Building Multilingual Semantic Networks with Non-Expert Contributions over the Web

Nathaniel Ayewah
Southern Methodist University
Dallas, 75275-0122
ayewah@engr.smu.edu

Rada Mihalcea
University of North Texas
Denton, 76203
rada@cs.unt.edu

Vivi Năstase
University of Ottawa
Ottawa, ON, Canada
vnastase@site.uottawa.ca

ABSTRACT

We present a system that allows non-expert Web users to contribute towards building a multilingual lexical resource. Our study focuses on the Romanian-English language pair, and the target resource is a Romanian WordNet strongly connected to the English WordNet. We use a bilingual dictionary, a monolingual definition dictionary and documents on the Web to build synsets, attach them a gloss, and provide some examples. The results of our semi-automatic acquisition system are judged by two human judges, and they are compared to automatic approaches to building a Romanian WordNet.

1. INTRODUCTION

In order to obtain a system that provides expertise in a specific domain, the knowledge of that domain must be made available in a format that the system can use. Developers of software often do not have the knowledge of such specific domains, and experts in the field do not have the knowledge to create such a knowledge base. This has led to a new trend, in which software developers write tools that allow experts to readily formalize their knowledge through the system provided, which then encodes this input in a format that a system can use [7].

Language is a field that all people are experts in. We offer them RSDNET – a tool freely available on the Internet, with a friendly interface, through which Web contributors participate in the construction of a multilingual semantic network, by validating automatically suggested synonym sets.

We present in this paper the paradigm behind this system, the implementation and the interface, the role of the user, and an analysis of the results obtained so far. The results gathered were analyzed by two human judges, and compared to results obtained in other similar endeavors.

2. RELATED PROJECTS

The idea of harnessing the knowledge of experts in a particular field in order to gather data has found many applications.

The Rapid Knowledge Formation project [7] is geared towards providing experts in various fields with tools that allow them to encode their knowledge in an intuitive way, without needing to acquire programming skills. This is realized by using a graphical interface, which the experts manipulate to form and link concepts [4].

Collecting data over the Web for a variety of AI applications is a relatively new approach. The basic idea behind the broad *Open Mind* initiative [16] is to use the information and knowledge obtainable from millions of Web users to create more intelligent applications. *Open Mind* projects include our own effort – *Open Mind Word Expert* [3] – to build lexically annotated corpora through volunteer contributions. They also include *Open Mind 1001 Questions* [2], which acquires knowledge and *Open Mind Common Sense* [15], a system that collects common sense statements from Web users.

The domain of expertise our project is focused on is language. All native speakers of a language are expert users of their mother tongue. A structured system can help them focus on particular aspects, and harness their knowledge towards the construction of interesting resources. *OpenMind Word Expert* provides such a system, allowing people all over the world to contribute towards building a corpus annotated with semantic information [3].

WordNet [13] is a lexical resource that is used frequently in the NLP community for word-sense disambiguation, question answering and summarization, and other tasks.

It's success has led to projects aimed at building equivalent resources for other languages [19].

[19] show the process of building a multilingual resource based on WordNet 1.5. An inter-lingual index (ILI) provides the connection among WordNet and all the other resources in various languages that are being built. For each language, a core WordNet is manually built for a set of common base concepts. These sets are then enriched with semantic links, and they are expanded in a top-down manner [20]. ILI is however strongly connected to the original WordNet 1.5 resource, making it difficult to port the multilingual network to new WordNet versions. Moreover, ILI does not have the capability of storing word-to-word relations within a synset, and therefore the use of this resource for multilingual applications (e.g. machine translation, cross language information retrieval) is not always straighforward.

[8] show a way of building a semantic network using a monolingual dictionary, and then merging this structure with WordNet, in order to enhance it with the semantic links that WordNet provides.

The *Euro WordNet* project covers languages from western and central Europe (French, German, Italian, Spanish, etc.). *BalkaNet* is a similar project, focused on languages from eastern Europe (Romanian, Bulgarian, etc.). As opposed to the *Euro WordNet* endeavor which emphasized the multilingual nature of the project, *BalkaNet* allows the projects for each language to develop on their own.

[14] propose an automatic way of building candidate synsets in the target language (Bulgarian) using *WordNet*, an English-Bulgarian dictionary and a Bulgarian-English dictionary. The candidate synsets (called e-sets) are built by translating each English word in a synset into Bulgarian using the English-Bulgarian dictionary, and then choosing from the possible senses of the word by cross-referencing the results using the Bulgarian-English dictionary. A function is used to evaluate the goodness of the e-sets. Ultimately, a linguist chooses from the proposed candidates. The algorithm proposed was found to work well with nouns.

[10] use a similar process as [14] to build a *Romanian WordNet*. The algorithm they employ covers nouns, adjectives and verbs. Again, two bilingual dictionaries are used to translate words in synsets, and perform word-sense disambiguation between possible senses. The system developed is language independent, and free for tryouts [18], [12]. We use this system to test the results of our acquisition experiments.

[1] propose a semi-automatic approach to building ItalWordNet, in which a system uses the English WordNet and a bilingual dictionary to propose a linguist supervisor possible synsets. The user can also input language-specific synsets through a special interface.

3. RESOURCES

One of the most important objectives targeted by the RSDNET system design is to facilitate the task of the non-expert contributor as much as possible. That is, rather than asking the user to look for external resources for word translations, definitions, and examples, we try to provide several such resources directly on the system Web site. With such resources linked directly from the RSDNET page, the task of the users is greatly simplified – they select the right information from a pool of readily available information and usually do not have to seek additional resources.

3.1 WordNet

WordNet is the primary information source that we use in RSDNET for the construction of a new semantic network. WordNet is a Machine Readable Dictionary developed at Princeton University by a group led by George Miller [13], [9].

WordNet covers the vast majority of nouns, verbs, adjectives and adverbs from the English language. The words in WordNet are organized in synonym sets, called *synsets*. Each synset represents a concept. WordNet 1.7 is the latest WordNet version and it was released in July 2001. It has a large network of 144,680 words, organized in 109,373 synonym sets, called *synsets*. Table 3.1 shows the number of nouns, verbs, adjectives and adverbs defined in WordNet 1.7, and the number of synsets for each of these parts of speech.

Part of speech	Words	Synsets
Noun	107,929	74.487
Verb	10,805	12,753
Adjective	21,364	18,522
Adverb	4,582	3,611
TOTAL	144,680	109,373

Table 1: Words and synsets in WordNet 1.7

WordNet also includes an impressive number of semantic relations defined across concepts (249,425 relations in WordNet 1.7). For instance, the following relations are explicitly encoded in WordNet:

- Hypernymy/hyponymy relation (ISA), as in *tree* ISA *plant*.
- Meronymy/holonymy relation (HASA), e.g. *car* HAS-PART *airbag*.
- Antonymy, defined for all parts of speech, e.g. *beautiful* (vs.) *ugly*.

- Entailment, which is a pointer defined only for verbs, as *limp* entails *walk*.
- Pertainimy, involves adjectives, adverbs and nouns, and groups together words that are related, as *parental* pertains to *parent*.

Note that semantic relations are defined among *concepts*, and not among *words*, and therefore the belief is that the same semantic relations hold in any language, independent of the *words* that are used to lexicalize a given *concept*. The goal of RSDNET is to identify, with the help of Web users, these *concept* lexicalizations specific to a given language (e.g. Romanian), and build a resource similar in structure to the original English *WordNet* in a much shorter period of time than if starting from "scratch".

3.2 Bilingual dictionaries

RSDNET uses bilingual dictionaries to suggest translations for a given English word in a WordNet synset. We use a combination of several dictionaries that were identified online. Currently, RSDNET uses an English-Romanian dictionary with about 75,000 entries, out of which about 40,000 are word-to-word translations, and the rest represent phrasal translations. This dictionary is used to suggest candidate translations in Phase 1 in the Web interface, as described in section 4.

3.3 Monolingual Dictionaries

Once synsets words have been selected, RSDNET attempts to suggests definitions and examples for all the words in the synset. To this end, we are using a monolingual Romanian dictionary, consisting of about 35,000 definitions for the most frequent words in the Romanian vocabulary. In future versions of RSDNET, we plan to use an augmented monolingual dictionary, by integrating the output of "DEX online," a collaborative effort for building an online alphabetic Romanian dictionary initiated by Cătălin Frâncu[1].

3.4 Romanian Corpus

RSDNET makes several suggestions for synset word examples, to complete the synset gloss. Examples are extracted from a 400 million words corpus, consisting of a collection of Romanian newspapers collected on the Web over a three years period (1999-2002). Alternatively, RSDNET users can use search engines to directly identify examples on the Web. The RSDNET interface includes links to several search engines (currently, we link to Google, AltaVista, Lycos), and search queries are automatically formed with the synset words, for increased efficiency.

[1]http://dex.francu.com

4. WEB INTERFACE

The strength and lure of Web data collection systems is the seemingly limitless availability of users that possess the knowledge the system aims to acquire. The RSDNET Web interface aims to maximize the benefit received from this resource by providing facilities to simplify the data collection process for the user, increase the contributions by each user and minimize the occurrence of errors. The interface simplifies the data collection process by dividing it into phases and providing suggestions whenever the user needs to provide input to the system. It also uses a scoring system to reward users with recognition and prizes for significant contributions. An administrative facility allows an exclusive group of experts to review the inputs and make corrections where necessary.

4.1 The Phases

The RSDNET interface uses four phases to guide the user to their final destination of capturing a Romanian synset. These phases are transparent to the user and allow him or her to focus on a small part of the problem. Briefly, these phases allow the user to:

- choose a concept or synset to define,
- find appropriate lexicalizations of the concept,
- develop or retrieve definitions and sample sentences that match the concept, and
- review the inputs to eliminate errors.

All the phases rely on information from WordNet, and this information is always visible to the user. This simplifies the exercise to one of translating from English to Romanian, though the user may be able to identify other expressions of the concept in the target language that do not arrive as direct translations of the English words in the original synset. Unfortunately, this feature reduces the pool of contributors to the system, since they must be bilingual or at least have a good understanding of English.

The preliminary phase, *Phase 0*, displays a list of random English synsets from WordNet. Beside the set of words, the system also displays the synset's gloss from WordNet, which describes its meaning. For example, RSDNET may display a synset with the words *diversion, deviation, digression, deflection, deflexion* and the gloss *turning aside (of your course or attention or concern): "a diversion from the main highway"; "a digression into irrelevant details"; "a deflection from his goal"*.

When running this system, it quickly becomes obvious that some of the synsets in WordNet cover concepts that are not familiar to all users. So this phase includes a facility that allows the user to request another random list of synsets for her to choose from. Once a synset is selected, and the user constructs the equivalent Romanian synset, the synset is removed from the pool of "available" synsets and moved into a different set containing synsets to be validated.

Phase 1 directs the user to specify the Romanian words that belong in the chosen synset either by translating the words in the English synset or by providing words that are not direct translations of any of the English words. To speed up the process, RSDNET uses an internal English to Romanian dictionary (Section 3.2) to translate the English words. These are only suggestions for the user because the system cannot determine if the translations are correct in the context of the synset. For example, RSDNET translates the word *plant* as *plantă* which is correct if the synset refers to *living organism*, but not if the synset refers to *industrial plant*. It is the role of the contributor to decide on the right translation for a given synset word. In the example described above for Phase 0, RSDNET correctly suggests the words *deviere, deviere, digresiune, abatere* respectively for the first four words.

Figure 1 shows a screen shot of *Phase 1*. As an added benefit of this phase, the translations validated by the user create relationships between English and Romanian words within the context of the synset. These relationships are stored in a database and could eventually lead to a semantic English-Romanian dictionary.

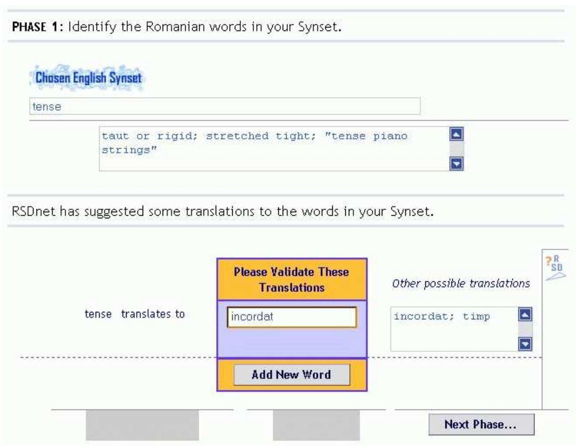

Figure 1: The top section of *Phase 1* keeps the English synset in view while the bottom section directs the user to make changes to Romanian translations if necessary.

In *Phase 2*, RSDNET uses an internal Romanian Dictionary (Section 3.3) and a textual corpus (Section 3.4) to suggest definitions and samples respectively. As in *Phase 1*, these suggestions may be out of context and the user needs to validate them or add new entries. *Phase 2* also provides a facility that allows the user to use popular search engines on the Web to retrieve sample sentences.

In the earlier *"diversion"* example, RSDNET does not suggest any definitions, so the user enters one: *o schimbare (în aten tie, a drumului, etc.)*. The system provides some samples from which the user selects *o deviere a drumului*, *o digresiune de la subiect*, and *o abatere de la calea cea dreaptă*.

Phase 3 directs the user to review the synset he or she has created to look for errors. Additionally, since Phases 1 and 2 allow the contributor to use simple html markup and to represent special characters using html, this phase gives a visual confirmation that the right markup has been used. From this phase, the user can finally submit his or her contribution to the RSDNET database.

In each of these phases, an online help system provides instructions for the non-expert user. This system also provides a reference for escape sequences that can be used to represent special Romanian characters. For example, the sequence '`\a`' is used to represent ă.

4.2 The Administrative System

A major concern with Web based data collection is the introduction of errors into the database because of a user's oversight, malicious intent or limited knowledge of the language. A color-coded administrative Web page was designed to allow select individuals to review and validate the entries into RSDNET, correcting or deleting them if necessary. Figure 2 shows a screenshot of this page, which also shows the administrator the original English synset and the relationships that have been created between English and Romanian words. A field in the database indicates which synsets have been validated.

4.3 The Scoring System

RSDNET, like similar projects at teach-computers.org, uses a rewards program to motivate users to make more contributions. When the user submits his or her synset in *Phase 3*, a score is computed based on the number of words in the synset. This is a very simple measure and can be thought of as a measure of the size of the synset. Future scoring schemes may consider the number of definitions and samples provided and penalize the user for incorrect entries. To attract new users and increase retention, we are giving away prizes, on a weekly or monthly basis.

4.4 Other Interface Features

Only users that are registered with RSDNET can improve their scores and win prizes. RSDNET provides a simple interface for registering with the system and updating personal information such as an email address.

The RSDNET interface also solicits feedback from contributors to look for ways to improve the system.

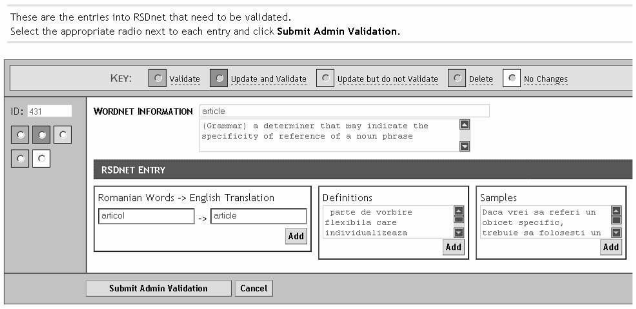

Figure 2: On the *Administrative Screen*, an expert can use the radio buttons on the left to make changes to the synset or remove it from RSDnet. The expert can also delete a field in the synset by leaving it blank.

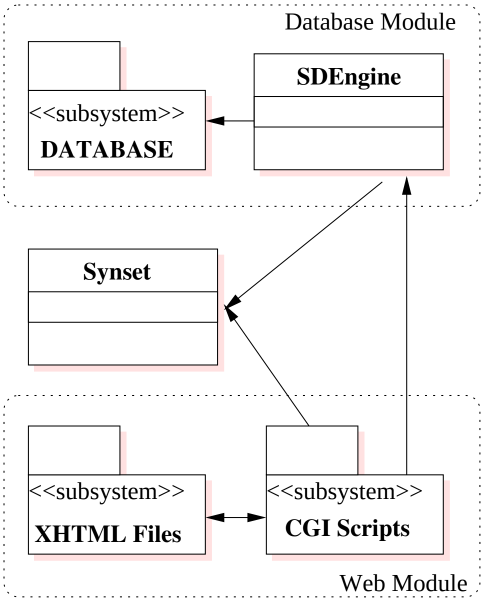

Figure 3: The High Level Design divides the system into two distinct modules which use a *Synset* object to encapsulate the information passed between them. *SDEngine* is a Perl object that contains all the queries used to access the database.

5. DESIGNING RSDNET

Figure 3 shows RSDNET's high level design which breaks the system into two modules: a Web module responsible for rendering and manipulating the RSDNET interface and a database module which controls all access to RSDNET's content. The one-directional arrow between the two modules indicates a 'client-server' relationship. The Web module (client) sends requests to the database module (server) to get information from the database. This design, along with a complete specification of the interface between the two modules, made it possible to develop and customize both modules simultaneously and relatively independently. The design aims to make it straightforward for other projects to interface with the modules.

The development of the RSDNET Web module focused on making it functional and extensible. This module is only a prototype and so, for example, does not provide secure login facilities to users adding entries to RSDNET. It does aim to be relatively fast and uses CGI scripts written in Perl which is ideal for rapid development and efficient at processing strings. To render the Web interface, this module uses xhtml files as templates which the CGI scripts populate with information from the database module. The look and feel of the interface is achieved using xhtml [6], cascading style sheets [5] and JavaScript.

The database module represents the more enduring aspect of RSDNET because it is more likely to be used in other projects especially when RSDNET becomes more comprehensive. It aims to be secure, reliable and well organized. A single Perl object, *SDEngine*, provides secure access to the database and contains all the queries that allow the Web interface to manipulate RSDNET's information. The database is designed as shown in Figure 4 to ensure that RSDNET can easily be used to perform many tasks including identifying Romanian synonyms (as a semantic dictionary), retrieving definitions for Romanian words (as a Romanian Dictionary) and providing English translations for Romanian words and vice versa (as a bilingual dictionary).

6. EVALUATION

We have compared the quality of the data obtained using RSDNET with data obtained from a system designed to automatically build a Romanian WordNet [10].

The comparison with automatically obtained data using GenSynsets [18] has led to a few observations. The fact that the ultimate judge in entering data is a human, bypasses most errors introduced by the lexical resources we use (the bilingual and the monolingual dictionaries). If for a certain word the dictionary does not provide a translation, the user can enter one himself. In the automatic approach, the system will produce no results for that particular synset, simply because the resources it has available are far from perfect. This is reflected in the difference between the accuracy numbers shown for GenSynsets in Table 2. The second set of results show a different run of the system when the dictionaries that the system used were manually edited to correct spelling and formatting errors. Also, GenSysets processes nouns, verbs and adjectives separately, and expects the dictionaries to provide separate entries for each of these parts of speech. We have used in the comparison the same dictionary that RSDNET uses, which

	RSDNET				GenSynsets			
	n	v	a	r	n	v	a	r
Correct synsets	96.6% (57)	100% (13)	92.3% (24)	100% (3)	18.5% (10) / 63% (34)	–	20.8% (5) / 71% (17)	–
Partially correct	3.3% (2)	0	3.8% (1)	0	1.8% (1) / 1.8% (1)	–	0 / 0	–
Erroneous	0	0	3.8% (1)	0	3.8% (1) / 3.8% (1)	–	4.1% (1) / 4.1% (1)	–
Missing	0	0	0	0	77.7% (42) / 33% (18)	–	75% (18) / 25% (6)	–
Total	59	13	26	3	54	–	24	–

Table 2: Results obtained with RSDNET for noun (n), verb (v), adjective (a) and adverb (r) synsets

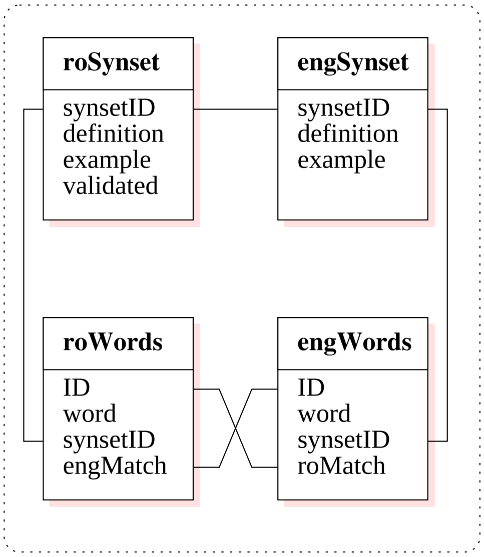

Figure 4: The Database maintains a one-to-many relationship between each concept (synset) and the lexicalizations (words) of the concept. At the same time, it maintains the relationships between English words and Romanian words within the context of a synset. A word can occur several times in each of the 'Words' tables if it belongs to more than one synset.

does not have part of speech information to allow us to separate the dictionary entries. Because of these issues, the automatic system produces more erroneous or has more missing synsets than it would with the appropriate dictionaries.

Table 2 shows the comparative results of RSDNET and the GenSynsets system, as evaluated by a human judge. 100 experimental synsets built using RSDNET have been manually validated by two human judges. RSDNET uses WordNet 1.7 as a reference, while GenSynsets was built to work with WordNet 1.6. A program automatically extracts the synsets in the 1.6 version of WordNet that correspond to the synsets translated using RSDNET. Some pairings between the two versions could not be made, and from the 100 synsets we have found 92 in the 1.6 version of WordNet. GenSynsets will work with these. Also, GenSynsets generates synsets only for adjectives and nouns, although theoretically the system also works for verbs [11].

7. WHAT'S NEXT

By proposing RSDNET, we choose a middle way between an automatic system, and a fully manual endeavour of building a semantic network of concepts. The pitfalls of the automatic approach come from the fact that it relies completely on imperfect lexical resources (namely dictionaries), which have a negative impact on the final results, as we have shown in section 6. The other extreme, a manual approach, is expensive in terms of time and human resources. We plan to compare the results of our semi-automatic acquisition with synsets created manually by Romanian linguists [17]. If the quality of our collection fares well in comparison with the one created by specialists (which is very likely, given the fact that the human judges validated the synsets collected until now with minor modifications, as was shown in table 2), RSDNET will be proven to be a worthwhile endeavour, given that the effort and the onus placed on the contributors is much less in the case of RSDNET.

Although RSDNET provides an environment for building a semantic network for Romanian based on a similar resource for English, the paradigm behind the system is generic enough to be applied for any pair of languages. The requirements are a resource for the original language to be modelled in the target language, and a bilingual and monolingual dictionaries for the target language. The existence of a corpus for extracting samples of usage would also be useful, but not indispensable.

The data collected in RSDNET does not consist only of synsets, but also of word pairs. The system keeps track of the English word and its Romanian translation, in the context of the synset to which the words belong. Such word-to-word translations could prove to be very useful in machine translation, cross language information retrieval, and other multilingual applications, since they show the lexicalization of a specific concept in Romanian and English.

8. ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of Ryan Farmer, Grace Kim, Elizabeth McLendon and Dr. Donald Evans from Southern Methodist University

who helped develop this system.

9. ADDITIONAL AUTHORS

Additional authors: Elizabeth McLendon, Southern Methodist University

10. REFERENCES

[1] L. Bentivogli, E. Pianta, and C. Girardi. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Myore, India, 2002.

[2] T. Chklovski. *Using Analogy to Acquire Commonsense Knowledge from Human Contributors*. PhD thesis, MIT, 2003.

[3] T. Chklovski and R. Mihalcea. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions", ACL 2002*, Philadelphia, July 2002.

[4] P. Clark, J. Thompson, K. Barker, B. Porter, V. Chaudhri, A. Rodriguez, J. Thomere, S. Mishra, Y. Gil, P. Hayes, and T. Reichherzer. Knowledge entry as the graphical assembly of components: The SHAKEN system. In *Proceedings of K-CAP 2001*, Victoria, BC, Canada, 2001.

[5] W. W. W. Consortium. Cascading style sheets, 2003. http://www.w3.org/Style/CSS/.

[6] W. W. W. Consortium. Hypertext markup language (html), 2003. http://www.w3.org/MarkUp/.

[7] DARPA. The rapid knowledge formation project, 2000. http://reliant.teknowledge.com/RKF/.

[8] X. Farreres, G. Rigau, and H. Rodriguez. Using WordNet for Building WordNets. In *Proceedings of the COLING-ACL 98 workshop on the Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, 1998.

[9] C. Fellbaum. *WordNet, An Electronic Lexical Database*. The MIT Press, 1998.

[10] F. Hristea. On the semiautomatic generation of WordNet type synsets and cluster. *Journal of Universal Computer Science*, 8(12):1047 – 1064, 2002.

[11] F. Hristea. On the semiautomatic generation of verb synsets in languages other than English. *Anals of the University of Bucharest*, ANO LII:75–86, 2003.

[12] F. Hristea. On the semiautomatic generation of WordNet type synsets and clusters with special reference to Romanian. *Building Awareness in Language Technology*, pages 113–140, 2003.

[13] G. Miller. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41, 1995.

[14] T. Nikolov and K. Petrova. Towards buildling Bulgarian WordNet. In *Proceedings of RANLP 2001*, pages 199–203, Tsigov Czark, Bulgaria, 2001.

[15] P. Singh. The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.*, Palo Alto, CA, 2002. AAAI.

[16] D. Stork. The Open Mind initiative. *IEEE Expert Systems and Their Applications*, 14(3):19–20, 1999.

[17] D. Tufis and D. Cristea. Methodological issues in building the Romanian WordNet and consistency checks in BalkaNet. In *Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation*, pages 35–41, Las Palmas, Spain, May 2002.

[18] G. D. Ungureanu, F. Hristea, and M. Popescu. GenSynsets, 2002. http://phobos.cs.unibuc.ro/roric/gensynsets.html.

[19] P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

[20] P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. The EuroWordNet Base Concepts and Top Ontolgy, 1998. Deliverable D017D034D036 EuroWordNet LE2-4003.