# Evaluating the Wordnet and CoRoLa-based Word Embedding Vectors for Romanian as Resources in the Task of Microworlds Lexicon Expansion

**Elena Irimia**
RACAI
Bucharest, Romania
elena@racai.ro

**Maria Mitrofan**
RACAI
Bucharest, Romania
maria@racai.ro

**Verginica Barbu Mititelu**
RACAI
Bucharest, Romania
vergi@racai.ro

## Abstract

Within a larger frame of facilitating human-robot interaction, we present here the creation of a core vocabulary to be learned by a robot. It is extracted from two tokenised and lemmatized scenarios pertaining to two imagined microworlds in which the robot is supposed to play an assistive role. We also evaluate two resources for their utility for expanding this vocabulary so as to better cope with the robot's communication needs. The language under study is Romanian and the resources used are the Romanian wordnet and word embedding vectors extracted from the large representative corpus of contemporary Romanian, CoRoLa. The evaluation is made for two situations: one in which the words are not semantically disambiguated before expanding the lexicon, and another one in which they are disambiguated with senses from the Romanian wordnet. The appropriateness of each resource is discussed.

## 1 Introduction

The work presented in this paper was carried out in the broader frame of the ROBIN[1] project, whose aim is to develop systems and services for using robots in various contexts occasioned by the emerging digital society we live in. Focused on different types of robots, from those specialized in assisting elderly people to software robots dedicated to autonomous or semi-autonomous car-driving, ROBIN has a sub-component that deals with the essential function of human-robot language communication in Romanian (ROBIN-Dialog[2]). The prototype system for verbal interaction with the robot was restricted to several microworlds (see section 3) and the Romanian language resources and tools that are under development at the moment are specifically targeted at describing and serving these microworlds. As a result, the robot should be able to communicate successfully with the human users on topics concerning the specified microworlds and to perform some tasks designated to it, all these activities involving spoken Romanian.

The robot used in this project is Pepper, created by Softbank Robotics[3] and designed to be mass-produced and to become an important actor, improving human everyday life by assisting in different activities. Therefore, Pepper was intended to receive widespread acceptance in society and its shape, size, look and behavior were customized to emulate sociability (Pandey and Gelin, 2018).

A system able to ensure the dialog between a robot and a human user combines different modules dedicated to automatic speech recognition (ASR) (translating the human's vocal message into text), natural language processing (NLP) with its tasks of analysis and synthesis, a dialog management (DM) system and automatic speech generation from text (text-to-speech, TTS) (Tufiş et al., 2019). Except for the DM module, all the other components are language dependent, thus they need training on Romanian data and use (at run-time) Romanian acoustic and language models and a Romanian lexicon enhanced with information about stress, syllabification and phonetic transcription. In the context of the ROBIN-Dialog project, the acoustic and language models could benefit from all the available bimodal training data (see (Barbu Mititelu et al., 2018) for the description of the speech component of the Reference Corpus of the Contemporary Romanian Language - CoRoLa), but tailoring the system to the specific

[1] http://aimas.cs.pub.ro/robin/
[2] http://www.racai.ro/p/robin/
[3] https://www.softbankrobotics.com/us/pepper

microworlds is necessary for preventing semantic ambiguities and misleading. This can be done by designing a wide enough lexicon to cover various ways of expressing the semantic content possible in the targeted microworlds but limited to the semantic fields of interest (e.g., avoiding out-of-context senses for polysemous words). The process of constructing a lexicon - balancing all the needs of the dialog system modules in this specific context - is the focus of this paper.

## 2 Related work

One of the important steps in human-robot language communication is addressing the problem of creating exhaustive lexicons on different topics, so as to enable the robot to process different ways of expressing the same topic. WordNet is one of the main resources used for the enrichment of different domain specific vocabularies. Hiep Phuc Luong et al. (2009) presented a semi-automatic approach used to disambiguate the senses present in WordNet in order to enrich the vocabulary for ontology concepts in the domain of amphibians.

Other important resources used in expanding the lexicons are the word embeddings vectors extracted from different corpora. The main hypothesis on which the current models of semantic word representations are based is that words occurring in similar contexts have similar meanings (Clark, 2015). Moreover, such representations, most of the times, get closer to human intuitions (Agirre et al., 2009). Therefore, pretrained word embeddings vectors are used to a wide variety of NLP tasks, including vocabulary expansion. For example, Leeuwenberg et al. (2016) and Pennington et al. (2014) demonstrated that word embeddings are able to capture synonyms and analogies. Ono et al. (2015) used synonym and antonym information extracted from thesauri together with distributional information obtained from large scale unlabelled data in order to train word embeddings to capture antonyms.

We are not aware of any work in which the results obtained by using these two resources to be evaluated and this is one of our aims in this paper.

In what follows we define a microworld (section 3), describe the extraction of the lexicon from the screenplays based on two microworlds (section 4), we explain how we have expanded this lexicon using the Romanian wordnet and CoRoLa-based word embeddings (section 5), then we analyze the results obtained and discuss their relevance (section 6) before concluding the paper.

## 3 Designing microworlds

We define a microworld as an extremely reduced universe that is confined to a well-delimited space, is anchored in time, contains a finite set of objects, is populated by some people and the robot, among which verbal exchanges occur. These exchanges are on topics connected to the microworld. These people know how to collaborate with the robot, while the robot is meant to learn how to collaborate with the people. The learning phase of the robot needs to cover the following topics: the space topology, recognizing the people in the microworld, understanding natural language and reacting to it, which presupposes the ability to formulate an oral response to a human's command or to execute the command within the microworld.

For the present paper, we focus on two microworlds imagined for the interaction with the robot, which is attributed an assistive role: a private home and a research laboratory. In the former, the robot will help people to take care of themselves: undergo some measurements of relevance for their condition (e.g., measure the blood glucose), communicate the value of a certain measurement at a specific time, keep track of them during a longer period of time, display their evolution for a specified period of time, remind people what medication to take, when to do it and even where the medication is, etc. In the latter microworld, the robot will be the host for visitors of the laboratory, greeting, welcoming known people, introducing itself to the new visitors. It will also transmit verbal messages from one person to another, provided that they are present in the laboratory.

After designing the microworlds, a first preparatory step in the process of teaching the robot to interact with people is the creation of a screenplay for each microworld, with verbal interactions and actions. Our focus here is the former, namely the possible dialogues in natural language (Romanian) between people and the robot. A set of possible actions the robot could do in each microworld was identified and possible topics for verbal exchanges corresponding to them were created. This is to be understood at a conceptual level, while all the possible ways of expressing these topics are registered as their lexicalized forms. The robot must be able to understand them all, that is why it has

to be taught a large vocabulary and numerous syntactic structures. For example, the following ways of asking the robot if it knows the person called George were identified in Romanian: "Îl ştii pe George?" / "Ştii cine e/este George?" / "Îl cunoşti pe George?".

The human-robot communication is confined to the entities and possible activities in the respective microworlds. The robot will never initiate the dialogue. It is there to help the human by answering a question or carrying out a task. The robot is not to be understood as a repository of world knowledge. It can only answer questions about the entities in the respective microworld (such as "Este George în sala 306?" (Is George in room 306?), iff George is known to the robot, i.e. the latter was trained to recognize John's face, and the robot knows where room 306 is in the space whose topology it was taught). The human will give the robot as much information as necessary for performing the task, will formulate it concisely, clearly, avoiding obscurity and ambiguity.

Consequently, the vocabulary used in such dialogues cannot be conceived as specific to a domain. Terms specific to a domain, such as "blood pressure", "blood glucose", etc. may occur in the microworld with Pepper playing an assistive role in a private home. However, they are terms that have penetrated the general language, are familiar to every speaker, thus their domain specificity being drastically reduced.

## 4 Extracting the lexicon

Based on the screenplays mentioned in section 3, an initial list of lemmas was created. The screenplays serve as a corpus that was processed with the TEPROLIN platform (Ion, 2018), using the TTL module to normalize, sentence split, tokenize, POS-tag and lemmatize the data. Then, a list of all the unique lemmas in this annotated corpus was extracted, to serve as a starting point for the enhancement process described in the next sections. We treat differently the content words and the function words: we teach the robot the limited list of all the function words in Romanian, but we want to control the (virtually unlimited) set of content words the robot has to deal with, to stay in the discourse microworlds. Therefore, we set apart a list of 190 content lemmas to work with in our experiments.

For the final form of the extended ROBIN lexicon (containing a comprehensive list of lemmas that need to be represented in our resource), we added:

- all the morphological variants (i.e., inflected forms) of the words that were in the initial lexicon and of the words that were extracted using the two resources, by looking-up in an in-house extensive lexicon of Romanian (TBL, comprising 1.2 million hand-validated entries);
- all the Romanian function words (pronouns, determiners, articles, prepositions, conjunctions and some numerals, recovered also from TBL, 2382 entries);
- the information about stress, syllabification and phonetic transcription, generated with the TTS (see (Stan et al., 2011)) module from TEPROLIN.

## 5 Expanding the lexicon

The lexicon extracted from screenplays, as presented in section 4, was expanded with the purpose of enhancing it with words capable of capturing the lexical and syntactic varieties of the language. In order to extend the lexicon, two resources were used: the Romanian WordNet (RoWN) (Tufiş and Barbu Mititelu, 2015) and precalculated word embeddings vectors based on the CoRoLa corpus (Păiş and Tufiş, 2018). From the initial lexicon we chose only those lemmas that occur both in RoWN and in CoRoLa. We call this subset $L$ and it contains 178 content lemmas. The difference between the whole set of content lemmas extracted from the screenplays and $L$ is represented by foreign words (e.g. *cool*), proper nouns (e.g. *George*), and several content words not implemented in RoWN (e.g. the adverb *româneşte* "in a Romanian way").

We ran another experiment in which we semantically disambiguated the words in L. For six of them, no sense implemented in RoWN is the one with which the respective words are used in the screenplays. Consequently, we obtained a smaller set of 172 disambiguated words, which we call $L'$ ($L' \subset L$).

### 5.1 Using RoWN

RoWN has been created since the BalkaNet project (Tufiş et al., 2004). During this project, the aim was to cover the initial Base Concepts set

from EuroWordNet (Vossen, 2002). All their hyperonym synsets from Princeton WordNet (Miller, 1995; Fellbaum, 1998) (PWN) were implemented into RoWN. The literals are translated and their list is enriched with the help of synonymy and other dictionaries; the synsets glosses are mainly taken from the corresponding Romanian explanatory dictionary entries or, when such definitions could not be found to match exactly the PWN sense, the Romanian glosses were the translation of the English ones. More than 400 concepts considered specific to the Balkan area were included in the BalkaNet wordnets as synsets for which a hypernym was found among the synsets already implemented in the wordnets (Tufiş et al., 2004). The further quantitative enrichment of RoWN targeted the lexical coverage of various corpora collected over time (Tufiş and Barbu Mititelu, 2015). At the moment RoWN contains 59,348 synsets in which 85,277 literals (representing 50,480 unique ones) occur, out of which 20,031 (i.e., 17,816 unique ones) are multiword literals, accounting for 23.5% of the total number of literals (i.e., 35.3% unique ones). The qualitative enrichment focused on in-line importing of the SUMO/MILO concept labels (Niles and Pease, 2001), connotation vectors for synsets (Tufiş and Ştefănescu, 2012), derivational relations (Barbu Mititelu, 2013) and annotation of verbal synsets with labels specific to various types of multiword expressions, adopting the same framework (the PARSEME annotation guidelines) (Barbu Mititelu and Mitrofan, 2019). RoWN can be queried at `http://relate.racai.ro/` and at `http://dcl.bas.bg/bulnet/`, the latter offering also the possibility of visualizing aligned wordnets (Rizov et al., 2015).

Since wordnets are rich knowledge bases in which words and synsets are linked by lexical and semantic relations, we used the Romanian wordnet to attain broader lexical and semantic coverage of the scenarios created for the two microworlds, by extracting from it words semantically related to the ones in the screenplays. We call *semantically related words* those words occurring in synsets that establish one of the following relations with the synset(s) to which the words in $L$ or $L'$ belong: hypernym, cause, entailment, similar_to, verb_group, also_see, near_participle, near_derived_from, near_eng_derivat, near_pertainym, near_antonym. All relations whose name is prefixed with *near_* are considered language specific. They exist in PWN without this "prefix", i.e. they are participle, derived_from, eng_derivat, pertainym, antonym, respectively. When transferred into the RoWN this prefix served as a way of signaling that for Romanian the relation may not hold, although some semantic relatedness exists.

We disregarded for our task the following relations: hyponym, instance_hypernym, instance_hyponym, member_holonym, part_holonym, substance_holonym, member_meronym, part_meronym, substance_meronym, attribute, domain_TOPIC, domain_REGION, domain_member_USAGE, domain_member_REGION, domain_USAGE, domain_member_TOPIC. The reason for disregarding hyponymy is that a hyponym cannot replace its hypernym (Cruse (1986) showed that implication is unilateral in the case of hyponymy). Kleiber and Tamba (1990) showed that in the case of holonymy-meronymy, the relation of implication holds only when the predicate expresses location or time: in the following examples, (1) implies (2) and both of them express location. However, (3) does not necessarily imply (4), where the same words are used without reference to a place.

(1) The fly is on the child's *elbow*.

(2) The fly is on the child's *arm*.

(3) The child's *elbow* is on the table.

(4) The child's *arm* is on the table.

That is why we disregarded all types of holonymy and meronymy in wordnet. Instances are not relevant for our microworlds, just like all domain-related relations: the scientific domain to which a word belongs (the domain_TOPIC and domain_member_TOPIC relations), the geographical or cultural domain of a concept (the domain_REGION and the domain_member_REGION relations) or the usage of a word (the domain_USAGE and the domain_member_USAGE relations)[4]. As can be noticed in the definition of our understanding of semantically related words, we do not explore the wordnet graph on more than one level to look for related words, so that to avoid expanding the lexicon with too general words or with words seman-

[4]Some of these relations are language-specific, so there is no need to consider them; they were automatically transferred from PWN, without checking their applicability to Romanian data.

tically too distant from the ones in *L*.

### 5.2 Using Word Embeddings Vectors

It is known that neural word representations have the ability to capture useful semantic properties and linguistic relationships between words (Bakarov, 2018). On the basis of the Romanian reference corpus CoRoLa, which contains almost 1 billion words distributed in different text types and domains, and using distributed neural language model word2vec (Mikolov et al., 2013), high quality word embeddings vectors were generated (Păiş and Tufiş, 2018). We extracted and used the first 10 nearest neighbours to a given lemma in the word embedding space (semantically similar lemmas). The neighbours were obtained by computing a similarity score between the given lemma and the rest of the words in the vocabulary. The similarity score was obtained by the calculation of the cosine of the angles between two vectors; the closer the score is to 1, the more similar the two lemmas are.

## 6 Analysis of the Words Extracted from the Two Resources

The aim of this analysis is to discuss the relevance of the words extracted using the resources described in section 5 above for the task of extending the lexicon coverage for the two screenplays. A word is considered *relevant* if one can imagine a sentence that could fit the screenplays either for rephrasing an existing sentence or for completing the screenplay with further exchanges.

Both resources face the challenge of overgeneration: words tend to have more senses in corpora, while in wordnets they occur with many if not all their senses. However, in the screenplays they are mostly used with one of their senses, as the microworld could be thought of as a closed, limited domain. Having the expansion of *L* as a purpose, we discuss the results obtained without semantically disambiguating the words in the initial lexicon and then the results obtained after semantically disambiguating them ($L'$) .

For the sake of clarity, let:

- *n*=178 be the number of lemmas in *L*

- $n'$=172 be the number of lemmas in $L'$

- *A* be the set of *n* lemmas from *L* together with the set of lemmas of their related words in RoWN (the number of related words for each initial lemma varies and depends on the number of synsets identified as relevant and on their length):

$L\_lemma_1$: $rownlemma_{1,1}$, ..., $rownlemma_{1,i}$, ...

$L\_lemma_2$: $rownlemma_{2,1}$, ..., $rownlemma_{2,j}$, ...

...

$L\_lemma_n$: $rownlemma_{n,1}$, ..., $rownlemma_{n,k}$, ...

- similarly, $A'$ be the set of $n'$ lemmas from $L'$ together with the set of lemmas of their related words in RoWN;

- *B* be the set of *n* lemmas from *L* that were identified in CoRoLa together with the set of lemmas extracted from the word-embedding vectors (the number of related words for each initial lemma is set to 10, see section 5.2):

$L\_lemma_1$: $welemma_{1,11}$, ..., $welemma_{1,10}$

$L\_lemma_2$: $welemma_{2,1}$, ..., $welemma_{2,10}$,

...

$L\_lemma_n$: $welemma_{n,1}$, ..., $welemma_{n,10}$

- similarly, $B'$ be the set of $n'$ lemmas from $L'$ that were identified in CoRoLa together with the set of lemmas extracted from the word-embedding vectors.

We applied the following set operations to the two resources in order to find:

1. lemmas that could be obtained from both resources ($A \cap B$, and $A' \cap B'$ respectively): $\forall$ $L\_lemma_i$, $\forall$ $rownlemma_{i,j}$, $rownlemma_{i,j}$ is in $A \cap B$ if $\exists$ *k* so that $rownlemma_{i,j}$ = $welemma_{i,k}$;

2. lemmas that were obtained from RoWN but not from word embeddings vectors ($A \setminus B$, $A' \setminus B'$ respectively) and lemmas obtained using word embeddings vectors but not RoWN ($B \setminus A$, $B' \setminus A'$ respectively): e.g. $\forall$ $L\_lemma_i$, for each $rownlemma_{i,j}$, $rownlemma_{i,j}$ is in $A \setminus B$ if there is no *k* so that $rownlemma_{i,j}$ = $welemma_{i,k}$;

In what follows we discuss the results of these set operations.

### 6.1 Relevance of different word types for the screenplays

In the process of expanding the initial lexicon with new words, different types of words can prove their usefulness. The relevance of synonyms is self-evident. Hypernyms are known to replace a word in a context (Cruse, 1986), so their relevance is also clear. As far as antonyms are concerned,

they may allow for rephrasing the sentence with a negative form of the verb in Romanian: here is an example with the antonyms *continua* (go on) and *înceta (stop)*:

(5) a. *Continuă* să mergi! (*Go on* walking!)

b. *Nu înceta* să mergi! (*Don't stop* walking!)

Here is a set of examples showing the relevance of words derived from the word in the initial lexicon: the pair is *căuta* (verb, *to search*) - *căutare* (noun, *search, searching*), where the latter is derived from the former:

(6) a. Am *căutat* în camera 3316. (I *searched* in room 3316.)

b. Am făcut *căutarea* în camera 3316. (I made the *search* in room 3316.)

### 6.2 Words found in both resources

In this section we look, on the one hand, at the intersection of the sets *A* and *B*, showing the results without previous semantic disambiguation of the words in *L* (see column A ∩ B), and, on the other hand, of the sets $A'$ and $B'$ (see column $A' \cap B'$). We started our analysis with this step because we assumed words identified by both resources are probably the most interesting ones in terms of similarity with the initial lemmas, as they are enforced by both resources. Initial lemmas in Table 1 are words from L, respectively from $L'$, for which the intersection of the set of words extracted from RoWN and of the set of words extracted from word embeddings is not null. Comparing the number of initial lemmas in the set intersections with the number of elements in *L* (178) and $L'$ (172), we notice that for only 64% (Table 1 line 1 column 2) of the words in *L* and for 49% (Table 1 line 1 column 3) of the words in $L'$ we found words common to the both resources. This brings us to the conclusion that the two resources complement each other, rather than confirming each other's decisions in our task.

Validating the words found in the two resources, we notice that the rate of acceptance is quite high (95% and 100%, respectively - see no. of validated words from the no. of found words in Table 1), which confirms our intuitions that words identified by both resources are highly probable candidates. Adding the disambiguation criterion brings the probability of finding a good word in the intersection almost to 100%, eliminating all the bad results.

The validated words for the experiment involving semantic disambiguation are a subset of the validated words in the experiment without disambiguation. One might have expected these sets to be identical, i.e. only the synsets to which the disambiguated words belong offer relevant related words. However, the explanation for accepting (in the non-disambiguated setting) related words to other senses of the initial lemmas is that we understand synonymy in a broader way: any word that may imply *any* syntactic reorganization of the sentences in the screenplay, as long as the compositional meaning of the sentences is *almost* the same[5].

| **Types of words** | **A ∩ B** | **A′ ∩ B′** |
|---|---|---|
| no. of initial lemmas | 114 | 85 |
| no. of found words | 211 | 140 |
| no. of validated words | 201 | 140 |
| % of validated words | 95 | 100 |
| no. of validated empty lists | 0 | 0 |
| no. of synonyms | 103 | 73 |
| no. of antonyms | 25 | 14 |
| no. of derivations | 54 | 42 |
| % of synonyms | 51 | 52 |
| % of antonyms | 12 | 10 |
| % of derivations | 27 | 30 |

Table 1: Nondisambiguated vs. disambiguated sets intersection.

Regarding the types of words that are found, most of them are synonyms of the words in the scenarios. More synonyms are found in the first experiment, which means that senses that were not chosen in the word sense disambiguation phase of our work could also contribute relevant words, even synonyms. For example, for the verb *considera* (consider) the following related words were found and validated in the first experiment: *aprecia*, *susţine*, *crede*, whereas after disambiguating the initial lemmas, the only related word found was *crede*. However, although *susţine* could be accepted only for some contexts, we consider that *aprecia* is definitely worth being included in the lexicon. One explanation for this situation is the fine granularity of wordnets, which makes some senses to be too closely related and expressed by the same words. As a consequence of this granularity, several senses of a word should have

[5]Compare this with the definition of synonyms in (Miller, 1995): "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value".

been accepted in the disambiguation task, while at the RoWN level, the synsets should have been richer, sharing more literals. Besides synonyms, antonyms[6] were also found, although with a low rate. The high number of derived words reported for both experiments shows the importance of derived words in rephrasing the same semantic content, recognized by the two resources.

### 6.3 Words found only in RoWN

The next group of results we present are those that were found using RoWN, but not in the word embeddings. The data is summarized in Table 2.

| **Types of words** | **A \ B** | **A′ \B′** |
|---|---|---|
| no. of initial lemmas | 178 | 172 |
| no. of extracted words | 5130 | 1651 |
| no. of validated words | 843 | 840 |
| no. of validated empty lists | 26 | 33 |
| no. of synonyms | 563 | 469 |
| no. of antonyms | 27 | 31 |
| no. of derivations | 45 | 48 |
| % of synonyms | 66 | 55 |
| % of antonyms | 3 | 3 |
| % of derivations | 5 | 5 |

Table 2: Related words found only in RoWN.

A first remark is the large number of related words extracted from RoWN: for each word around 27 words, on average (see line 2 in Table 2), were extracted, due to the high number of relations used. However, many useless words (84%, see the no. of validated words as a percent of the no. of found words in Table 2) were extracted in the first experiment, whereas, as expected, the situation improved in the second experiment, in which only half of the extracted words were useless. We analyzed the invalidated words extracted: some of them are extracted by means of lexical not of semantic relations (see the discussion about relations prefixed with *near_* in the RoWN in subsection 5.1). Others are hypernyms that would seem unnatural in the screenplays, contrary to the linguistic expectations. The same inadequate usage characterizes some verbs from the same group as some initial verbal lemma. Although with these relations we also extract words that are useless for our task, we cannot eliminate them from the list of relations we need for expanding the lexicon, because they also return good words. We could not come up with any heuristic for deciding when to accept such relations and when to neglect them.

It is noteworthy that synonyms represent more than half of the total number of useful related words found in RoWN. Given the reduced average synset length in RoWN (that is 1.46, see (Tufiş et al., 2013)), we infer that the words in $L$ and $L'$ belong to longer synsets. This is something one could have expected, given the rather general character of most words occurring in the screenplays (see section 3 above for a discussion about the vocabulary of microworlds). Such words, belonging to the core vocabulary used by all people, are known to develop synonyms, derived words, to enter more expressions, to be semantically rich.

From the number of validated empty lists in Table 2 we understand that for those words no extracted word could be accepted as semantically related.

### 6.4 Words found only with word embeddings vectors

| **B \ A statistics** | **B \ A** | **B′ \A′** |
|---|---|---|
| no. of initial lemmas | 178 | 172 |
| no. of extracted words | 1600 | 1554 |
| no. of validated words | 737 | 656 |
| no. of validated empty lists | 21 | 19 |
| no. of synonyms | 46 | 40 |
| no. of antonyms | 47 | 32 |
| no. of derivations | 101 | 81 |
| % of synonyms | 6 | 6 |
| % of antonyms | 6 | 5 |
| % of derivations | 14 | 12 |

Table 3: Nondisambiguated vs. disambiguated B-A statistics.

For 178 initial lemmas, $B \setminus A$ extracted 1660 (and $B' \setminus A'$ extracted 1554) supposedly similar words from CoRoLa using word embeddings, from which 737 (and 656 respectively) were validated. We notice that although the number of extracted words is reduced considerably compared to the ones extracted from wordnet in the non-disambiguated setting, the number of validated words is lower, but close (737 vs. 843, 656 vs. 840). This implies that the two resources quantitative contribution to expanding the lexicon is similar, and, if done in the disambiguated setting, in-

[6] See (Ono et al., 2015) for extracting antonyms using word embeddings.

volves much less validation effort. While the differences in numbers and percents for the contribution of antonyms is negligible, what is evident in the data is that most of the synonyms come from the wordnet (see the 66% percent from Table 2 versus the 6% percent from Table 3) and most of the derivations come from the corpus (see 12-14% in Table 3 versus 5% in Table 2). Examples of initial lemmas whose list of extracted words abounds in derivated words are "robot" (robot) and "cântări" (weigh)[7]:

- *robot*: *robotiza*, computer, *robotic*, *roboţel*, *robotizat*, *robotică*, *robotizare*;

- *cântări*: *recântări*, gram, *recântărire*, greutate, *cântărit*, *cântărire*.

## 7 Conclusions

The experiments presented here prove the adequacy of RoWN and CoRoLa-based word embeddings for expanding a lexicon so as to ensure a wider lexical and syntactic coverage, meant to ensure the ability of a robot to understand humans in specific microworlds.

We worked with a list of 178 non-disambiguated initial lemmas (L) and with a list of 172 disambiguated initial lemmas (L') and we obtained a number of 1,694 unique lemmas ($A \cap B$) and, respectively, a number of 1,287 unique lemmas ($A' \cap B'$), extracted from RoWN and CoRoLa. The amount of validation work is substantially decreased in the disambiguated setting (even with the supplementary disambiguation costs) and, while such a solution is preferable in similar tasks, the loss in interesting, valid extracted words corresponding to different senses of the lemmas has to be taken into account. A solution would be to accept more senses for a specific lemma in the disambiguation phase, when the human validator considers it necessary. Words identified as related by the two resources are most probably good candidates, while in the disambiguated setting the probability of their usefulness is close to 100%.

As far as the contribution of different relations in wordnet is concerned, the way in which the task was formulated seems to have determined the acceptance of mainly synonyms (even if in a larger sense than that accepted by the wordnet projects), antonyms and words derived from the initial ones. Although a hyponym can be replaced by its hypernym, the need for precision can prevent this, whereas larger contexts would encourage this replacement as a means of avoiding repetition, which was not our concern in this experiment, as we did not focus on context, but on single sentences. The majority of synonyms was extracted from the wordnet, while the derivatives are mostly obtained from the corpus.

[7] Only the italicized words are derivationally related to the given ones.

## References


Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27, 2009.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. In *arXiv preprint arXiv:1801.09536*, 2018.

Verginica Barbu Mititelu. 2013. Increasing the effectiveness of the Romanian Wordnet in NLP applications. *CSJM*, vol. 21, no. 3, 320–331.

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of LREC 2018*, Japan, p.1178-1185.

Verginica Barbu Mititelu and Maria Mitrofan. 2019. Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet. In *Proceedings of the 10th Global WordNet Conference*, Wroclaw, Poland, (this volume).

Alan D. Cruse. 1986. *Lexical Semantics*. Cambridge, CUP.

Christiane Fellbaum. 1998, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Paul Grice. 1975. Logic and conversation. In Cole, P.; Morgan, J. *Syntax and semantics. 3: Speech acts*. New York: Academic Press. pp. 4158.

Radu Ion. 2018. TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In *Proceedings of the 13th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iaşi, 22-23 November 2018

Georges Kleiber, and Irène Tamba. 1990. L'hyponymie revisité: inclusion et hiérarchie. *Langages*, no. 98: L'hyponymie et l'hyperonymie, Larousse.

Artuur Leeuwenberg, Mihaela Vela, Jon Dehdari, and Josef van Genabith. 2016. A minimally supervised approach for synonym extraction with word embeddings. In *The Prague Bulletin of Mathematical Linguistics*, 111-142, 2016.

Hiep Phuc Luong, Susan Gauch, and Mirco Speretta. 2009. Enriching concept descriptions in an amphibian ontology with vocabulary extracted from wordnet In *22nd IEEE International Symposium on Computer-Based Medical Systems*, 1-6, 2009.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

George A. Miller. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, 2–9.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 984–989.

Amit Kumar Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of its Kind. *IEEE Robotics Automation Magazine*: 40–48.

Vasile Păiş and Dan Tufiş. 2018. Computing Distributed Representations of Words using the CoRoLa Corpus. In *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, vol. 19: 185–191.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543, 2014.

Borislav Rizov, Tsvetana Dimitrova, and Verginica Barbu Mititelu. 2015. Hydra for Web: A Multilingual Wordnet Viewer. In *Proceedings of the 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, Iaşi, Romania, 19–30.

Rion Snow, Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng. 2007. Learning to Merge Word Senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*: 1005–1014.

Mirco Speretta, and Susan Gauch. 2008. Using text mining to enrich the vocabulary of domain ontologies In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, 549–552, 2008.

Adriana Stan, Junichi Yamagishi, Simon King, and Matthew Aylett. 2011. The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate In *Speech Communication* vol.53 442-450.

Dan Tufiş, Dan Cristea and Sofia Stamou. 2004. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* Journal on Information Science and Technology, Special Issue on BalkaNet, Romanian Academy, 7 (1-2), 7–41.

Dan Tufiş and Dan Ştefănescu. 2012. Experiments with a differential semantics annotation for WordNet 3.0. In *Decision Support Systems* vol.53, no. 4, 695–703.

Dan Tufiş, Verginica Barbu Mititelu, Dan Stefanescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language Resources and Evaluation*, 47: 1305–1314.

Dan Tufiş, and Verginica Barbu Mititelu. 2015. The Lexical Ontology for Romaniany. In Nuria Gala, Reinhard Rapp and Gemma Bel-Enguix (eds.), *Language Production, Cognition, and the Lexicon*: 491–504.

Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Radu Ion, and George Cioroiu. 2019. Making Pepper Understand and Respond in Romanian. In *Proceedings of CSCS22* (in press).

Piek Vossen. 2002. *EuroWordNet general document version 3.* Report, University of Amsterdam.