R}}$ relations instantiated for the direct descendant *SemU* (*SemU* child) and the $\tilde{\text{R}}$ relations inherited from its direct ancestor (*SemU* father). In fact, *prototypical* relations should be consistent when implemented for children *SemUs*. Here, consistency simply means that the target of the $\tilde{\text{R}}$ relations for both father and child *SemUs* share the same semantic type.

Relation mandatory value	**Cardinality value**
Yes	min 1, max 1
RecYes	min 1
No	min 0, max 1
RecNo	min 0

Table 1: Relation cardinality

Cardinality also contributes to answer question 2: in fact, if a relation $\tilde{\text{R}}$ is not recursive, we can specify for children *SemUs* the relation $\tilde{\text{R}}$ with a different target.
From both a theoretical and implementative perspective, points 1 and 2 represent the same problem and can be solved with the same strategy. In the database, in fact, we add records for both father and child *SemUs*. The difference is that, in 1, the record(s) for the child *SemU* are simply added to the father's ones, while in 2, the record for the child *SemU* overrides the father's ones.
To address this situation we introduce the notion of *scope* of relations. This concept allows both to answer point 3 and to manage relation overriding issues. Let us consider the following (abstract) example:

$$SemU_{father} \quad \tilde{R} \quad Target_{father}$$

Since $SemU_{child}$ *is-a* $SemU_{father}$ then:

$$SemU_{child} \quad \tilde{R} \quad Target_{father}\,, s = 0$$

[4]Total number of relations in SIMPLE-CLIPS excluding the *is-a* relations.

[5]Actually, we have decided not to implement semantic consistency among the target *SemUs* of inherited and new relations. This constraint would bound the set of possible values for target *SemUs*. We assume the consistency of lexicographers' encoding.

$$SemU_{child} \quad \tilde{R} \quad Target_{child}\,, s = 0$$

In the database, both information are recorded and displayed to the user. This is correct for point 1, but wrong for point 2, since only the record about $SemU_{child}$ has to be displayed. By using the scope, we can force the software to display only the desired information:

$$SemU_{father} \quad \tilde{R} \quad Target_{father}\,, s = 0$$

$$SemU_{child} \quad \tilde{R} \quad Target_{father}\,, s = 0$$

$$SemU_{child} \quad \tilde{R} \quad Target_{child}\,, s = +1$$

Thanks to the positive scope of the third record, only the $SemU_{child}$ with $Target_{child}$ information will be displayed to the user.[6]

4.1. Example with a recursive relation

In this section we use the following notation:

$$Rel_{name}\textit{(source,target)}$$

Suppose to analyze the semantic entry for *'doctor'*. At a given taxonomy level, say j, we have the following prototypical relation:

is_the_activity_of ('doctor', 'treat')

Since in the entry for 'surgeon', we have the relation: *is-a (surgeon, doctor)*, at level $j + 1$, we have the following inherited relation:

is_the_activity_of ('surgeon', 'treat')

Since this relation is recursive, we can add another prototypical relation of the same type:

is_the_activity_of ('surgeon', 'operate')

The table 2 below summarizes the above situation:

Relation	**type**	**scope**
is_the_activity_of (doctor, treat)	saved in db	0
is-a (surgeon, doctor)	saved in db	0
is_the_activity_of (surgeon, operate)	saved in db	0
is_the_activity_of (surgeon, treat)	saved in db	0

Table 2: Inheritance representation for recursive relations

No semantic conflict exists between the target of the inherited and the added telic relation since the semantic units 'treat' and 'operate' share many semantic properties, namely semantic type (PURPOSE_ACT), hyperonym : [*is-a ('to operate', 'to act')*, *is-a ('to treat', 'to act')*] and domain of use (Health_and_Medicine).

[6] scope $= 1$ means local validity. In such a case, the software finds out that the relation $\tilde{R}$ has been defined at child level with scope $= 1$ and shows to the user this *local* relation instead of inherited one.

4.2. Example with a non-recursive relation

Domain restriction and scope of relations play a crucial role in non-recursive relations.
In this situation, the relation $\tilde{R}$ defined for $SemU_{child}$ overrides the inherited $\tilde{R}$ relation. In other words, we can say that the relation $\tilde{R}$, when instantiated for $SemU_{child}$, further specifies the semantic unit target of the relation, while the inherited information (for $\tilde{R}$) is not relevant. This situation occurs when the $SemU_{father}$ is underspecified for relation $\tilde{R}$: an *underspecified* relation stands for a relation for which the target *SemUs* can assume one or more values within a given set of possible values. In this case, to better characterize the $SemU_{child}$, we have to specify the target of the $\tilde{R}$ at child level. Moreover, this more specified relation has to be valid for each child instance.
Let us consider the following example, from particle physics. In physics, a lepton is a sub-atomic particle with specific properties such as charge, spin and so on. There are three "families" of lepton, among which the most common is the *electron*. The electron is coupled to a nearly massless neutral particle called *neutrino*. The charged lepton (electron) has two possible spin states and so, two possible helicity states, while the neutrino is observed to be only *left-handed*.
In such a case, we may say that the helicity for electrons is "degenerated", meaning that the helicity can be one of the two possible values: *right-handed* and *left-handed*.
In SIMPLE, we could implement the following relations:

$$has\_helicity(lepton, degenerated)$$

Let us consider electrons and neutrinos: they are both leptons, but neutrinos can have *only* one of the allowed values for helicity. Since *is-a (neutrino, lepton)*, we have the following inherited relation:

has_helicity(lepton,degenerated)

This relation must be specified, since all neutrinos are *left-handed*:

has_helicity(neutrino, left-handed)

The table 4.2. below describes the above situation:

Relation	**type**	**scope**
has_helicity(lepton, degenerated)	saved in db	0
is-a (neutrino, lepton)	saved in db	0
has_helicity(neutrino, left-handed)	saved in db	1
has_helicity(neutrino, degenerated)	inherited	0

Table 3: Inheritance representation for non-recursive relations

In this case, the targets of the relation *has_helicity* are in an *is-a* relation and the target value for the child *SemU* has to be one of the allowed father *SemU*.

5. The SIMPLE-CLIPS Database and the Implementation of Inheritance

The implementation of the inheritance principle had a significant positive side effect, viz. a considerable enhancement of consistency of the SIMPLE-CLIPS database. A correct inheritance of properties being in fact highly dependent on a coherent encoding, a preliminary 'cleaning' and harmonization of the lexical data was in fact performed, with a rigorous consistency check of semantic relations and, in particular, of hyperonymic links. As to a possible coverage extension, it can reasonably be assumed that the insertion of new data should not generate further inconsistencies. Actually, lexicographers will now be less prone to encoding errors, as they only need to explicit specific (orthogonal) links in a word entry, but no more the relations already defined for their parent nodes.
On the other hand, a first cost-free benefit of implementing the inheritance principle is internal to the lexical resource. In fact, provided they bear the formal *is-a* relation, the $28,500$ SIMPLE-CLIPS entries not fully encoded will inherit orthogonal relations. Well then, right recently hyperonymic links were added to a large number of those entries, in the framework of the ILC project for linking the SIMPLE-CLIPS and the ItalWordNet databases. Besides, in the event of importing in the SIMPLE-CLIPS database new semantic units (along with their hyperonymic link) from other lexical resources, the new entries will freely acquire these inherited relations.

5.1. Exploiting inheritance in the enriched database

The extension of the SIMPLE-CLIPS database with new links holding between events and their participants and among co-participants in events is being carried out in a costless and low-effort semi-automatic way. This is possible thanks to the extraordinary richness of information of the SIMPLE model and the possibility offered by the lexicon management tool to investigate every single feature of the lexical data. Using existing syntactic and semantic information, the pairs of word senses candidate to a new relation are in fact automatically identified and extracted, through a tangle of queries and constraints (Ruimy and Toral, 2008).
The drawback of this enrichment process is, however, the considerable increase of the relation number in the database. More than $5,000$ new relations were in fact instantiated for the six relation types encoded so far, viz. *involved_agent*, *involved_location*, *involved_instrument*, *involved_result*, *role_instrument* and *role_location*. The instantiation of these new links, and particularly the one relating events to their agents, has determined a further exponential growth of information redundancy in the lexicon and has considerably increased the size of a large number of entries[7]. Some of them would now be scarcely manageable if all properties were explicitly represented. The enriched SIMPLE-CLIPS lexicon offers therefore an ideal testbed for assessing the impact of using the inheritance principle.
Essentially, our reasoning is the following one. Let A be the set of *SemUs* target of a given relation $\tilde{R}$; let K be the cardinality, i.e. the number of elements of A. The *is-a* relation creates a quotient set on A, by defining equivalence classes. This means, for example, that it is possible to find, in the K elements, M direct ancestors which divide the set A in M subsets. The relation $\tilde{R}$ can be instantiated only for these M direct ancestors[8].

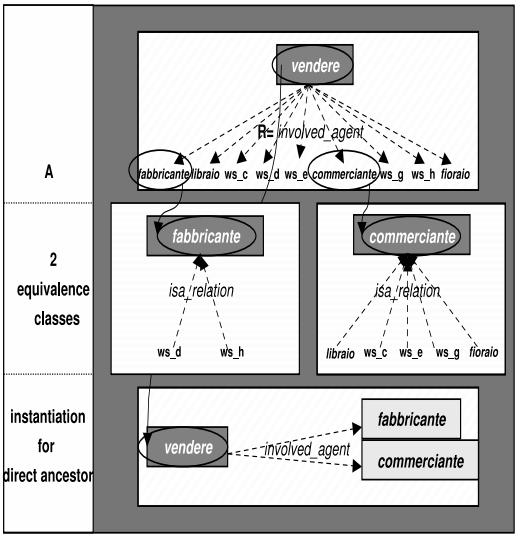

Figure 1: The *is-a* relation creates a quotient set

Let us take, for example, the verb *vendere*, 'to sell'. *Vendere* is the target *SemU* of the telic relation *is_the_activity_of* (which corresponds to the EuroWordNet relation *role_agent*) in 67 entries of profession-denoting nouns, among which 'bookseller', 'florist' but also more generic terms such as 'shopkeeper' or 'sales assistant'. Now, 62 out of these 67 entries are hyponyms of these last two words. Implementing the inheritance, these entries will inherit this telic relation from their hyperonym and 62 explicitly encoded relations will therefore be removed. The same holds for 63 building-denoting nouns, which will inherit from their hyperonym 'shop' the *role_location* relation whose target is *vendere*. As to the lexical entry for *vendere*, see figure 1, it includes now 69 semantic relations, among which the 67 newly added *involved_agent* ones. Implementing the inheritance, the verb will only be related to the most generic terms such as 'shopkeeper' or 'sales assistant' and the like and, again, 62 relations will turn inherited. So, while *vendere* was involved in 74 relations in the original database and in 273 ones after the addition of new links, thanks to the inheritance of properties, it will now be effectively involved in only 23 relations while the other 250 links will be inherited.

[7] This is, in particular, the case of high frequency verb entries, e.g. 165 *involved_agent* relations in the entry of the verb *lavorare* '(to work)'

[8] For direct ancestor we intend the nearest common parent for a list of elements of A.

6. Conclusions

This paper reports on the practical application of the inheritance principle to the *is-a* taxonomy of a semantically rich Lexical Resource, SIMPLE-CLIPS. The aim is to lighten its semantic representation by reducing the number of encoded relations, without losing any piece of information. By implementig inheritance, we avoid representing explicitly all relations for a semantic entry: those that are inherited from an ascendant do not need to be encoded.

A prediction calculation on the impact of introducing inheritance as regards space occupancy has been carried out, which yelds a significant space reduction of 22%. This is corroborated by its actual application that reduces the number of explicitly encoded relations in this lexicon by $18,4\%$.

Applying inheritance poses some issues to the Lexical Resource. However, these have been carefully studied and successfully solved, as discussed and clarified through examples. It is important to mention that besides the space reduction, the implementation of inheritance provides two notable advantages. On one hand inheritance enhances the consistency of the lexicon. On the other, it allows to gain further knowledge with respect to the one explicitly encoded by inference.

Finally, we have demonstrated the effectiveness of the inheritance principle by applying it to the enriched SIMPLE-CLIPS database.

7. References

Ted Briscoe, Valeria de Paiva, and Ann Copestake, editors. 1993. *Inheritance, Defaults, and the Lexicon*. Cambridge University Press, Cambridge.

Alessandro Lenci, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, and Antonio Zampolli, 2000. *SIMPLE Linguistic Specifications*.

N Ruimy and A. Toral. 2008. More semantic links in the SIMPLE-CLIPS database. In *Proceedings of the 6th Language Resources and Evaluation Conference*, Marrakech, Morocco.

N. Ruimy. 2007. Enhancing SIMPLE semantic relations: A proposal. In *Proceedings of 3rd Language & Technology Conference*, Fundacja Uniwersytetu im A. Mickiewicza, Poznań.

A. Toral, M. Monachini, and R. Muñoz. 2007. Automatically converting and enriching a computational lexicon ontology for NLP semantic tasks. In *Proceedings of 3rd Language & Technology Conference*, Fundacja Uniwersytetu im A. Mickiewicza, Poznań.