Cross-Lingual Validation of Multilingual Wordnets

Dan Tufiş, Radu Ion, Eduard Barbu, and Verginica Barbu

Institute for Artificial Intelligence, 13, Calea 13 Septembrie, 050711,
Bucharest 5, Romania
Email: {tufis, radu, eduard, vergi}@racai.ro

Abstract. Incorporating Wordnet or its monolingual followers in modern NLP-based systems already represents a general trend motivated by numerous reports showing significant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet or BalkaNet, represent one step further with great promises in the domain of multilingual processing. The paper describes one possible way to check the quality (correctness and completeness) of the interlingual alignments of several wordnets and pinpoints the possible omissions or alignment errors.

1 Introduction

Semantic lexicons are one of the most valuable resources for a plethora of natural language applications. Incorporating Wordnet or its monolingual followers in modern NLP-based systems already represent a general trend motivated by numerous reports showing significant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet and the ongoing BalkaNet, which adopted the Princeton Wordnet [1] as an interlingual linking device, represent one step further with great promises in the domain of multilingual processing. A general presentation of the BalkaNet project is given in [2]. The detailed presentation of the Romanian wordnet, part of the BalkaNet multilingual lexical ontology, is given in [3,4]. The EuroWordNet is largely described in [5].

Depending on the approach in building the monolingual wordnets included into a multilingual lexical semantic network and on the idiosyncratic properties of each language, the semantic alignment of the wordnets may be pursued and validated in several ways. We distinguish among syntactic and semantic validation methods.

Syntactic validation methods are concerned with checking whether a wordnet is structurally well-formed with respect to a set of rigorously and formally described restrictions such as: all the literals in a synset should have a legal sense identifier or, no literal with the same sense should appear in more than one synset or, there should be no dangling or unlinked synsets, and many others. Such kinds of errors are easy to spot, although not necessarily very easy to correct (especially when they are due to different granularity of the language resources used to build the wordnets). Semantic validation methods (in this context) rely on the notion of semantic equivalence between the word senses in two or more languages used to express the same concept.

2 Assumptions and the Basic Methodology

One fundamental assumption in the study of language is its compositional semantics. Compositionality is a feature of language by virtue of which the meaning of a sentence is

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 332–340.

a function of the meanings of its constituent parts (going down to the level of the constituent words). With this tarskian approach to meaning, our methodology assumes that the meaning building blocks (lexical items-single or multiple word units) in each language of a parallel text could be automatically paired (at least some of them) and as such, these lexical items should be aligned to closely related concepts at the ILI level. That is to say that if the lexical item W^i_{L1} in the first language is found to be translated in the second language by W^j_{L2}, common intuition says that it is reasonable to expect that at least one synset which the lemma of W^i_{L1} belongs to, and at least one synset which the lemma of W^j_{L2} belongs to, would be aligned to the same interlingual record or to two interlingual records semantically closely related.

As a test-bed, we use the wordnets developed within the BalkaNet European project and the "*Nineteen Eighty-Four*" parallel corpus [6] which currently includes four relevant languages for BalkaNet (with the prospects of extending the corpus to all the BalkaNet languages). This project aims at building, along the lines of EuroWordNet lexical ontology, wordnets for five new Balkan languages (Bulgarian, Greek, Serbian, Romanian and Turkish) and at improving the Czech wordnet developed in the EuroWordNet project. The methodology for semantic validation assumes the following basic steps:

A) given a bitext T_{L1L2} in languages L1 and L2 for which there are aligned wordnets, one extracts the pairs of lexical items that are reciprocal translations: $\{<W^i_{L1}W^j_{L2}>^+\}$;

B) for each lexical alignment of interest, $<W^i_{L1}W^j_{L2}>$, one extracts the synsets in each language that contain the lexical items of the current pair and respectively their ILI projections. For every lexical item recorded in the monolingual wordnets there will result two lists of ILI labels, one for each language, L^1_{ILI} and L^2_{ILI}. Based on the content evaluation of these two lists, several lines of reasoning might be followed highlighting various problems related to: the implementation of one or the other of the two wordnets, the alignment to the ILI; different sense granularity among wordnets; lexical gaps; wrong translation in the bitext, etc.

The first processing step is crucial and its accuracy is essential for the success of the validation method. A recent shared task evaluation (`http://www.cs.unt.edu/~rada/wpt`) of different word aligners, organized on the occasion of the Conference of the NAACL showed that step A) may be solved quite reliably. The best performing word alignment system [7] produced lexicons, relevant for wordnets evaluation, with an aggregated F-measure as high as 84.26%.

The content evaluation of L^1_{ILI} and L^2_{ILI} assumes a definition for the semantic distance between ILI records. Our system uses Siddharth Patwardhan and Ted Pedersen's WordNet-Similarity PERL module, a WN plug-in implementation of the five semantic measures described in [8].

3 Interlingual Validation Based on Parallel Corpus Evidence

If we take the position according to which word senses (language specific) represent language independent meanings, abstracted by ILI records, then the evaluation procedure of wordnets interlingual alignment becomes straightforward: in a parallel text, words which are used to

translate each other should have among their senses at least one pointing to the same ILI or to closely related ILIs. However, both in the EuroWordNet and in BalkaNet the ILI records are not structured, so we need to clarify what "closely related ILI" means. In the context of this research, we assume that the *hierarchy preservation* principle [4] is sound. This principle may be stated as follows:

if in the language L1 two synsets M_1^{L1} *and* M_2^{L1} *are linked by a (transitive) hierarchical relation H, that is* $M_1^{L1} H^n M_2^{L1}$ *and if* M_1^{L1} *is aligned to the synset* N_1^{L2} *and* M_2^{L1} *is aligned to* N_2^{L2} *of the language L2 then* $N_1^{L2} H^m N_2^{L2}$ *even if* $n \neq m$ *(chains of the H relation in the two languages could be of different lengths).* The difference in lengths could be induced by the existence of meanings in the chain of language L1 which are not lexicalized in language L2.

Under this assumption, we take the *relatedness* of two ILI records R1 and R2 as a measure for the *semantic-distance* between the synsets Syn1 and Syn2 in PWN that correspond to R1 and R2. One should note that every synset is linked (EQ-SYN) to exactly one ILI and that no two different synsets have the same ILI assigned to them. Furthermore, two ILI records R1 and R2 will be considered closely related if *relatedness*(R1, R2)=*semantic-distance* (Syn1, Syn2)$\leq$ k, where k is an empirical threshold, depending on the monolingual wordnets and on the measure used for evaluating semantic distance.

Having a parallel corpus, containing texts in k+1 languages (T, L_1, $L_2 \ldots L_k$) and having monolingual wordnets for all of them, interlinked via an ILI-like structure, let us call the T language as the target language and L_1, $L_2 \ldots L_k$ as source languages. The parallel corpus is encoded as a sequence of *translation units* (TU). A translation unit contains aligned sentences from each language, with tokens tagged and lemmatized as exemplified in Figure 1 (for details on encoding see `http://nl.ijs.si/ME/V2/msd/html/`).

```
<tu id="Ozz.113">
 <seg lang="en">
    <s id="Oen.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
        <w lemma="be" ana="Vais3s">was</w>      ... </s>
 </seg>
 <seg lang="ro">
    <s id="Oro.1.2.23.2"><w lemma="Winston" ana="Np">Winston</w>
     <w lemma="fi" ana="Vmii3s">era</w>     ...     </s>
 </seg>
 <seg lang="cs">
    <s id="Ocs.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
    <w lemma="se" ana="Px---d--ypn--n">si</w>  ...   </s>
  </seg>
  . . .
</tu>
```

Fig. 1. A partial translation unit from the parallel corpus

We will refer to the wordnet for the target language as T-wordnet and to the one for the language L_i as the i-wordnet. We use the following notations:

T_word = a target word;

T_word_j = the j-th occurrence of the target word;
eq_{ij} = the translation equivalent (TE) in the source language L_i for T_word_j;
EQ = the matrix containing translations of the T_word (k languages, n occurrences):

Table 1. The translation equivalents matrix (EQ matrix)

	Occ #1	Occ #2	...	Occ #n
L_1	eq_{11}	eq_{12}	...	eq_{1n}
L_2	eq_{21}	eq_{22}	...	eq_{2n}
...	...	...	...	...
L_k	eq_{k1}	eq_{k2}	...	eq_{kn}

TU_j = the translation unit containing T_word_j;
EQ_i = a vector, containing the TEs of T_wordin language L_i: (eq_{i1} eq_{i2} ... eq_{in})
More often than not the translation equivalents found for different occurrences of the target word are identical and thus identical words could appear in the EQ_i vector. If T_word_j is not translated in the language L_i, then eq_{ij} is represented by the null string. Every non-null element eq_{ij} of the EQ matrix is subsequently replaced with the set of all ILI identifiers that correspond to the senses of the word eq_{ij} as described in the wordnet of the i-language. If this set is named IS_{ij}, we obtain the matrix EQ_ILI which is the same as EQ matrix except that it has an ILI set for every cell (Table 2).

Table 2. The matrix containing the senses for all translation equivalents (EQ_ILI matrix)

	Occ #1	Occ #2	...	Occ #n
L_1	IS_{11} ={ILI_p| ILI_p identifies a synset of eq_{11}}	IS_{12} ={ILI_p| ILI_p identifies a synset of eq_{12}}	...	IS_{1n} = {ILI_p| ILI_p identifies a synset of eq_{1n}}
L_2	IS_{21} ={ILI_p| ILI_p identifies a synset of eq_{21}}	IS_{22} {ILI_p| ILI_p identifies a synset of eq_{22}}	...	IS_{2n} = {ILI_p| ILI_p identifies a synset of eq_{2n}}
...	...	...	...	...
L_k	IS_{k1} ={ILI_p| ILI_p identifies a synset of eq_{k1}}	IS_{k2} {ILI_p| ILI_p identifies a synset of eq_{k2}}	...	IS_{kn} = {ILI_p| ILI_p identifies a synset of eq_{kn}}

If some cells in EQ contain empty strings, then the corresponding cells in EQ_ILI will obviously contain empty sets. Similarly, we have for the T_word the list T_ILI = (ILI_{T1} ILI_{T2}... ILI_{Tq}).

The next step is to define our target data structure. Let us consider a new matrix (see Table 3), called VSA (Validation and Sense Assignment).

with VSA_{ij} = T_ILI $\cap$ IS_{ij}, if IS_{ij} is non-empty and $\perp$ (undefined) otherwise.

The i^{th} line of the VSA matrix provides valuable corpus-based information for the evaluation of the interlingual linking of the the i-wordnet and T-wordnet.

Ideally, computing for each column j the set SA_j (sense assignment) as the intersection $ILI_{1j} \cap ILI_{2j} \ldots \cap ILI_{kj}$ one should get at a single ILI identifier: SA_j=($ILI_{T\alpha}$), that is the j^{th} occurrence of the target word was used in all source languages with the same meaning,

Table 3. The VSA matrix

	Occ #1	Occ #2	...	Occ #n
L_1	VSA_{11}	VSA_{12}	...	VSA_{1n}
L_2	VSA_{21}	VSA_{22}	...	VSA_{22}
...	...	...	...	...
L_k	VSA_{k1}	VSA_{k2}	...	VSA_{kn}

represented interlingually by $ILI_{T\alpha}$. If this happened for any T_word, then the WSD problem (at least with the parallel corpora) would not exist. But this does not happen, and there are various reasons for it: the wordnets are partial and (even the PWN) are not perfect, the human translators are not perfect, there are lexical gaps between different languages, automatic extraction of translation equivalents is far from being perfect, etc.

Yet, for cross-lingual validation of interlinked wordnets the analysis of VSAs may offer wordnet developers extremely useful hints on senses and/or synsets missing in their wordnets, wrong ILI mappings of synsets, wrong human translation in the parallel corpus and mistakes in word alignment. Once the wordnets have been validated and corrected accordingly, the WSD (in parallel corpora) should be very simple. There are two ways of exploiting VSAs for validation:

Horizontal validation (HV): the development team of i-wordnet (native speakers of the language L_i with very good command of the target language) will validate their own i-wordnet with respect to the T-wordnet, that is from all VSA matrixes (one for each target word) they would pay attention only to the i-th line (the VSA(L_i) vector).

Vertical validation (VV): for each VSA all SAs will be computed. Empty SAs could be an indication of ILI mapping errors still surviving in one or more wordnets (or could be explained by lexical gaps, wrong translations etc.) and as such, the suspicious wordnet(s) might be re-validated in a focused way. The case of an SA containing more than a single ILI identifier could be explained by the possibility of having in all i-languages words with similar ambiguity.

We exemplify the two types of validation by considering English as the target language and Romanian and Czech as source languages. At the time of this writing the Romanian wordnet contains 11698 synsets (encoding 23571 literals), all linked to ILI records. The Czech wordnet is twice as large (25240 synsets and 37451 literals).

HV: The case study language is Romanian. For the validation purposes we selected a pool of 733 English common nouns appearing in Orwell's *Nineteen Eighty-Four* (out of 3167), because all their senses were implemented in the Romanian wordnet. There were 4319 occurrences of these words in the English part of our corpus and we built, as described in the previous section, 733 VSA vectors.

Almost half of the 4319 VSA_{ij} in the 733 vectors were empty. According to the procedure discussed in the previous section, when a VSA_{ij} contains an empty set, it means that none of the senses of the word eq_{ij} could be mapped (via ILI) to any of the senses of the target word. Although the analysis is not complete yet, we identified the following main explanations:

1. T_word and eq_{ij} are not related and the error is attributable to the human translator who used a wrong translation for T_word; we spoted only one such error (*darts/damă*) but systematically used four times.

2. T_word and eq_{ij} are not related and they were wrongly extracted as a translation pair by the word alignment program. By inspecting the TU_j it was easy to recognize this case and correct it; although these errors were not related to Wordnet development, and less than 15% of the analysed empty VSA_{ij} cells could be attributed to word-alignment errors, identifying them was beneficial for further development of the word aligner.
3. the right sense is defined for eq_{ij} but it has a wrong ILI identifier (it is wrongly mapped on ILI). By inspecting TU_j and sense glosses for eq_{ij}, the i-wordnet developer may easily identify the wrong mapping and correct it appropriately. This case is very relevant for the wordnet development and we estimate around 20% of the empty VSA_{ij} cells being explained by wrong mappings.
4. the synset linked to the relevant ILI record does not include the literal eq_{ij}, meaning that not all senses of eq_{ij} are defined in the i-wordnet and it happened that one of the missing senses was used in the TU_j. This situation is easy to recognize by a native speaker and the obvious solution is to add the eq_{ij} literal (indexed with the new sense number) to the proper synset. We estimate that this case (incomplete synsets) is responsible for almost 25% of all empty VSAs cells.
5. although none of the senses of T_word and eq_{ij} points to the same ILI identifier, one could identify a sense of T_word linked to ILI_α and a sense of eq_{ij} linked to ILI_β so that ILI_α and ILI_β are closely related. Closely relatedness was considered based on a maximum of two link traversals. This is what we call a *near-miss* interlingual linking. This case was the most frequent (we estimate it to more than 35%). The near-misses might be explained either by the translator's use of a more general or more specific Romanian word for the English word (e.g. because of lexical gaps or stylistic reasons) as in case of *prettiness/frumuseţe*, *bureaucrat/funcţionar*, *dish/farfurie*, *throat/gât*, etc. or by a misguided ILI mapping in the Romanian wordnet (still close enough) such as: *emotion/emoţie*, *hero/erou*, *event/eveniment* and several other real cognates. While translation licenses are inherent, coping with them is very important for the WSD task. The relatedness measure is an effective approach to decide which senses the T_word and eq_{ij} might have. The near-misses due to wordnet builders must be corrected. Most near-misses due to mapping errors show quite a regular pattern: when mapping a Romanian synset, the lexicographer had always as options at least two ILI records characterised by very similar glosses. As expected, looking up the PWN synsets corresponding to these ILI records, more often than not they were located in the same proximity (one hyponym/hypernym or meronym/holonym relation). Without additional information and based on subjective reasoning, lexicographers' introspection was wrong in several cases.

VV: The vertical validation is exemplified for English-Romanian-Czech**.** In order to see the potential of vertical validation procedure, we conducted a very small experiment on Romanian and Czech building the VSA for the T_world *country.*The 20 occurrences of the word *country* were translated in Czech by *země* (13 times), *venkov* (twice), *stát* (twice), *vlast* (twice), and once it was not translated. In Romanian, the occurrences of *country* were translated by the words *ţară* (12 times), *tărâm* (5 times), *stat* (twice) and once it was translated by a pronoun. The distinct triples of non-null mutual translations were the following:

1. <country ţară země> occurring eight times;
2. <country stat stát> occurring twice;

3. <country ţară vlast > occurring twice;
4. <country ţară venkov> occurring twice;
5. <country tărâm země> occurring five times.

Computing SAs for all triples above we obtained complete disambiguation for the first two of them (ten occurrences), all corresponding to the ILI record 171-07034213-n. The disambiguated translations of these 10 occurrences of *country* were:

1') <country:1 ţară:1 země:3>;
2') <country:1 stat:1.1a stát:3>.
The remaining triples generated empty SAs. However, they were disambiguated as near-misses as follows:
3') <country:1 ţară:1 vlast:1> – vlast:1 is a hyponym of země:3 and <country:1 ţară:1 země:3> is uniquely interpretable as 171-07034213-n. The contexts of these occurrences were:"...they betrayed their country..." and "...you betray your country...". This example show a near miss due to a lexical gap: neither English nor Romanian uses a single word for the concept of *own country*, unlike Czech.
4') <country:4 ţară:5 venkov:1> – both *country*:4 and *ţară*:5 are linked to the ILI record 171-07121548-n which is closely related to the one corresponding to ILI record 172-07121859-n standing for *venkov*:1. This latter ILI record is lexicalized in English by *countryside*, the first sense of which is a hyponym of *country*:4(rural area).
5') Finally, the third group of reciprocal translations was the most interesting. All the five occurrences were in the context of "...Golden Country..." (the fantasy land Winston Smith, the main character in "*Nineteen Eighty-Four*", was dreaming of). Between English and Romanian the near-miss was disambiguated as (country:5 tărâm:1) corresponding to the ILI record 171-06996512-n. Between English and Czech, the VSA_{ij}(country, země) = (171-07034213-n 171-06771212-n) and as such the near-miss was partially disambiguated as ((country:1 země:3)(country:3 země:6)). Since the distances between country:1 and country:5 or between country:3 and country:5 were beyond our considered threshold, the global near-miss could not be disambiguated. The conclusion we reached was that in the Czech wordnet there should be another sense for *země* (in the same synset with oblast:1, území:2 and prostor:2) in order to license translations as in the example below:
In his waking thoughts he called it the Golden Country/V duchu ji nazýval Zlatá země

4 Conclusions

This preliminary experiment shows that using translation equivalents extracted from a test-bed parallel corpus may precisely pinpoint various problems in the wordnets structuring and interlingual linking. A thorough quantitative and qualitative evaluation will follow the syntactic validations of the BalkaNet wordnets.

Recently the wordnets of the Balkanet project have been remapped on an ILI that corresponds to PWN2.0.

The methodology we discussed in this paper has been implemented in a Java program called *WSDtool.* In the present stage of the project we use it as a multilingual wordnet checker and specialized editor for error correction. Once the wordnets are validated, *WSDtool* can be

used to consistently sense-tag the entire multilingual parallel corpus (hence the name). For the most part, the sense tagging can be accomplished fully automatically; in those cases where it cannot, the human annotator is offered a small set of options from which to choose, thus reducing the likelihood of error. In the Appendix there is a commented snapshot from a horizontal validation session (English-Romanian) with WSDTool.

References

1. Fellbaum, Ch. (Ed.) (1998) WordNet: An Electronic Lexical Database, MIT Press.
2. Stamou, S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002): BalkaNet A Multilingual Semantic Network for the Balkan Languages, in Proceedings of the 1st *International Wordnet Conference*, Mysore.
3. Tufiş, D., Cristea, D. (2002): Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet, In Proceedings of *LREC2002 Workshop on Wordnet Structures and Standardisation*, Las Palmas, Spain, May, 35–41.
4. Tufiş, D., Cristea, D.: Probleme metodologice în crearea Wordnet-ului românesc şi teste de consistenţă pentru BalkaNet, în Tufiş, D., F. Gh. Filip (eds.) *Limba Română în Societatea Informaţională – Societatea Cunoaşterii*, Editura Expert, Academia Română, (2002) 139–166.
5. Vossen, P. (Ed.) (1999): *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer Academic Publishers, Dordrecht.
6. Erjavec, T., Lawson A., Romary, L. (eds.) (1998): East Meet West: A Compendium of Multilingual Resources. TELRI-MULTEXT EAST CD-ROM, 1998, ISBN: 3-922641-46-6.
7. Tufiş D., Barbu A. M., Ion R. (2003): A word-alignment system with limited language resources, Proceedings of the *NAACL 2003 Workshop on Building and Using Parallel Texts*; Romanian-English Shared Task, Edmonton, Canada, 36–39 (also available at: `http://www.cs.unt.edu/~rada/wpt/index.html#proceedings/`).
8. Budanitsky, A., Hirst, G. (2001): Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the *Workshop on WordNet and Other Lexical Resources*, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, June.

Appendix

The snapshot illustrates a horizontal validation (English-Romanian), the selected target word being "*shop*" and its translation equivalents in Romanian being displayed on the right part of the main screen. The first occurrence of "shop" appears in the Ozz.69 translation unit and clicking in the VSA cell corresponding to this occurrence on the **Check** and **Go** buttons several windows are opened:

1. the top most window shows the translation unit Ozz.69 with the translation equivalents highlighted (shops ↔ magazinele).
2. the partial networks in the Princeton Wordnet and Romanian Wordnet with the corresponding synsets as barycenters (right top and bottom left corners of the main window). Next to the barycenters are the entries in the two wordnets: [shop(1), store(1)] ↔ [magazin(1), prăvălie(1)].

The VSA cell exemplified contains one single ILI-record number (ENG171-03661978-n), signifying full disambiguation of the translation pair <shop, magazin>. The single common ILI-record number is pointed by the senses *shop(1)* and *magazin(1)*.

The VSA cell below the one exemplified contains the same ILI-record and everything discussed above holds true.

However, the VSA cell corresponding to the third occurrence of *"shop"* (visible at the bottom left corner of the main window) is empty. This occurrence of the target word was not translated in Romanian aligned sentence.

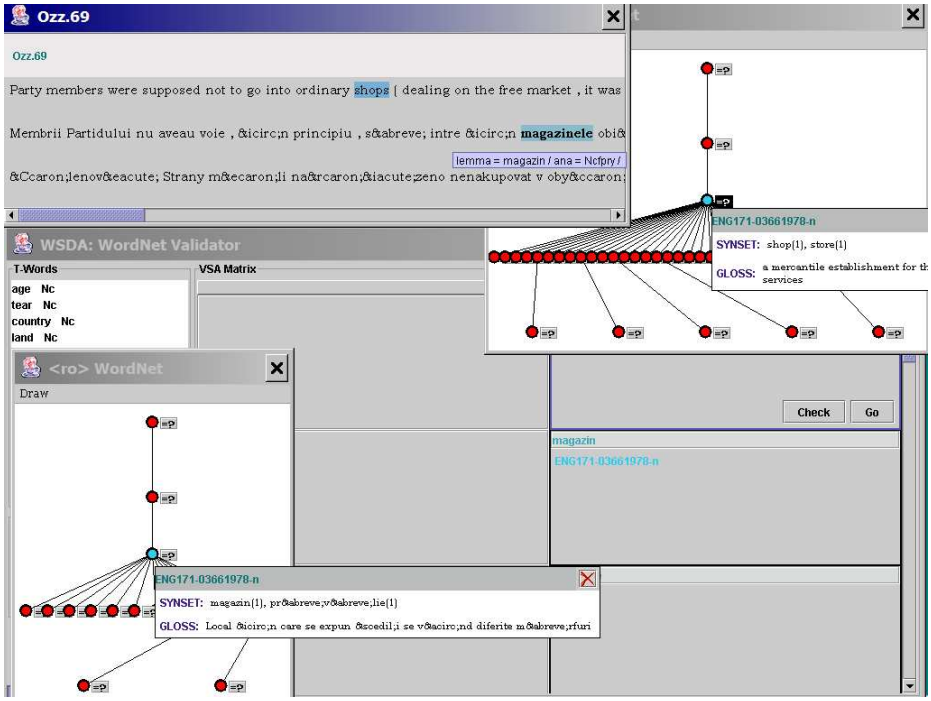

Fig. 2. A snapshot from a WSDTool HV session:
T-word is "shop", L_1 is Romanian, eq_{11} is "magazin" and VSA_{11} is {ENG171-03661978-n}