Cross-lingual Synonymy Overlap

Anca Dinu[1], **Liviu P. Dinu**[2], **Ana Sabina Uban**[2]
[1]Faculty of Foreign Languages and Literatures, University of Bucharest
[2]Faculty of Mathematics and Computer Science, University of Bucharest
anca_d_dinu@yahoo.com, liviu.p.dinu@gmail.com, ana.uban@gmail.com

Abstract

We investigate in this paper the degree of overlap between synonym sets of translated word pairs across three languages: French, English and Romanian. We use for this purpose a French Synonym Dictionary, a Romanian Synonym Dictionary, Princeton's WordNet and Google Translate API. We build a database containing pairs of (translated) words from the three languages, along with their corresponding synonym sets. We use it in order to gain insight into the synonym overlap for each language pair, and thus, into their degree of common concept lexicalization, by various queries. While the overall percentage of common synonyms is (expectedly) quite small (averaging ~6% across all language pairs), the percentage of hard synonyms pairs (pairs that have at least one common synonym), reaching ~62%, is significant. This is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on corresponding English versions of these lexical databases. Another interesting query topic was obtaining distributions of hard synonym pairs, function of their part of speech: hard synonyms were most frequent among verbs for English, and among adjectives for Romanian and French.

Keywords: cross-lingual synonyms, French, Romanian, database

1 Introduction

We investigate in this paper the degree of overlap between synonym sets of translated word pairs in three different languages, namely French, English and Romanian. The main idea is to test whether the synonym sets of pairs of translated words are still semantically related, that is to measure the degree of synonym overlap.

Synonymy is a lexical semantic relation, that is, a relation between meanings of words. By definition, synonyms are 'words or expressions of the same language that have the same or nearly the same meaning in some or all senses' (Inc., 2004). Cross-linguistically, the question that we try to answer in this paper is how much of this common meaning is shared by pairs of translated words. Since synonymy closely associates different lexicalizations of the same concept (which is language-specific), the overlap between synonym sets across a pair of languages expresses a kind of concept lexicalization overlap.

Cross-lingual synonym sets prove to be useful in tasks such as, for instance, automatic translation of web pages. Since search engines are using more of the Latent Semantic Indexing, which associates keywords of an article or a page with its synonyms within the domain covered by the keywords, one needs to take into consideration the synonym set of the translated keywords and the overlap of two languages synonym sets.

2 Related Works

There are various NLP applications using synonyms, one of the most notable being automatic synonym detection or extraction (Wang and Hirst, 2011; Wang et al., 2010; Mohammad and Hirst, 2006; Bikel and Castelli, 2008), a. o., which in turn can help in tasks including machine translation, information retrieval, speech recognition, spelling correction, or text categorization (Budanitsky and Hirst, 2006).

A multilingual approach based on word alignment of parallel corpora proved to have (Van der Plas et al., 2011) higher precision and recall scores

Proceedings of Recent Advances in Natural Language Processing, pages 147–152,
Hissar, Bulgaria, Sep 7–9 2015.

for the task of synonym extraction than the monolingual approach. Other work on semantic distance between words and concepts (Mohammad et al., 2007) emphasise on the advantages of multilingual over the monolingual treatment.

3 Data and Tools

For Romanian language, we used a synonym dictionary (Dicționarul de sinonime al limbii Române, by Luiza Seche and Mircea Seche), which contains about 45.000 words and 230.000 synonym pairs. For English language we employed Princeton's WordNet, version 3.0, which contains about 150.000 words and 250.000 synonym pairs. For French language we used the synonyms dictionary developed by the CRISCO research centre, which contains almost 50.000 words and 400.000 synonyms relations. As a translation tool we used Google Translate API. We stored the data in a MySQL database.

4 Methodology

In the pre-processing step, we extracted and cleaned the data in the Romanian and French dictionary, and removed multiword expressions, obtaining 42.277 Romanian words with a total of 230.445 synonym pairs, 44.884 English words with a total of 145.898 synonyms, and 39.564 French words with a total of 344.600 synonyms. Of these, we analyzed the words for which translations were available using the Google Translate API; the number of such words for each language is illustrated in Table 1 below.

	Total words	Translation pairs		
		EN	FR	RO
EN	44.884	-	25.048	19.454
FR	39.564	19.302	-	20.209
RO	42.277	19.654	23.207	-

Table 1: Number of words and translation pairs

As a pre-processing step, Romanian words were stripped of accents (though in normal usage of the language Romanian characters don't usually have accents, in the dictionary some words are marked with accents to indicate their pronunciation), but the diacritics were left as they were found. The translations obtained with Google Translate API needed to be cleaned by removing non-alphanumeric characters and by matching the case to the translated word's case (lowercase if original word was lowercase, capitalized if original word was capitalized). Articles were also removed from the nouns among synonyms and translations for all languages, as well as infinitive markers from the verbs (*a* for Romanian, *to* for English), and sometimes pronouns for the Romanian verbs, such as *i* (*a i se năzări*) or *o* (*o șterge*), so as to ensure the canonical dictionary form of the verb. Reflexive pronouns (*se*) were kept, because they mark reflexive verbs (which may have a different meaning than their non-reflexive variant). To make sure the translations returned by the Google Translate API are valid dictionary words (since the API does not guarantee this), we only accepted for each language translations which we could find as words or synonyms in our dictionaries for that language, and discarded the rest.

Synonymy was considered a symmetric property - that is, for each *(w, s)* word-synonym pair found in the dictionaries, *(s, w)* was added as a synonym pair as well. Translation was treated as symmetric as well: for any word-translation pair *(w, t)* from *language A* to *language B* as found using the Google Translate API, *w* was considered to be the translation of *t* from *language B* to *language A*. This assumption was used to fill in missing data where translations for some words in certain languages were not found by the API.

For each of the Romanian, French and English words in the dictionaries, we obtained their synonym sets. For the English words, the synonyms were extracted from WordNet, where words are organized in synonym sets (or "synsets"), the synonyms of an English word were considered to be all the words in the union of all the synonym sets that include that word.

In the case of homonyms or polysemantic words, we merged all the synonyms for each sense of the word together, thus obtaining unique word forms across the entire word set (for either of the three languages), each associated with one synonym set.

We extracted information on each word's part of speech. In the Romanian synonym dictionary, possible parts of speech are {noun, verb, adjective, adverb, pronoun, article, interjection, numeral, preposition, conjunction}. In WordNet, words can have one of 4 parts of speech: {noun, verb, adjective, adverb}. In the French dictionary, possible parts of speech are {noun, verb, adverb, adjective,

interjection, onomatopoeia, function word}. Considering we treated homonyms as the same word, for words where different senses of the word have different parts of speech, the word was considered to have multiple parts of speech.

For each pair of languages among the three languages analyzed, we generated word-translation pairs, we then computed statistics on their respective synonym sets, measuring overlaps between sets of synonyms from two perspectives: first translating the original word's synonyms in order to find their overlap with the translation's synonyms, and then translating the translation's synonyms in order to find their overlap with the original word's synonyms, resulting in two basic methods for measuring the synonyms' overlap.

Here are the steps we followed to obtain the statistics for word pairs and synonym sets, for a given pair of languages *language A* and *language B*, where *language A* and *language B* are both one of the three languages analyzed (English, French, Romanian): for each word in *language A*'s synonym dictionary:

1. We found its set of synonyms in *language A* (using *language A*'s synonym dictionary);
2. We obtained the word's translation into *language B* (given by Google Translate API);
3. We also obtained the set of synonyms for the *language B* translation (using *language B*'s synonym dictionary);
4. Finally, we found the translations in *language B* of the words in the *language A* set of synonyms (given by Google Translate API);

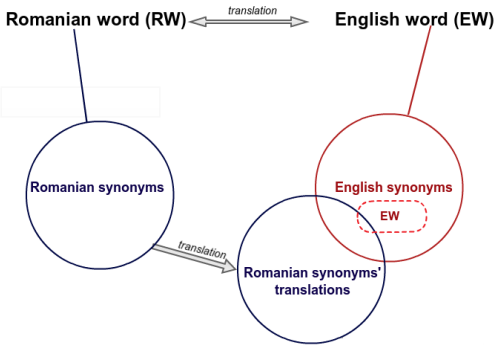

Figure 1: The method (for Romanian-English)

In order to test the overlap of *language A* - *language B* synonym sets, we counted the number of common words present in the synonym sets (consisting of words in *language B*) as computed above, for each word-translation pair. This process, exemplified for Romanian-English, is depicted in figure 1.

We applied the same algorithm the other way around. For each *language B* word the translation of which is found as an entry in the *language A* synonyms dictionary, one obtains its synonym set, its translation in *language A*, the synonym set for this translation and the translation into *language A* of the synonym set of the original *language B* word, then counts the common words present in these two resulted synonym sets (consisting of words in *language A*).

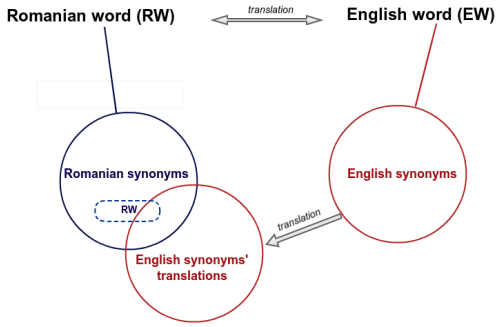

Figure 2: The method (for English-Romanian)

For measuring the intersections we used two methods: the first including only the synonyms of the two words (original *language A* word and its *language B* translation) and their translations, and the other including, along with the synonyms, the original target words as well (marked in the figures with the dotted border). We computed the overall percentage of common synonyms across synonym sets for all word pairs: for each word-translation pair, we measured the size of their joint synonym sets, as well as the size of these sets' overlap, as described above. We added these measures for all word pairs, and obtained the ratio of the number of common synonyms to the total size of all synonym sets.

We also counted the number of word-translation pairs for which at least one common synonym was found, or the synonym overlap contained at least one synonym (using any of the measures described above). These word pairs (along with their respective synonyms) will be called *hard synonyms*.

We organized the data in a MySQL database, in order to gain ease of access and to be able to instantiate various queries. The database consists of

two tables: the first is the Word table - containing all words (words in either language, that have an entry in the dictionary or were just found as synonyms), as well as information on their translation, language and part of speech. There is a uniqueness constraint on the pair of columns (word, language), reflecting the uniqueness of word forms described above. The second table is WordsSynonyms - containing synonymy relations as references to pairs of words in the Word table.

This database structure straightforwardly allows for queries such as, for instance, queries on synonym set overlap, function of the word pair's part of speech tag.

Other queries may also be formulated in order to compute various statistics on words and their synonyms, such as average number of synonyms for words, function of their language or part of speech.

An example of such a query, that extracts the common synonyms for the Romanian-English word pair *nebunie - madness*, is depicted in figure 3 below.

```
mysql> SELECT rw.word AS "RO word", tw.word AS "EN translation",
    ->        rsw.word AS "RO synonym",
    ->        tsw.word AS "Common EN synonym" FROM (
    ->                SELECT * FROM Word
    ->                WHERE is_headWord AND language="RO"
    ->             ) AS rw
    ->             JOIN WordsSynonyms AS rs
    ->                  ON rw.id=rs.word_id
    ->             JOIN Word AS rsw
    ->                  ON rs.synonym_id=rsw.id
    ->             JOIN WordsSynonyms AS ts
    ->                  ON (ts.word_id=rw.translation_EN_id AND
    ->                     ts.synonym_id=rsw.translation_EN_id)
    ->             JOIN Word AS tw
    ->                  ON rw.translation_EN_id=tw.id
    ->             JOIN Word as tsw ON rsw.translation_EN_id=tsw.id
    ->             WHERE rw.word="nebunie";
+---------+----------------+--------------+-------------------+
| RO word | EN translation | RO synonym   | Common EN synonym |
+---------+----------------+--------------+-------------------+
| nebunie | madness        | țicnealã     | folly             |
| nebunie | madness        | mișelie      | folly             |
| nebunie | madness        | scrânteală   | craziness         |
| nebunie | madness        | zârgheală    | folly             |
+---------+----------------+--------------+-------------------+
```

Figure 3: An example of a database query

5 Results

The overall percentage of synonym overlap ranges from 4% to around 9% and is highest for the English-French and the French-Romanian language pairs: 9,29% for English-French (from a total of 319.624 words in both synonym sets, a total of 29.703 words are common), and 6,95% for French-Romanian (26.303 words are common from a total of 378.604 synonyms). These results were obtained using the second method described in the previous section, (e. g. including the target words in the synonym sets).

The average percent of hard synonym pairs is approximately 46,6% - with high scores for French-Romanian and Romanian-French, as well as English-French. The total number of hard synonyms for French-Romanian is 10.870 covering 53,79% of all 20.209 word pairs, while for Romanian-French the proportion of word pairs that are hard synonyms is 44,01%, and 62,02% for English-French. This is encouraging, since hard synonyms may have potential use tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on corresponding English versions of these lexical databases. The percentages for Romanian and English are slightly lower (around 30%), as are those for the French-English language pair. Table 2 and 3 show the proportions of synonyms overlaps and hard synonym pairs respectively, for each of the language pairs considered and each of the two methods.

lang A	lang B	HS % (1)	HS % (2)
RO	FR	31,04%	44,01%
FR	RO	34,22%	53,79%
RO	EN	20,12%	33,36%
EN	RO	24,92%	46,85%
FR	EN	30,53%	39,86%
EN	FR	38,75%	62,02%

Table 2: Hard synonyms

lang A	lang B	Overlap%(1)	Overlap%(2)
RO	FR	3,79%	5,15%
FR	RO	4,51%	6,95%
RO	EN	3,05%	4,89%
EN	RO	3,67%	6,86%
FR	EN	3,31%	4,20%
EN	FR	5,96%	9,29%

Table 3: Total synonyms overlap

The distribution of hard synonym pairs, according to their part of speech, was also computed. The highest percentages of hard synonyms among words with a certain part of speech were obtained, in the case of language pairs including English (French-English, Romanian-English and their reversed analogues) for verbs, with as many as 74,03% of English verbs analyzed being part of an English-French hard synonyms pair (9.100 of 12.293 verb pairs). For French-English and English-French adverbs had the lowest pro-

portion of hard synonyms - 51,45% and 62,52% respectively, whereas for English-Romanian and Romanian-English, nouns (50,14%) and adjectives respectively (37,24%) had the lowest percentages of hard synonyms. This hierarchy may look surprising at a first glance. One possible explanation is that particular object lexicalization varies more across languages than more abstract concepts (such as properties or events) lexicalization. It can be argued that these numbers support the hypothesis that language acquisition proceeds from general (abstract) concepts towards particularizations, and not the other way around (from particular cases towards generalizations).

	RO - FR	FR - RO
HS%	57,50% adj	78,88% adj
	53,57% noun	74,78% verb
	52,77% verb	70,76% noun
	52,14% adv	70,56% adv

Table 4: Distribution of hard synonyms across parts of speech for Romanian - French pairs

	RO - EN	EN - RO
HS%	49,60% verb	55,63% verb
	49,58% adv	55,48% adj
	42,17% noun	51,81% adv
	37,24% adj	50,14% noun

Table 5: Distribution of hard synonyms across parts of speech for Romanian - English pairs

	FR - EN	EN - FR
HS%	62,20% verb	74,03% verb
	54,04% adj	68,20% noun
	52,52% noun	67,51% adj
	51,45% adv	62,52% adv

Table 6: Distribution of hard synonyms across parts of speech for French - English pairs

For French-Romanian, on the other hand, (as well as for its reverse), the highest proportion of hard synonyms was found among adjectives: 78,88% of French adjectives are hard synonyms. Since some of the words in our database can have multiple parts of speech, the distribution of most common tuples of parts of speech that occur toghether for the same word among hard synonym pairs was also computed. The (adjective, noun) tuple was found to be especially rich in hard synonyms for the French-Romanian and Romanian-French word pairs (with 74,91% of French words that are both adjective and noun being part of a French-Romanian hard synonym pair). Table 7, 8 and 9 show the most common such part of speech tuples found among hard synonyms for each language pair.

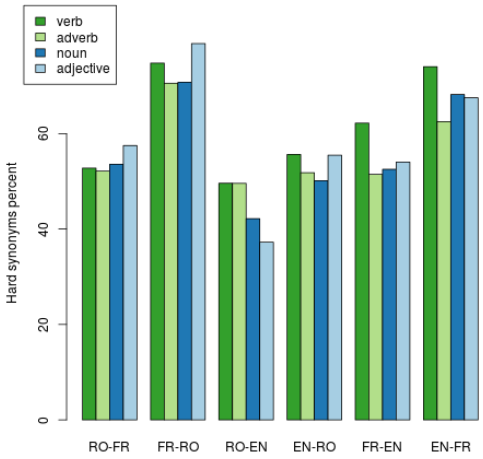

Figure 4: Hard synonyms proportion across parts of speech and language pairs

	RO - FR	FR - RO
HS%	53,63% adj,noun	74,91% adj,noun
	51,05% adj,adv	68,41% adj
	48,66% adj	65,04% verb
	44,75% noun	63,72% adv
	43,77% verb	59,71% noun

Table 7: Distribution of hard synonyms across words with multiple parts of speech, for most frequent combinations for French - Romanian pairs

6 Future Works

We leave for further research applying the same algorithm at deeper levels like synonym of syn-

	RO - EN	EN - RO
HS%	43,85% adv	63,49% adj,adv
	43,08% adj,adv	59,38% adj,verb
	40,06% verb	57,94% adj,noun,verb
	36,00% noun	50,77% adj,noun
	34,92% adj,noun	49,48% verb

Table 8: Distribution of hard synonyms across words with multiple parts of speech, for most frequent combinations for Romanian - English pairs

	FR - EN	EN - FR
HS%	58,28% verb	78,90% adj,noun,verb
	50,00% adj,noun	77,61% adj,adv
	45,58% noun	68,14% noun,verb
	45,47% adj	66,02% adj,noun
	44,20% adv	65,17% verb

Table 9: Distribution of hard synonyms across words with multiple parts of speech, for most frequent combinations for French - English pairs

onyms. Also, it would be interesting to test the distributional properties of the hard synonyms (as opposed to non-hard synonyms) on a parallel corpus. What one might hope to observe is a higher rate of co-occurrence of hard synonyms, since they express a common cross-lingual lexicalization of the same concept. Hard synonyms are also susceptible to be more reliable than non-hard synonyms with regard to the correlation between automatic word similarity judgements and human word similarity judgements.

7 Conclusions

We have presented a cross-lingual synonym overlap analysis for pairs of languages among three languages: French, English and Romanian, which can be quite straightforwardly extended for any other pair of languages. We have built a database containing pairs of (translated) words from the two languages along with their corresponding synonym sets and their synonym overlap set. Furthermore, we used it in order to gain insight into the synonym overlap of the three languages, and thus, into their degree of common concept lexicalization, by various queries. While the overall percentage of common synonyms is (expectedly) quite small (with an average of about 6% across all language pairs), the percentage of hard synonyms pairs (pairs that have at least one common synonym), as high as ~60%, is significant. This is encouraging for further use of this special kind of word translated pairs in tasks such as automatic enhancement of lexical databases (such as WordNet) for less resourced languages such as Romanian, based on corresponding English versions of these lexical databases. Another interesting query topic was obtaining distributions of hard synonym pairs, function of their part of speech: results varied with languages used in analysis: verbs had the biggest synonym overlap percentage for English hard synonyms (paired with any other of the two remaining languages), whereas adjectives and words that can be both adjectives and nouns were the most common for Romanian and French.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. Research by UEFISCDI, PNII-IDPCE- 2011-3-0959.

References

Daniel M Bikel and Vittorio Castelli. 2008. Event matching using the transitive closure of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 145–148. Association for Computational Linguistics.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Merriam-Webster Inc. 2004. *Merriam-Webster's collegiate dictionary*. Merriam-Webster.

Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 35–43. Association for Computational Linguistics.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *EMNLP-CoNLL*, pages 571–580.

Lonneke Van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.

Tong Wang and Graeme Hirst. 2011. Refining the notions of depth and density in wordnet-based semantic similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1003–1011. Association for Computational Linguistics.

Wenbo Wang, Christopher Thomas, Amit Sheth, and Victor Chan. 2010. Pattern-based synonym and antonym extraction. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 64. ACM.