ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 125–145

Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets

Dan CRISTEA[1], Cătălin MIHĂILĂ[2], Corina FORĂSCU[1],
Diana TRANDABĂŢ[1], Maria HUSARCIUC[1],
Gabriela HAJA[3], Oana POSTOLACHE[1]

[1] "Al. I. Cuza" University of Iaşi, Faculty of Computer Science
[2] Romanian Academy, Research Institute for Artificial Intelligence – Bucharest
[3] "A. Philippide" Institute for Romanian Philology, Romanian Academy – Iaşi

1. Introduction

The paper reports difficulties encountered during the alignment of synsets between English and Romanian. Reasons for these difficulties are inconsistencies found both in Princeton WordNet (we will refer to it from now on with PWN), on one part, and in our sources, on the other part, the difference in criteria based on which senses were recorded in PWN and in our sources, and the inherent differences in the lexicalization of concepts in the two languages.

The "word senses" problem, i.e. understanding when we can say that a sense is distinct from another, with its sub-problem of finding criteria for sense distinction, is a much debated one, see for instance [10], and has important consequences in the alignment of resources. Palmer et al. say recently [15]: "*even today, in spite of the proliferation of dictionaries, there is no methodology by which two lexicographers working independently are guaranteed to derive the same set of distinctions for a given word*", and then: "*The inherent fluidity of language ensures that a definition of a word is a moving target*".

One of the most frequent critiques brought to PWN is that sometimes word senses are so close together that a distinction is hard to be made, even by humans. Gonzalo et al. [8] confirm the too fine-grained sense distinction in both PWN and EWN for a majority of applications. They propose different types of sense clustering criteria applied to Information Retrieval. Mihalcea and Moldovan [13] also recognising the problems induced by the too fine-grained sense distinction in PWN, propose (language independent) principles for automatic transformation of PWN into a coarser grained dictionary, without affecting its semantic relations. A related problem is that of lexical

gaps. Bentivogli et al. [1] present a general methodology for handling lexical gaps when building aligned multilingual wordnets.

In acquiring the Romanian wordnet (ROWN) we have used a mixed merge-expand model, which is largely explained in [19], in this volume. We have collected our own synsets from different sources, tried to align the PWN synsets onto them, but, when necessary, have modified them in order to best match the linguistic concepts that had to be exemplified. In this paper we will exemplify the main situations encountered and show the solutions, wherever possible. We think that some decisions are general enough for being transposed to other languages in similar attempts to build PWN-aligned wordnets.

In this paper we will use the terms *sense*, *meaning* and *concept* as explained in [18], in this volume.

2. Word Senses in PWN

There are mainly two problems that anyone who immersed in building a new wordnet aligned with PWN faces constantly during this activity: the difference in the granularity of word senses in that person's native language compared to English and the conceptual gaps in one or the other language. In essence, these two issues have a common origin: the conceptual diversity of cultures, which is reflected in language. It is hard to know whether concepts evolved from words or word senses from concepts. It is obvious that words are names for concepts. However, in the case of young people or of realities that do not belong to one's cultural environment, it often happens that concepts are derived from the words that designate them. In many cases, children learn words and only later attach concepts to them. People enrich their conceptual diversity by exposure to language, which enables them to conceive new sensations or to form new mental representations of concrete things they have never seen. This is the case, for instance, of fabulous mythological creatures whose memory has been long gone from collective consciousness; yet, we are able to reconstruct concepts for them only because literature has preserved their names. Il may well happen that what we imagine now be very different from the original creatures. In any case, this new image is the result of our need to attach concepts to words we already have. In this respect concepts evolve from words. On the other hand, there is no individual who, while experiencing feelings or while thinking, had not been stricken by the difficulty to adequately express his feelings or thoughts because of the poverty of language. This need of human mind to use a name for every new sensation discovered led to the association of new concepts with existing words, based on every kind of analogy, some even mistaken, some completely invented, or to the extension or reduction of the existent sense of a word to designate larger or smaller concepts as well. All these had as a result more and more word senses. In this respect, word senses evolve from concepts. From this profitable vicious circle, languages have emerged and are continuously evolving. Human activities of all kinds have contributed to the interaction between human communities, which has allowed the development of a dense linguistic conceptual map. Then, interaction between languages, in language families or outside them, by processes of import and loan translation, contributed to

the acquisition of a rather rich intersection of word senses. However, differences exist and they are manifested in the granularity of senses and in conceptual gaps between pair of languages.

The granularity-problem is proven by the experimental evidence that humans, even professionals not only novices, diverge in opinions when asked to assign senses to words in contexts. In people's minds, multiple senses of a polysemous word are not strictly divided, but overlap. Dictionaries are artificial in pretending that one sense ends where another sense begins. This is true for homonyms, but not polysems. But this is a general problem of any dictionary, and WN is not more polysemous than standard collegiate dictionaries are (the Oxford English Dictionary, for instance, is much more prodigious in this respect). In a study intended to compare different dictionaries, including PWN, Fellbaum et al. [6] show up significant differences with respect to entries of polysemous words. Editors and lexicographers' decisions influence the number of senses to be included and the granularity down to which certain senses are broken up. The verb sense clusters, which form up synsets, currently in PWN have been created largely on the basis of lexicographic intuitions. PWN misses an explicit relation between close senses. In PWN all senses of a word are equally distanced. Not even the distinction among homonyms is marked, and so much less among polysems. In [6], a method of enhancing sense clustering based on the analysis of inter-annotator disagreement is presented. But until an alternative approach for representing senses, as for instance an underspecified methodology ([16], [2], [3]), will be implemented into a dictionary, the best we can do is to show how closely senses are related. This is what is happening now in EWN with the sense groupings, where closely related senses are distinguished from less closely related senses [5].

The gap-problem induces the impossibility to always align meanings between two languages. When the alignment conceptual backbone, in a consortium like EWN or BalkaNet, is a bag of concepts like ILI, a solution to the gap-problem would be to design ILI as the union of concepts contributed by all languages (correctly speaking, languages contribute with meanings, which are seen as concepts as soon as they are included in ILI). If that had been the case, then ILI, augmented also with the ontology (hyper/hypo relations), would have become a universal linguistic ontology[1]. In this case the meanings of one particular language would have been aligned sparsely (here and there), but exactly (there is always one concept in ILI with which the meaning of a synset of one language matches exactly), into this backbone ontology. Then finding the correspondences of synsets between pairs of languages would be straightforward: start on a synset S1 (as a collection of individual word senses) of a language L1, go through the mapping directly into ILI, find the closest point there which has a mapping link with the language L2 and report the corresponding synset S2 (again a collection of word senses) in L2.

Unfortunately, there are some tricky problems with this strategy and EWN, and consequently also BalkaNet, had to decline it: 1). badly differentiated senses in the ILI may lead to unintended mismatches across languages and 2). the status

[1] Linguistic ontology, and not only ontology, because only concepts that correspond to word meanings (therefore for which there exist words, in at least one language, that have the corresponding senses) are there.

of the ILI-concepts is not the same. Some concepts are more fundamental than others. Some languages have handy lexicalizations to refer to conceptualizations that are analytically expressed in other languages (as compounds that can systematically derive new concepts from existing concepts). For example, in Germanic languages there are verbal compounds that combine a manner verb with a resultive state[2]: *to open by pulling*, *to open by pushing*, *to open by kicking*, *to open by tearing*, *to open by turning*, etc. ..., *to close by pulling*, *to close by pushing*, etc. ..., *to clean by pulling*, *to clean by pushing*, etc. ... This would inflate the ILI to huge proportions. The discussion at the end of EWN was to either add these concepts but mark the status as being productively derivable from basic concepts, or to link the lexicalization in these languages to these concepts with multiple complex equivalence links (plus an indication that these links are exhaustive). Any of these solutions proved to require an adaptation of the database, which was not feasible close to the end of the project. Furthermore, it was discovered that most of the concepts that could not be linked to the ILI are of this nature and that the intersection across the languages in terms of lexicalizations that could not be linked would have been very low. Finally, the EWN consortium did not agree to transform ILI into a superset.

3. Lexical Resources

The main dictionary that we used for mapping the PWN onto the Romanian WN is *The Romanian Explanatory Dictionary* (EXPD) [4], second edition – one of the lexicographic references for contemporary Romanian, realized by "Iorgu Iordan" Institute of Linguistics of the Romanian Academy. With about 56 000 entries, including homographs, the second edition is republished in 1998. Words in EXPD can have major senses, related senses, and senses that depend on a major sense. When applicable, idioms, phrases and/or collocations reveal special usages of the title word in combinations with other words. Morphological information, examples, etymological references, indications on popular, regional and archaic forms and the domain of applicability are also included.

The electronic format of the EXPD in the XML standard (XML-EXPD) started to be developed in the context of the CONCEDE European project and the WEB-LEX project of the Romanian Academy and evolved gradually from an initial variant with 23 000 word entries to the present variant with more than 63 000 entries. Details of the Document Type Definition (DTD) of the XML-EXPD are given in [17] and [19].

As explained in [19], the development of the ROWN was based on different lexicographic sources and followed a mixed expand-merge model. A source of candidates for the Romanian synsets was the second edition (1999) of *The Romanian Dictionary of Synonyms* (SYND) [21]. The dictionary observes two major principles that give the structure of each synonymic series. At the base level a synonymic series lists word senses that are semantically equivalent in a given context – the semantic hierarchy criterion. The functional values inside a series are recorded at the references level in order to indicate the stylistic, diachronic and regional usage of the synonyms.

[2]We thank Piek Vossen to provide these examples in a dialogue with one of the authors of this paper.

More than 35 000 words were grouped in about 22 000 synonymic series. SYND was transposed electronically and formatted in XML (the resource is known as XML-SYND). Because ROWN was intended to offer a lexical coverage of general use in Contemporary Romanian, the archaisms and regionalisms from SYND were removed.

Other resources used were: an English-Romanian dictionary (ERD), also brought into the XML format (XML-ERD) and a collection of translation equivalents extracted automatically from an English-Romanian parallel corpora (XML-ERTE) [20]. All these resources were integrated in an interface, which is largely explained in [19].

4. Overtaking Mapping Difficulties

In this section we will present difficult cases of sense mapping. Among the main causes that made mapping hard in these cases were: the difficulty to perceive correctly the English senses by a non-native speaker of English, the too fine-grained sense grid of the PWN compared to the Romanian resources, the different principles underlying PWN and the Romanian linguistic sources and the inherent differences between the two languages.

Glosses are far from being perfect in many cases in PWN, as reported by some authors [9]. A constant cause of trouble, in this respect, is the occurrence of the entry word in the gloss (circular or recursive definitions), a situation that would be harmless only if the word belongs also in a hypernym synset, where it would have a more general sense, and the definition would observe the *genus proximus* paradigm.

The difference of granularity between PWN and EXPD caused many problems. The literal "break" has 27 senses in PWN and it has also three derived expressions: *break_into_fragments, break_into_parts* and *break_up* (the last one having itself 18 different senses). In some cases we refined the meanings from EXPD and in others we added other meanings. There where we had to add new senses we also translated the PWN glosses.

As explained in [19], in this volume, when a sense is perceived as missing in EXPD, our linguists were supposed to assign x to a literal, as the sense number. If, among the senses already available, the human annotator can find one close enough to what he intends to express but which does not have exactly the same meaning, he can assign that number of sense followed by a $.x$, for instance $2.x$. In a later stage, a procedure will process all $.x$ markings and generate automatically sense numbers, simultaneously with the automatic generation of a new EXPD, which will display a much finer grained sense distinction.

In the following we will try a classification of problems encountered during the PWN to ROWN mapping process. As matters of notation, in order to differentiate between hypernyms and hyponyms/troponyms we will use the sign => to prefix hypernyms and <= to prefix hyponyms/troponyms, when illustrated in line with the text, we will place synsets between square brackets [...], sometimes in an abridged form, and we will differentiate between English and Romanian words in our examples by placing English in Courier and Romanian in *Italics*.

Example 1.

Difficulty to map senses due to: similar glosses, no examples of usage and unique hypernym.

Solution: mapping based on Romanian cognates and similarity between a noun and the verb it is derived from.

Consider the following synsets that display very close senses:

```
fan(1) -- (a device for creating a current of air by movement of a
   surface or surfaces)
ventilator(1) -- (a device (such as a fan) that introduces fresh air
   or expels foul air)
blower(1) -- (a device that produces a current of air)
```

Although all three synsets above do not have words in common, the sense distinction is hard to perceive because of the similarity in glosses (all three are described to be devices that produce a current of air) and their unique direct hypernym:

```
device -- (an instrumentality invented for a particular purpose;
  "the device is small enough to wear on your wrist";
  "a device intended to conserve water").
```

Romanian includes more words for devices that produce airflows in different settings: *uscător*, *uscător_de_păr*, *foen* (or *foehn*), *suflantă*, *ventilator*, *suflător*. However, *uscător*, *uscător_de_păr*, *foen* (or *foehn*) is the synset mapped onto the only hyponym synset for `blower(1)`:

```
hand blower, blow dryer, blow drier, hair dryer, hair drier --
   (a hand-held electric blower that can blow warm air onto the hair;
   used for styling hair)
```

It follows that we can eliminate from this debate *uscător_de_păr*, *foen* (or *foehn*). Although ambiguous between `hair dryer` and `dryer`, *uscător* has the meaning of instrument that circulates air in order to dry some materials and should also be eliminated. We remain to decide among *suflantă*, *ventilator*, and *suflător*. Among them, *ventilator(1)*, used to cool air in a room, should be mapped onto `ventilator(1)`, and `blower(1)`, being derived from the verb `to blow` (*a sufla*) should mainly be like a *suflantă(1)* or a *suflător(x)* and can be mapped accordingly. It remains `fan(1)`, which could be mapped, inconvincibly, onto an invented sense for *ventilator*: *ventilator(x)*.

Example 2.

Difficulty due to: intersecting synsets and lack of hypernym (top-most PWN synsets).

Solution: mapping based on judging hypernyms (where existent) and glosses.

```
commemorate(1), mark(4) -- (mark by some ceremony or observation;
   ‘‘We marked the anniversary of his death’’)
   => observe, celebrate, keep -- (celebrate, as of holidays or rites;
```

```
   "Keep the commandments"; "celebrate Christmas";
   "Observe Yom Kippur")
commemorate(2), remember(8) -- (call to remembrance; keep alive
   the memory of someone or something, as in a ceremony;
   ‘‘We remembered the 50th anniversary of the liberation of
   Auschwitz’’; ‘‘Remember the dead of the First World War’’)
commemorate(3), memorialize(2), memorialise(2), immortalize(1),
   immortalise(1), record(5) -- (be or provide a memorial to a
   person or an event; ‘‘This sculpture commemorates the victims
   of the concentration camps’’; ‘‘We memorialized the Dead’’)
   => remind -- (put in the mind of someone;
   "Remind me to call Mother")
```

Identification of different hypernyms is an important clue in distinguishing among very close meanings. In the above example, in which all three synsets are semantically very close, hypernyms can differentiate only two of the three. We find that `[commemorate(1), mark(4)]` is one way of `[observe, celebrate, keep]`, while `[commemorate(3), memorialize(2), memorialise(2), immortalize(1), immortalise(1), record(5)]` is one way of `[remind]`. This distinction allows the mapping of the first synset above onto the Romanian equivalents: *comemora(1)*, *celebra(2)*, as in a ceremony (*Am comemorat terminarea războiului – We have commemorated the end of the war*), and, respectively, the third synset above onto *comemora(x)*, *marca(2.x)*, which expresses the reason of a monument erected in the memory of something or somebody, as in *Stela comemorează (marchează) căderea zidului* – `The stele immortalize the fall of the Berlin wall`. As for the second synset, its meaning cannot be distinguished based on a hypernym, since it has none, but the gloss and the inclusion of `remember` in its synset helps to map it with *comemora(x)*, *aduce_aminte(x)*, *rememora(x)*, an act of remembering with distinguished feelings (as in *Am comemorat (ne-am adus aminte de, am rememorat) victimele atentatului de la World Trade Center* – `We have commemorated/remembered the victims of the World Trade Center attempt`).

Example 3.

Difficulty due to: identical synsets and hypernyms, glosses differing by just one word, poorly differentiating examples and a common hyponym/troponym.

Solution: mapping judged on slight differences in glosses and examples, and on the verbs subcategorisation frames.

The following synsets are identical in literals, have almost identical glosses and very similar examples:

```
beat(8), flap(3) -- (move with a thrashing motion; ‘‘The bird flapped
   its wings’’; ‘‘The eagle beat its wings and soared high into the
   sky’’)
beat(15), flap(4) -- (move with a flapping motion; ‘‘The bird’s wings
   were flapping’’)
```

The confusion is accentuated also by a common troponym: `[clap] -- (strike the air in flight; of the wings of certain birds)`. However, the PWN sentence frames reveal that the verbs of the first synset are transitive while the second are intransitive. If this proves correct, than the Romanian synsets should be:

bate(x), lovi(x) for `[beat(8), flap(3)]`
(se) bate(x), se lovi(x), fâlfâi(x) for `[beat(15), flap(4)]`

So, one can say: *Aripile pasării băteau (loveau) cu putere aerul*, as in the first example, or *Vântul bate frunza-n dungă* (in a free translation: `The wind beats the leaves`), and *Pânza bate (fâlfâie) în vânt. Pânza se bate (se loveşte) de catarg* (`The foils beat(15)/flap(4) in the wind.`; still: `The foils are beating(8)/flapping(3) against the mast`, i.e. a natural translation obliges to a sense shift.).

Example 4.
Difficulty due to: identical synsets, almost identical glosses.
Solution: mapping due to transitivity and discretisation of EXPD sub-senses.

```
paint(1) -- (make a painting; "he painted all day in the garden"; "He
  painted a painting of the garden")
  => create -- (pursue a creative activity; be engaged in a creative
  activity; "Don't disturb him -- he is creating")
   => act, move -- (perform an action, or work out or perform (an
   action); "think before you act"; "We must move quickly"; "The
   governor should act on the new energy bill"; "The nanny acted
   quickly by grabbing the toddler and covering him with a wet towel")
paint(3) -- (make a painting of; "He painted his mistress many times")
  => represent,interpret -- (create an image or likeness of, in art)
   => re-create -- (create anew; "Re-create the boom of the West
   on a small scale")
    => make, create -- (make or cause to be or to become; "make a
    mess in one's office"; "create a furor")
```

Among the four one-literal synsets of `paint`, senses 1 and 3 are harder to differentiate. The distinct hypernyms do not help much. However, the difference of one preposition in the glosses and the sentence frames leave us with the conclusion that `paint(1)` is intransitive, while `paint(3)` is transitive, the direct object expressing the object painted. Although the second example of the first synset contradicts this supposition, the corresponding hypernyms do strengthen it. As in English, Romanian does not have distinct lexicalization for these two senses, but EXPD considers one as a derived sense of the other: *picta(1.1)*, as for being gifted with the tallent to paint, and *picta(1)*, as for executing a painting. They were aligned, respectively, to the two synsets above.

Example 5.
Difficulty due to: recursive glosses, same word in synsets related hypernymically.
Solution: one synset mapped, one synset left open.

```
fall(14) -- (to be given by assignment or distribution; "The most
  difficult task fell on the youngest member of the team"; "The onus
  fell on us"; "The pressure to succeed fell on the youngest student")

fall(20), light(5) -- (fall to somebody by assignment or lot; "The
  task fell to me"; "It fell to me to notify the parents of the
  victims")
  => fall(21), return, pass, devolve -- (be inherited by; "The estate
  fell to my sister"; "The land returned to the family"; "The estate
  devolved to an heir that everybody had assumed to be dead")
```

Not only the presence of the same word (`fall`) in the synset and in the gloss (this is what we call a recursive gloss), but the same literal (`assignment`) appears, apparently as an important clue, in both glosses. To make the task of distinguishing meanings even harder, `[fall(14)]` has neither hypernyms nor troponyms (it is a case of an isolated synset), while for `fall(20)` the hypernym contains again the same word, `fall`. In Romanian, *cădea*, *cădea_pe_umerii_cuiva*, *reveni* can all be used in utterances that are translations of both kinds of the exhibited examples, so the decision was to align `[fall(20), light(5)]` onto this synset and to leave the other synset unmapped.

Example 6.

Difficulty due to: closely related synsets, very similar examples, no differentiating hypernyms (case A), no differentiating troponyms.

Solution: mappings based on different hypernyms (case B) and transitivity; expanding applied (case A).

Kilgariff and Yallop [11], based on observations acquired on a small corpus, have found that it is invalid to assume that lexicographically close senses are also hierarchically close. High-level categories in the hierarchy often relate to different facets of the same object or event.

One example of this kind is the verb `snap` with three of the senses as follows:

```
snap(6), crack -- (make a sharp sound; "his fingers snapped")
snap(7) -- (move with a snapping sound; "bullets snapped past us")
snap(10), click, flick -- (cause to make a snapping sound; "snap your
   fingers")
```

The sense `snap(6)` falls under:

```
sound, go -- (make a certain noise or sound; "She went 'Mmmmm'"; "The
   gun went 'bang'")
```

while the senses `snap(7)` and `snap(10)` have the same hypernym:

```
move -- (move so as to change position, perform a nontranslational
   motion; "He moved his hand slightly to the right").
```

If a non-native speaker understands quite easily the distinction between senses 6, on one hand, and 7 and 10, on the other, due to the sharp separation induced by their hypernyms, `[sound, go]` and `[move]`, respectively, it is much more difficult to distinguish between senses 7 and 10. Moreover, none has troponyms. The gloss of `snap(10)`, confirmed also by the sentence frames, however, distinguishes this last synset as the causative (transitive) variant of `snap(7)`. The Romanian synset *plesni(8.2)*, *pocni(x)* equally maps all principal meanings, so the solution is, simply, to expand by translation the English synsets.

Example 7.

Difficulty due to: dangerously close glosses, misleading hypernyms, intersecting hypernymic paths.

Mapping based on: good differentiating examples.

```
drink(2), booze(1), fuddle(2) -- (consume alcohol; ‘‘We were up
   drinking all night.’’)
   => use, habituate -- (take or consume (regularly or habitually);
   "She uses drugs rarely")
    => consume, ingest, take in, take, have -- (serve oneself to, or
    consume regularly; "Have another bowl of chicken soup!"; "I don’t
    take sugar in my coffee")
drink(5), tope(1) -- (drink excessive amounts of alcohol; be an
   alcoholic; ‘‘The husband drinks and beats his wife’’)
   => consume, ingest, take in, take, have -- (serve oneself to, or
   consume regularly; "Have another bowl of chicken soup!"; "I don’t
   take sugar in my coffee")
```

The path from these two synsets to the common synset `[consume, ingest, take in, take, have]` includes two hypernym links from `[drink(2), booze(1), fuddle(2)]` and one from `[drink(5), tope(1)]`. The first synset reveals to be a troponym of `[use, habituate]`, from which it inherits the meaning of regularly/habitually drinking. Still the glosses and the attached examples seem to give contradictory details. Based on them, one can state that if one drinks excessively and regularly, then he `drinks(5)` while `drink(2)` is more like an occasional drinking, although not less excessive. Our lexicographers decided to give more credit to the glosses/examples then to the hierarchy. Hence, appropriate examples that better differentiate between the two meanings are: `Don’t marry him because he drinks.` for `drink(5)`, and `Even if I drank all night, in the morning I was all right.` for `drink(2)`. The corresponding Romanian synsets are: *bea(x), fi_beţiv(1)*: *Nu l-a luat de bărbat pentru că bea.* and *bea(3)*: *A consuma băuturi alcoolice.*

Example 8.

Difficulty due to: practically identical glosses.

Mapping based on: differentiating examples, different hypernyms.

```
play(25), play_on(2) -- (perform music on (a musical instrument);
   "He plays the flute"; "Can you play on this old recorder?")
```

```
   => sound -- (cause to sound; "sound the bell"; "sound a certain
   note")
play(34) -- (play (music) on an instrument; "The band played
   all night long")
   => perform -- (give a performance (of something); "Horowitz is
   performing at Carnegie Hall tonight"; "We performed a popular
   Gilbert and Sullivan opera")
```

Since glosses are helpless in this case, the slight difference evidenced by examples should have been confirmed by the hypernymic paths. Our intuition was to map `play(25)`, `play_on(2)` on *a cânta(x)*, with the meaning of possessing the skill to play an instrument, as in *Enescu cânta la vioară de la 4 ani.* (`Enescu played the violin since he was four years old`), and to map `play(34)` to *a cânta(x), a interpreta(x), a executa(1)*, with the meaning of performing music at a certain moment. Both interpretations are confirmed by the immediate hypernyms. To do the mapping we had to notate sub-senses in order to account for the absence of these senses in EXPD.

Gildea [7], referring to the same verb `to play` in WN, notices the complex interaction between verb senses and alternation criteria.

Example 9.

Difficulty due to: scarceness of glosses, irrelevant examples, unique hypernym.

Solution: mapping based on information acquired from other sources.

```
braise(1) -- (cook in liquid; ‘‘braise beef’’)
   => cook -- (transform and make suitable for consumption by heating;
   "These potatoes have to cook for 20 minutes")
boil(2) -- (cook in boiling liquid; ‘‘boil potatoes’’)
   => cook -- (transform and make suitable for consumption by heating;
   "These potatoes have to cook for 20 minutes")
```

Longman [12] defines `braise` as `to cook meat or vegetable slowly in a small amount of liquid in a closed container`, a meaning which is not lexicalized in Romanian in a single word, but is usually used as an expression: *fierbe_înăbuşit(x)*. For `boil(2)`, the meaning is clear: *fierbe(2)*.

Example 10.

Difficulty due to: recursive glosses, partly identical hypernyms, non-differentiating hyponyms/troponyms.

Solution: distinction failed.

```
bulge(1), pouch, protrude -- (swell or protrude outwards; "His eyes
   bulged with surprise")
   => change shape, change form, deform -- (assume a different shape
   or form)
   <= bulk -- (stick out or up; "The parcel bulked in the sack")
bulge(2), bag -- (bulge out; form a bulge outward, or be so full as
   to appear to bulge).
```

```
   => stick out, protrude, jut out, jut, project -- (extend out or
   project in space; "His sharp nose jutted out"; "A single rock
   sticks out from the cliff")
   <= protuberate -- (form a rounded prominence; "The starved child's
   belly protuberated")
protrude, pop, pop out, bulge(3), bulge out, bug out, come out --
   (bulge outward; "His eyes popped")
   => change shape, change form, deform -- (assume a different shape
   or form)
```

Again we face recursive glosses. Then, `bulge(1)` and `bulge(3)` are placed under the same hypernym – `[change]`, while `bulge(2)` is placed under `[stick out]`. The immediate hypothesis is that one hierarchy has to deal with a (volitive) active act, while the other with a passive act, and that the active meaning should be the ones under `[change]`. Another possibility to make a distinction[3] is that between a permanent or stative interpretation (as a nose or a rock) and a temporary one (as when a soccer ball in a bag makes the bag bulge out). Still, a non-English-native reader could hardly exploit the information in these PWN records in order to distinguish between the meanings `bulge(1)` and `bulge(3)`, since the examples are much confusing either: `"His eyes bulged with surprise"` versus `"His eyes popped/bulged"`. The most closed Romanian equivalents are: *(se) bulbuca(1.1)*, *face_ochii_mari(x)* (when related to eyes), *(se) umfla(x)*, *ieşi_în_relief(x)*, but the mapping decision is left open.

Example 11.
Difficulty due to: too large a synonymic Romanian series.
Solution: mapping based on the splitting of initial synonymic series into sub-series.

The WN synset:

`feud(1) -- (a bitter quarrel between two parties)`

was initially mapped onto the SYND-based synonymic suite:

animozitate(1.1), ceartă(2.1), conflict(1.1), dezacord(1), dezbinare(1.2), diferend(1), discordie(1), discuţie(2.2), disensiune(1), divergenţă(1), gâlceavă(1.2), învrăjbire(1.2), neînţelegere(2.1), vrajbă(1.1), zâzanie(1.1)

which is more a semantic cluster then a synset displaying one meaning. Eventually, we had to dramatically decrease the number of literals from many synsets in order to perform an accurate mapping. Thus, *feud(1)* was finally mapped onto a synset with a single literal – *dihonie(1)*. The following classification shows how diversely this cluster disperses in the PWN hierarchy:

[*animozitate(1.1)*]⇔`[animosity(1), animus(1), bad blood(1)] - a feeling of ill will arousing active hostility ...=>[psychological feature]`

[*dezacord(1), diferend(1.x), discordie(1), divergenţă(1.1)*]⇔ `[discrepancy(1),`

[3]Suggested by Christiane Fellbaum.

`disagreement(2), divergence(4), variance(4)] -- a difference between conflicting facts or claims or opinions ...=>[abstraction]`

[*divergenţă(1), fricţiune(2.1)*]⇔`[clash(2), friction(1)] -- a state of conflict between persons=>[conflict(4)]=>[state]`

[*conflict(1.1)*]⇔`[conflict(4)] -- a state of opposition between persons or ideas or interests=>[state]`

[*dezacord(1.x), diferend(1), disensiune(1), divergenţă(1.x), neînţelegere(2.1)*]⇔ `[discord(4), discordance(2)] -- strife resulting from a lack of agreement=>...=>[conflict(1)] ...=>[act]`

[*dihonie(1)*]⇔`feud(1) -- a bitter quarell between two parties =>[conflict(1)] ...=>[act]`

[*conflict(2), controversă(1.1), discuţie(2.2), dispută(1.1), neînţelegere(2.2)*]⇔ `[dispute(1), difference(3), difference of opinion(1), conflict(7)] -- a disagreement or argument about something important...=>[act]`

[*dezbinare(1.2), sciziune(1)*]⇔`[disunion (1)] -- the termination or destruction of union...=>act`

The rest: *ceartă(2.1)*, *gâlceavă(1.2)*, *învrăjbire(1.2), vrajbă(1.1)* and *zâzanie(1.1)* are yet unmapped.

A similar example is for:

`design(7) -- (intend or have as a purpose; "She designed to go far in the world of business")`

which was mapped initially onto:

afla(5), concepe(1), crea(1.1), elabora(1.1), gândi(3), imagina(1.2), inventa(1.1), născoci(1.1), plăsmui(2), proiecta(1.1), realiza(2.1), scorni(1.2)

and eventually onto *plănui(1.2)*. Since SYND authors understand synonymy in a different way than the PWN authors, we eventually had to dramatically decrease the number of literals from many synsets in order to perform an accurate mapping.

To realise the dimension of the decrease in number of literals per synset we made a statistic starting in August 2003. At that time, the number of synsets was 14,407 and the average number of literals per synset was 1.922. At a later stage, with 14,699 synsets, the average number of literals per synset decreased to 1.814. In this moment, with ROWN comprising 15,763 synsets, the number of literals per synset has fallen to 1.783. If we take the PWN as a standard, with its 1.760 literals per synset, then our wordnet is approaching this average from above.

Example 12.
Difficulty due to: lexical gaps.
Solution: mapping to an empty synset.

`bevy(2) -- (a flock of quail)`

We have a word for `flock`: *stol*, but we don't have a word for a specific kind of flock, like a flock formed by quail. On the other hand, to map `bevy(2)` with *stol_de_prepeliţe* would violate the principle based on which expressions are accepted in a dictionary (either because their meaning cannot be inferred from the meanings of the component words – the case of idiomatic expressions, or because their frequency is much higher than others formed under the same composition rules).

A similar thing happens with the synset:

```
cork(1), cork up -- (close a bottle with a cork)
   => plug, stop up, secure -- (fill or close tightly with or as if
   with a plug; "plug the hole"; "stop up the leak")
```

We have a word for `plug(2)` which is *astupa(x)*, *înfunda(x)*, *pune_dop(x)* but to cork is to plug something (as a bottle) with a specific kind of plug, a cork (*un dop de plută*), a concept which is not lexicalized in Romanian.

The consortium voted for the non-lexicalised synsets, in the Balkanet progress meeting in Bucharest (August 2003). Before this date the PWN concepts that are non-lexicalised in Romanian have been mapped onto Romanian synsets that were considered their closest approximations.

Example 13.

Difficulty due to: Romanian synonymy series does not match all examples associated to the PWN synset.

Recommended solution: mapping to a synset having a Romanian-specific hyponym.

The problem signalled here seems to be due to a coarser granularity of a PWN synset compared with the possible Romanian equivalents. Consider the synset:

```
violate(3) -- (destroy; "Don't violate my garden"; "violate
   my privacy")
```

The Romanian closest synonymy series is *viola(3), încălca(x), călca(7.x), profana(x), pângări(1.2), păta(3.1)*. Still, there is a differentiation, because while *viola(3)* can be followed by both abstract and concrete objects (*proprietatea* – `the property`, *grădina* – `the garden`), *încălca(x)* can only accept an abstract noun as an object (*intimitatea* – `the privacy`, *proprietatea* – `the property`, *drepturile* – `the rights`) and *călca(7.x), profana(x)* accept only concrete objects (*grădina* – `the garden`, *mormîntul* – `the grave`). So, **a încălca grădina* and **a călca intimitatea* are both defect. Certainly, a possible solution would be to disregard these selectional restrictions and to put up a synset as the one suggested, but a better alternative would be to map the PWN synset `[violate(3)]` onto the ROWN synset [*viola(3)*], to which to attach two troponyms, the synsets: [*încălca(x)*] and [*călca(7.x), profana(x)*].

Example 14.

Difficulty due to: systematic sense unifying (condensing) glosses of a certain class of nouns in EXPD.

Solution: refinement (splitting) of senses.

Very often EXPD defines activity nouns such as `the action/activity of ...and its result` (*acţiunea/activitatea de a ... şi rezultatul ei*). Examples are: *amestecare(1), amendare(1), brăzduire(1), chenăruire(1), chemare(1), cumpănire(1), dispreţuire(1), dispersare(1), lepădare(1), limitare(1)* etc. The list can continue as almost every printed page in EXPD contains such a definition for a noun.

In these cases the solution was to split the Romanian senses by the procedure described at the beginning of this section and in [19], in this volume.

Example 15.

Difficulty due to: EXPD considers figurative or derived senses attached to a principal sense, there where PWN indicate different senses.

Solution: mapping by discretisation of EXPD sub-senses.

The verb `devour` has four senses in PWN: one related to destruction, one to pleasure and two to the act of eating:

```
devour(1) -- (destroy completely; "Fire had devoured our home")
   => destroy, ruin -- (destroy completely; damage irreparably;
   "You have ruined my car by pouring sugar in the tank!";
   "The tears ruined her make-up")

devour(2) -- (enjoy avidly; "She devoured his novels")
   => enjoy, bask, relish, savor, savour -- (derive or receive
   pleasure from; get enjoyment from; take pleasure in;
   "She relished her fame and basked in her glory")

devour(3), down, consume, go through -- (eat immoderately;
   "Some people can down a pound of meat in the course of one meal")
   => eat -- (take in solid food; "She was eating a banana";
   "What did you eat for dinner last night?")

devour(4), guttle, raven, pig -- (eat greedily; "he devoured
   three sandwiches")
   => eat -- (take in solid food; "She was eating a banana";
   "What did you eat for dinner last night?")
```

The most closely related Romanian equivalent is the cognate verb *devora*, (both English and Romanian, derived from the same French *dévorer*, and Latin *devorare*) which is defined in EXPD with one principal sense – related to the act of eating: *mânca_cu_lăcomie, înghiţi_pe_nemestecate*. EXPD records the sense related to destruction as a derived figurative one, with the synonymy class: *consuma, mistui, înghiţi*, and with a similar example to that of `devour(1)` *Focul a mistuit casa* (`Fire has devoured the house`). This sense is further refined in a phrasal construction which exemplifies the pleasure sense: *a devora o carte* (`to devour a book`). Hence the explicit refinement of senses in EXPD is to be aligned to the discrete independent range of senses of PWN (similar to Example 4).

Example 16.

Difficulty due to: missing of principal or derived senses in EXPD for words that gained figurative senses that became more frequent that the original sense.

Solution: notation of new senses.

```
manipulate(1), pull strings, pull wires -- (influence or control
  shrewdly or deviously; "He manipulated public opinion
  in his favor")
```

The Romanian equivalent, expressed by the verb *manipula* has in EXPD just one sense, which is that of `holding something in one's hand and move it`, which is the PWN sense 2 of the same verb. Moreover, as frequently happens, in contemporary Romanian (in English also, as indicated by the senses order) the figurative sense became more frequent than the original, recorded, sense.

Example 17.

Difficulty due to: meanings that in Romanian are not expressed by single literals, but by idiomatic expressions.

Solution: expressions introduced.

The synset:

```
reckon(6), count(8) -- (take account of; ‘‘You have to reckon with our
   opponents; ‘‘Count on the monsoon’’)
```

does not have a one word equivalent in Romanian. However, the corresponding meaning is expressed by some compounds, although EXPD does not record them in neither of the verbs around which they are formed: *ţine(35.1.x), avea_în_vedere(1), lua_în_considerare(1), ţine_cont_de(1).*

Example 18.

Difficulty due to: obscene words/expressions omitted in EXPD.

Solution: equivalents introduced.

An old direction suggested by Iorgu Iordan, one of the most significant Romanian linguists, was against the inclusion of extremely obscene words in a normative dictionary. We do not have the nerve to contest this decision, but, since words in a dictionary are only words, they cannot harm people more than read in a novel or listen in everyday usage (if one would consider to eliminate all words that have the potential to harm in language use, we would probably remain with very few). To map concepts such as:

```
roll in the hay, love, make out, make love, sleep with, get laid,
   have sex, know, do it, be intimate, have intercourse, have it away,
   have it off, screw, fuck, jazz, eff, hump, lie with, bed, have a go
   at it, bang, get it on, bonk -- (have sexual intercourse with;
   "This student sleeps with everyone in her dorm"; "Adam knew Eve";
   "Were you ever intimate with this man?")
```

or

`cock, prick, dick, shaft, peter, tool - obscene terms for penis`

we have added in ROWN some equivalent slang words.

Example 19.
Difficulty due to: artificial meanings of expression-type entries in PWN.
Solution: translation of expressions.

Included, probably, for reasons of separating certain closely related classes of troponyms, sometimes PWN considers, rather artificially, expression-type entries that are neither common use lexicalizations, nor idiomatic expressions. For instance, it includes entries for:

`change intensity -- (increase or decrease in intensity)`

to group under it, as troponyms, seven related meanings, or:

`change taste -- (alter the flavor of)`

to group three related meanings, but do not include entries for `change location, change age, change colour`, etc., of the related 393 synsets that are troponyms of

```
change, alter, modify -- (cause to change; make different; cause a
   transformation; "The advent of the automobile may have altered the
   growth pattern of the city"; "The discussion has changed my
   thinking about the issue")
```

In these cases the adopted solution was to translate the expression: *a(-şi) schimba, modifica intensitatea* for `change intensity`, or *a(-şi) shimba gustul* for `change taste`, although, perhaps a solution more in accordance with the principles of wordnets would have been to map them to unlexicalized entries (lexical gaps).

5. Discussion and Conclusions

In this paper we have reviewed the main problems encountered during the alignment of PWN synsets onto ROWN synsets from the perspective of a wordnet developer who is a non-native speaker of English. Nineteen different types of difficulties have been inventoried and their corresponding solutions have been indicated.

Although all problems have been discussed in correlation with these two languages, English and Romanian, and their corresponding wordnets, one completed and one under construction, we believe that some of the causes analysed are general and could be faced by the great majority of wordnet builders. The choice of one or another of the presented solutions depend, of course, on the particular types of linguistic sources employed, on the language peculiarities, but, to a large extend, are also subjective, as they could be influenced by the linguist's personal "touch" of the language. However, knowing these types of solutions in advance could speed-up and enhance the results of this extremely elaborate and time consuming process. We hope that our experience could be useful to new wordnet projects. As we were not prepared for these problems, we had to invent ad-hoc solutions each time we faced them. Moreover, the activity in

a project of this extent is always carried out by large teams of linguists and computer scientists, and their solutions to similar issues could be very personal if they are not warned beforehand how to proceed. It is our belief that if we had known from the very beginning the types of problems possible to arise, the elaboration process of this large resource for the Romanian language would have been smoother, more efficient, and perhaps also better.

To synthesize, the problems that we faced can be classified in three large groups. In the synthesis that follows, in order to abstract away from the pair of languages English-Romanian, we will call English and PWN the source language and SouWN, respectively, while Romanian and ROWN, the target language and TarWN, respectively.

1. Problems caused by the difficulty to differentiate among different meanings in SouWN.

   These problems have as origin the following causes, all originating in the SouWN:

   - identical (ex. 3, 4) or intersecting synsets (all the others);
   - similar or non sufficiently differentiating glosses (ex. 1, 3, 4, 5, 8, 9);
   - recursive glosses (ex. 5, 10);
   - non-differentiating examples (ex. 1, 3, 5, 6, 9);
   - lack of differentiating hypernyms (ex. 2, 5) or identical hypernyms (ex. 3, 6, 9, 10);
   - lack of differentiating or confusing hyponyms/troponyms (ex. 5, 6, 10);
   - lexical intersection between synsets and their hypernyms (ex. 5);
   - contradiction between the information contributed by the hypernyms and that contributed by the examples (ex. 7).

   In cases when a differentiation was possible for the target language, the following sources of information contributed to the decision:

   - SouWN differentiating glosses (ex. 2)
   - SouWN differentiating hypernyms (ex. 2, 6, 8);
   - cognates in source and target language (ex. 1);
   - similarity between a noun and the verb it is derived from in both source and target language (ex. 1);
   - matching of target language words in the translation of examples of SouWN (ex. 3, 7, 8);
   - verb transitivity (ex. 4, 6);
   - other lexical sources of the target language (ex. 9);
   - the SouWN synset, used for mere translation (expand) (ex. 6).

2. Problems caused by the different principles underlying the SouWN and the target language lexical sources.

   These problems had as origin the following causes:

   - all word senses in SouWN are equal, while the target language lexical sources have main and derived senses, which can go more and more specialised (ex. 4, 15);
   - the target language sources for synonym series display a much coarser granularity than the SouWN synsets (ex. 11);
   - systematic sense clashing in the target language linguistic sources (e.g. an activity and its result) (ex. 14);
   - new figurative senses not recorded in the target language linguistic sources (ex. 16);
   - idiomatic expressions missing in the target language linguistic sources (ex. 17);
   - extremely obscene words and expressions ignored in the target language linguistic sources (ex. 18);
   - artificial expressions in SouWN (ex. 19).

   Solutions considered:

   - discretisation of the sub-senses given by the target language linguistic sources (ex. 4, 15);
   - refinement of existing senses given by the target language linguistic sources (ex. 14);
   - inclusion in TarWN of new senses in addition to those given by the target language linguistic sources (ex. 16);
   - splitting of the synonymy classes proposed by the target language linguistic sources into SouWN-like synsets (ex. 11);
   - inclusion in TarWN of new idiomatic expressions in addition to those found in the target language linguistic sources (ex. 17);
   - inclusion in TarWN of equivalents of obscene words/expressions (ex. 18);
   - translation of expressions, entries of SouWN (ex. 19).

3. Problems caused by the intrinsic differences between the source language and the target language.

   These problems had as origin the following causes:

   - source language meanings are missing in the target language (ex. 12);
   - target language meanings are missing in the source language (ex. 13);
   - source language one-word meanings have a compound equivalent in the target language (ex. 17)

Solutions considered:

- SouWN synsets mapped onto empty synsets (target language lexical gaps) (ex. 12);
- inclusion of unmapped synsets as hyponyms/troponyms of mapped synsets in the TarWN (ex. 13);
- inclusion in the TarWN of new idiomatic expressions (ex. 17).

A general look at the conclusions above shows that the first two groups of difficulties we have individualized are due to the nature of the available resources. Practically, only the third group of problems derives from differences in the nature of the considered languages being thus subject to interpretation. We believe that the solutions we have proposed in the latter case respect both the linguistic reality and the technical necessity, and should therefore be adopted in case of further wordnet elaborations.

Acknowledgement. The authors are grateful to our colleagues from the Institute of Artificial Intelligence of the Romanian Academy in Bucharest (RACAI, in BalkaNet), Dan Tufiş, Ana-Maria Barbu, Verginica Mititelu, Eduard Barbu and Luigi Bozianu, who, not only collected and mapped synsets as we did, but had also the task to resolve conflicts, a much specialized and time consuming task. Our gratitude goes also to Christiane Fellbaum and Alexandra Cornilescu who gave us extremely useful feedbacks to a first draft of this paper.

References

[1] BENTIVOGLI, L., PIANTA, E., PIANESI, F., *Coping with lexical gaps when building aligned multilingual wordnets*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[2] BUITELAAR, P.P., *CoreLex: Systematic Polysemy and Underspecification*, Ph.D. Dissertation, Brandeis University, 1998.

[3] BUITELAAR, P.P., *Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification*, in *Proceedings of the ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*, Seattle, USA, 2000.

[4] * * *, *DEX – Dicţionarul Explicativ al Limbii Române* (in Romanian: The Explanatory Dictionary of Romanian Language), second edition, Univers Enciclopedic Publishing House, Bucharest, 1996.

[5] EDMONDS, P., *SENSEVAL, The evaluation of word sense disambiguation systems*, ELRA Newsletter, **7**(3), 2002.

[6] FELLBAUM, C., PALMER, M., DANG, H.T., DELFS, L., WOLF, S., *Manual and Automatic Semantic Annotation with WordNet*, invited paper, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[7] GILDEA, D., *Probabilistic Models of Verb-Argument Structure*, in *Proceedings of COLING-2002*, Taipei, Taiwan, 2002.

[8] GONZALO, J., CHUGUR, I., VERDEJO, F., *Sense Clusters for Information Retrieval: Evidence from SemCor and the EuroWordNet InterLingual Index*, in *Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality*, in conjunction with ACL-2000, Hong Kong, China, 2000.

[9] HARABAGIU, S., MILLER, G., MOLDOVAN, D., *WordNet 2 – a morphologically and semantically enhanced resource*, in *Proceedings of SIGLEX-99*, University of Maryland, 1999.

[10] KILGARIFF, A., *I don't believe in word senses*, research report ITRI-97-12, ITRI – University of Brighton, also published in Computers and Humanities, **31**(2), 1997.

[11] KILGARIFF, A., YALLOP, C., *What's in a thesaurus?*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[12] RUNDELL, M. et al. (eds.), *Longman Dictionary of Contemporary English*, Third edition, Summers, D., Harlow: Longman, 1995.

[13] MIHALCEA, R., MOLDOVAN, D.I., *Automatic Generation of a Coarse Grained WordNet*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[14] MORATO, J., MARZAL, M. Á., LLORÉNS, J., MOREIRO, J., *WordNet Applications*, in *Proceedings of GWC-2004*, Brno, Czech Republic, 2004.

[15] PALMER, M., DANG, H.T., FELLBAUM, C., *Making fine-grained and coarse-grained sense distinctions, both manually and automatically*, to appear in Natural Language Engineering.

[16] PUSTEJOVSKY, J., *The Semantics of Lexical Underspecification*, in Folia Linguistica, 1998.

[17] TUFIŞ, D., CRISTEA, D., *RO-BALKANET – ontologie lexicalizată, în context multilingv, pentru limba română* (in Romanian: RO-BALKANET – lexicalised ontology for Romanian in a multilingual context), in Dan Tufiş, Fl. Gh. Filip (Eds): *Limba Română în Societatea Informaţională – Societatea Cunoaşterii* (in Romanian: Romanian language in the Information Society – Knowledge Society), Expert Publishing House, Bucharest, 2002.

[18] TUFIŞ, D., CRISTEA, D., STAMOU, S., *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*, in this volume, 2004.

[19] TUFIŞ, D., BARBU, E., BARBU-MITITELU, V., ION, R., BOZIANU, L., *The Romanian Wordnet*, in this volume, 2004.

[20] TUFIŞ, D., BARBU, A.M., *Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries*, in Romanian Journal of Information Science and Technology, **4**, no. 3–4, 2001.

[21] SECHE, L., SECHE, M., *Dicţionarul de Sinonime al Limbii Române* (in Romanian: Dictionary of Synonyms of Romanian Language), second edition, Univers Enciclopedic Publishing House, Bucharest, 1999.