Romanian Semantic Role Resource

Diana Trandabăț[1,3] and Maria Husarciuc[1, 2]
[1]Faculty of Computer Science, "Al. I. Cuza" University of Iaşi, Romania
[2]Faculty of Letters, "Al. I. Cuza" University of Iaşi, Romania
[3]Institute for Computer Science, Romanian Academy
E-mail: dtrandabat@info.uaic.ro, mhusarciuc@gmail.com

Abstract

Semantic databases are a stable starting point in developing knowledge based systems. Since creating language resources demands many temporal, financial and human resources, a possible solution could be the import of a resource annotation from one language to another. This paper presents the creation of a semantic role database for Romanian, starting from the English FrameNet semantic resource. The intuition behind the importing program is that most of the frames defined in the English FN are likely to be valid cross-lingual, since semantic frames express conceptual structures, language independent at the deep structure level. The surface realization, the surface level, is realized according to each language syntactic constraints. In the paper we present the advantages of choosing to import the English FrameNet annotation, instead of annotating a new corpus. We also take into account the mismatches encountered in the validation process. The rules created to manage particular situations are used to improve the import program. We believe the information and argumentations in this paper could be of interest for those who wish develop FrameNet-like systems for other languages.

1. Introduction

Annotated language resources have became a must in natural language processing, especially for supervised learning (training and evaluation), unsupervised learning (evaluation), hand-crafted systems (evaluation), etc. Quality control is an important issue, since annotations, in order to be used as gold standard for evaluation, need to be very accurate. Interannotator agreement metrics have been developed (an overview is presented by Artstein & Poesio (2005)), but the major problems remain the temporal, financial and human resources needed in order to ensure a (near) perfect corpus. What if we have short deadlines and limited human and financial possibilities? A good solution would be to use existing language resources, built with considerable efforts for a specific language, and import them for a new language. In this paper we will militate for this idea by presenting the building of the Romanian semantic role resource through the import of the English FrameNet annotation.

Fillmore (1968) divides the language representation into Surface Structure (the syntactic knowledge) and Deep Structure (the semantic knowledge). The language process begins at the Deep Structure level with a non-verbal representation (an idea or a thought) and ends in the Surface Structure, as we express ourselves. The semantic roles (Case Notions) are representations at a semantic level of the lexical arguments. This inventory of cases includes universal concepts, possible innate, reusable in all languages, such as Agent, Instrument, Dative, Experiencer, Locative, Object, etc.

The paper is structured as follows: Section 2 gives a very brief overview of the English FrameNet resource, Section 3 presents the realization of the semantic structures database for the Romanian language, Section 4 presents the evaluation of the import method, and the last section discusses some possible applications and final conclusions.

2. The FrameNet project

FrameNet is a lexicographic research project which produced a lexicon containing very detailed information about the English predicational words (verbs, nouns and adjectives). The basic unit of analysis is the *semantic frame*, a "script-like structure of inferences, linked by linguistic convention to the meanings of the lexical units", defined as a type of event or state (Backer & al., 1998). A frame has a definition, a set of *frame elements FEs* (semantic roles) and a set of *lexical units LUs* which participate at its actualization. A lexical unit is a predicational word for which combinatory properties (the semantic frame) applies. The frame elements represent valences for a target predicational word and can thus be mandatory for the verb lexico-semantic realization, named *core frame elements*, or facultative, named *non-core frame elements*. Usually, the core FEs correspond to the direct arguments of a verb and assure the semantic correctness of the enunciation, while the non-core FEs represent the adjuncts, the modifiers of a verb, completing the enunciation with additional information. Figure 1 presents an example for the semantic frame *Commerce_sell*:

Frame elements (semantic roles):
Core FE: *Buyer*, *Seller*, *Goods*
Non - core FE: *Duration*, *Manner*, *Means*, *Money*, *Place*, *Purpose*, *Rate*, *Reason*, *Time*, *Unit*

Lexical units:
Verbs: *retail*, *sell*, *vend*
Nouns: *retailer*, *vendor*.

Example:
$[He]_{Seller}$ will probably $[sell]_{Target}$ $[her]_{Buyer}$ $[the\ book]_{Goods}$ $[for\ \$15]_{Money}$.

Figure 1: Example of Frame elements and lexical units for the target verb *sell*

FrameNet contains more than 10,000 lexical units, more than 6,100 of which are fully annotated, in approx. than 800 semantic frames, exemplified in more than 135,000 annotated sentences. In recent years, attempts to create lexical entries for languages other than English using the frame semantic approach have been undertaken. Japanese FrameNet, German FrameNet and Spanish FrameNet are currently under development. The Romanian FrameNet project belongs to the general tendency in the current research generated by the multilingual character of the knowledge society.

3. Romanian Semantic Role Resource: Building from Scratch or Importing?

The starting point for the German, Japanese and Spanish FN creation was the manual annotation at FE level of existing corpora for each language. This process is time and resource consuming. The approach we adopted was the direct import of the English annotation by translating the sentences in FrameNet. An overview of the two methods is presented in table 1.

Annotation of a new corpus
1. finding a corpus; 2. establishing an annotation schema (could be the same used in the English FrameNet project); 3. creating or deciding for an annotation software; 4. training at least two annotators (in order to be able to perform the interannotator agreement); 5. annotation process; 6. computing interannotator agreement and review of the mismatching cases.
Import of the annotation
1. translating FrameNet sentences; 2. aligning the English with the Romanian sentences; 3. running the import program; 4. validation of the data and review of the mismatching cases.

Table 1: Overview of the two methods for creating a semantic frames resource

The decision for the second method was based on a set of preliminary comparing tests. Thus, for the first method, we considered that we have the corpus, the schema, the software and two very well trained annotators, with good semantic frames knowledge, and that we only need to worry about the annotation process itself. Our tests revealed that a person can annotate an average of 30 medium sized sentences per hour. For a target of 100.000 sentences, we computed around 3500 hours, i.e. 20 months, considering 8 hours a day, 5 days a week working time. The main problem with this approach was the lack of a definite list of possible semantic roles. Therefore, different annotators can give different names (*agent* or *seller* or *vendor* for instance) to the same role, confusing the corpus quality metrics.

For the import method, the main time consuming task is the translation. However, a professional translator can translate up to 60 sentences an hour, even faster if translation memory is used. But the real gain is that the corpus can be split to several translators (cheaper and easier to find than semantic annotators), thus finishing the translation in about three months. After the automatic alignment and import, a single annotator is needed to perform the validation of the created corpus, focusing on cases where the alignment was not 1:1 (only 15% of the total number of sentences). Considering those calculations, the fact that we didn't had two annotators to work for 20 months just on semantic annotation, and the belief that once we have the import program, every other language could benefit from it and transfer annotations for its own language, we created a Romanian FrameNet based on the English annotation.

The intuition behind the importing program (presented in (Trandabăț, 2007)) is that most of the frames defined in the English FN are likely to be valid cross-lingual, since semantic frames express conceptual structures, language independent at the deep structure level. The surface realization, the surface level, is realized according to each language syntactic constraints.

The Romanian semantic role resource creation started with manual translation of 1094 sentences from the English FN (110 randomly selected sentences and the *Event* frame). This frame was selected due to its rich frame to frame relations, as presented below:

Relation	**With frame**	**LUs**	**Number of sentences**
Inheritance	Becoming	124	108
	Change_of_consistency	10	68
	Committing_crime	2	41
	Death	17	210
	Experience_bodily_harm	14	221
	Process_resume	1	0
	Process_start	10	14
	Process_stop	7	37
	Rotting	10	69
Subframe	Change_of_state_scenario	0	0
Using	Process_end 67	10	67

Table 2: Frame to frame relations within the *Event* frame

The next step was the automatic alignment at word level of the Romanian versions with the English ones, followed by the import of the English annotation, a manual validation aiming at detection the mismatching cases (Husarciuc & al., 2005) and an optimization process which, based on inference rules, corrects the automatic annotation.

In our approach, using the XML files of the English annotated sentences, we automatically created a set of XML files containing FE annotated sentences for the Romanian language. The import program (Fig. 2) uses as input files the annotations of the English lexical units and the aligner's output files.

The import algorithm is a simple one, focusing on:

- reading the English XML files and the alignment files;
- labelling each English word with the corresponding semantic role (FE) converting the character indexes into a word level annotation;
- mapping the English words with the aligned Romanian

correspondences;
- writing an output XML file containing the Romanian annotated corpus.

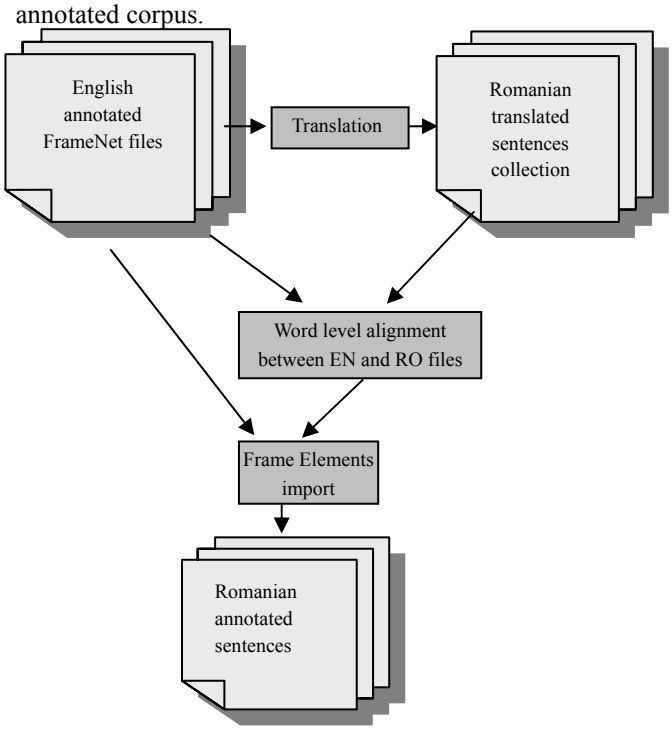

Figure 2: Creation of a quick parallel Romanian-English FrameNet subset

An example of import is presented in figure 3, where the first sentence is a Romanian translation of the second, English, sentence.

[Incidentul]$_{Event}$ A APĂRUT [după o dispută între individ şi personal]$_{time/cause}$ [la o filială a Băncii Irlandeze din Cahir]$_{Place}$.

[The incident]$_{Event}$ OCCURRED [after a dispute between the man and staff]$_{time/cause}$ [at a branch of the Bank of Ireland in Cahir]$_{Place}$

Figure 3: Example of imported sentence

4. Evaluation

The translations have been realized by professional translators, so the errors propagated in the corpus should be minimal. The reported problems related mainly to the lack of the context for English sentences, which generated different translation variants. However, if the English semantic frame is considered, this problem is surmountable. For the alignment process we used the aligner developed by the Romanian Academy Centre for Artificial Intelligence (Tufiş & al., 2005) with 87.17% reported precision and 70.25% recall. The aligner results were manually validated before entering the annotation import program.

The first results of the annotation import show an overall accuracy of 85%. The validation focuses on detecting the cases where the import has failed, trying to discover if the problems are due to the translation or to the semantic or syntactic specificities of Romanian that couldn't be captured in the aligner. Evaluating the correspondences between the semantic roles in the two languages, we have found that the automatic import of the double annotation in Romanian, realized on the basis of a simple inferring rule, is generally valid. However, there are some mismatches, whose causes, partially analyzed in (Husarciuc & al., 2005), are (1) the double annotation, (2) the existence of imbricate frame elements (FEs), (3) the unexpressed semantic frames, (4) the problems in the translation of the target-words, and (5) the lack of total correspondence between English and Romanian frames.

4.1. Double Annotation

In the English FrameNet, a FE is double annotated if and only if, due to semantic ambiguity, its role in the sentence cannot be precisely established. The double annotation applies only to the non-core frame elements, due to the fact that the same phrase can refer to multiple circumstances (peripheral roles) of an event. When a semantic element is double annotated in English, the same generally holds also for Romanian. The most frequent case of double annotation is for the *Time/Cause* roles, since almost any temporal specification involves a cause and/or a goal. In Figure 4, where the prepositional phrase is annotated both with a temporal and a causal semantic role, we have a succession of cause – effect relations: **Cause$_1$** (cutting the throat) => **Effect$_1$** = **Cause$_2$** (bleeding) => **Effect$_2$** (death).

Metodele tradiţionale cer ca [animalul]$_{Prot}$ [să sângereze]$_{Cause}$ până la MOARTE [după ce i s-a tăiat gâtul]$_{Time/Cause}$; o procedură agonizantă şi dezagreabilă durând între unul şi două minute .

Traditional methods require that [the animal]$_{Prot}$ [bleed]$_{Cause}$ to DEATH [after having its throat cut]$_{Time/Cause}$; an agonising and unsavoury procedure lasting between one and two minutes .

Figure 4: Example of double annotation

We represented a general lexicalized rule for the annotation of groups susceptible to express a double circumstance, temporal or causal (e.g. "after having its throat cut"), based on the lexical correspondence of several ambiguous prepositions, as in Figure 5:

```
# TC_en and TC_ro are lists of elements that can introduce
both temporal and causal phrases:
    TC_en = ("after", "when" ...);
    TC_ro = ("după ce", "când" ...) ;
    FOR each element x in Tc_en {
    IF (x has the correspondent y in Tc_ro)
        THEN {
# mark with FE Time/Cause the subordinate introduced by x
in English and by y in Romanian
        Time:PSub[x] => Time:PSub[y]
        Cause:PSub[x] => Cause:PSub[y]
        }
```

Figure 5. Lexicalized rule for double annotation transfer

The manual validation will resolve the cases in which a FE double annotated in English must be simple annotated in Romanian, if any. Until now, all the revised double annotation cases were cases where the ambiguity was kept in Romanian.

4.2. Imbrications

There are cases when a word can be part of two semantic elements without being double annotated. The imbrications process is common in the English annotations mainly in the possessive noun phrases (e.g. [[*his*]$_{\text{Prot}}$ ankle]$_{\text{BodyPart}}$). In Romanian, when the possessive pronoun is placed before the verb as a reflexive pronoun, the imbrications disappear (Fig 6).

[Când şi-a revenit după atac]$_{\text{Time/Cause}}$, a căzut şi [*şi*]$_{\text{Exp}}$-a RUPT [mâna]$_{\text{BodyPart}}$.

[When she got over the stroke]$_{\text{Time/Cause}}$ [she]$_{\text{Exp}}$ fell and BROKE [[*her*]$_{\text{Exp}}$ hand]$_{\text{BodyPart}}$.

Figure 6: Imbricate FEs in English, but not in Romanian

Even if we don't have an absolute correspondence between the whole FE *BodyPart* form English into Romanian, the noun *mâna* (*hand*) is correctly annotated in Romanian as representing the *BodyPart* frame.
The import of the annotation works also when the Romanian target-word is a gerund followed by a reflexive pronoun and a noun phrase, as in the following example:

[Josef Jakobs]$_{\text{Prot}}$ a aterizat într-un câmp de cartofi în North Stifford , Essex , căzând greu şi RUPÂNDU-**[*şi*]$_{\text{Prot}}$ [glezna]$_{\text{BodyP}}$** .

[Josef Jakobs]$_{\text{Prot}}$ landed in a potato field in North Stifford , Essex , falling heavily and BREAKING **[[*his*]$_{\text{Prot}}$ ankle]$_{\text{BodyP}}$** .

Figure 7: Another example of imbricate English FEs not present in Romanian

Although apparently similar to the English structure, in the Romanian sentence, the frame elements are not imbricate, but successive, since the regent of the pronoun **şi**, is not the noun *glezna* (*ankle*), but the gerundive verb.

4.3. Unexpressed Semantic Frames

A FE can be expressed in English, but implicit in Romanian, or vice-versa (Fig. 8). If the first case poses no problems to the transfer, the second one supposes importing roles unexpressed in English.

(a) [Sângele]$_{\text{Undergoer}}$ se ÎNGROŞĂ [spre capătul fibulei zdrobite]$_{\text{Place}}$.
[Blood]$_{\text{Undergoer}}$ had CONGEALED [thickly]$_{\text{Manner}}$ [on the end of the smashed fibula]$_{\text{Place}}$.

(b) LĂSAŢI-[vă]$_{\text{Protagonist}}$ [de fumat]$_{\text{Process}}$.
QUIT [smoking]$_{\text{Process}}$.

Figure 8: Frame Elements expressed in English but unexpressed in Romanian (a) and vice-versa (b)

In the (b) example above, the English verb *to quit* is translated by *a se lăsa*, where the reflexive pronoun *se* expresses the person that makes and supports, in the same time, the action, therefore being the *protagonist* of the action.
A situation apart is represented in Figure 9, where "uneaten food", had been translated by "mâncare" (the exact translation of English noun *food*), because the adjective corresponding of *uneaten* has, in Romanian, the same root with the noun *mâncare (food)* and his utilization in such a case would be inappropriate.

Nu uita că [orice resturi de **mâncare**]$_{\text{Und}}$ se vor DESCOMPUNE [în bazinul tău]$_{\text{Place}}$ şi vor murdări apa .

Remember that [any traces of **uneaten food**]$_{\text{Und}}$ will DECOMPOSE [in your tank]$_{\text{Place}}$ and foul the water .

Figure 9: Partial correspondence between English and Romanian frame

The absence of the adjective in the Romanian sentence is imposed not only by the syntactical or morphological specificity of the language, but also by some pragmatic reasons. Somehow, any food is uneaten (yet). So "uneaten food" is strongly sensed as pleonastic by native Romanian speakers. The expression refers to an insignificant quantity of food left, by antithesis with the food already ingested, implicitly mentioned. The semantics of this frame element differently lexicalized in the two languages is nevertheless the same.

4.4 Problems in the Translation of Target-Words

The problems imposed by the morphological conversion during translation (the nominalization is the most frequent phenomenon) can be solved using conversion rules, as shown in (Husarciuc & al., 2005). A specific case concerns the problems imposed by the translation of a phrasal verb. There are some situations when the entire structure is considered a single lexical unit (LU) and others when a similar morpho-syntactical structure is annotated as a multi-word expression (MWE), composed by two or more LUs (see Ruppenhofer & al., 2002). When the phrasal verb is considered a single LU, its translation into Romanian is either a collocation or a simple verb (as in Figure 10).

Many times [I]$_{\text{Addressee}}$'m called up [by a local doctor]$_{\text{Communicator}}$ and asked to do what is *called* a DV -- a domiciliary visit to assess the mental state of an individual .

De multe ori am fost contactat [de un medic din zonă]$_{\text{Communicator}}$ şi [mi]$_{\text{Addressee}}$ s-a cerut să fac ceea ce se *cheamă* o VD -- vizită la domiciliu, pentru evaluarea stării mentale a unui individ.

Figure 10: Multi-word expression translation

Nevertheless, problems occur during the annotation import

of the idiomatic expressions, which generally don't appear as a single LU. There are also cases when the same expression has different meanings, depending on the contexts, problematic especially when they represent (parts of) core FEs. How to deal with sentences such as:

1) [I]$_{Avenger}$'ll get even [with her]$_{Offender}$.

2) Cheat me and [I]$_{Avenger}$'ll get even.

where the same expression (*get even*) is translated differently?

1) Vom fi chit.

2) Înşeală-mă şi [eu]$_{Avenger}$ voi face la fel.

In these cases, of interest is the solution found by Moszczynski (2007), who classifies MWEs in "units that should be processed before syntactic analysis" and "expressions whose recognition should be combined with the syntactic analysis process", e.g. phrasal verbs. But not all situations that work for Polish also apply to Romanian. A supplementary considered solution is to use a bilingual database, containing the English and Romanian MWEs in a lemmatized form, which would help the alignment process.

4.5. The lack of total correspondence between frames

In the English FrameNet, similar sentences can serve as examples for different, though related, frames. In (Ruppenhofer & al., 2006), the relation between *Communication* and *Contacting* frame is illustrated by two sentences that are apparently semantically equivalent:

(a) I e-mailed him my new phone number. (Contacting frame)

(b) I communicated my new phone number to him by e-mail. (Communication frame)

The FrameNet authors explain the different classification since only the second sentence "entails that the Recipient received the message", because "with Contacting, no actual successful communicative act is implied, only the successful completion of acts which could establish the communication" (Ruppenhofer & al., 2006, p. 17). The Romanian translation of both sentences is similar due to the absence in Romanian of a simple verb corresponding to *e-mail*:

I-am trimis prin e-mail noul meu număr de telefon.

The solution is to include the Romanian sentence into the most general frame (i.e., the *Communication frame*).

Using the above-presented particular situations, a formalism that simplifies the work of a human annotator, and afterwards configures the activity of an automatic machine, can be created.

5. Conclusions

In this paper, we have presented a fast method for the realization of a Romanian semantic role corpus. The main purpose of creating a quick semantic annotated database is using it as training corpus for automatic labelled semantic frames detection. The import method was preferred to the 'classical' creation by hand of a manually annotated corpus because of its time sparing and possible automation. We currently investigate the possibility of using a translation engine for the most resource consuming task, namely the translation of the English sentences.

The resulted resource can also be used as a verifying resource for the syntactic annotation. FrameNet comes, besides frame elements, with a syntactic analysis of each the sentences. This annotation can also be imported, but it is not representative, since the syntax represents the surface level, thus the one with language specificities. Therefore, the Romanian sentences are syntactically parsed at the alignment stage. The comparison of the two annotations is a very useful to create a syntax transfer model.

6. References

Artstein, Ron, Poesio, Massimo. (2005). Kappa cubed = alpha (or beta). *Technical Report NLE Technote 2005-01*, Essex, UK.

Backer, Collin F., Fillmore, C., Lowe, J. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*. Montreal, Canada.

Fillmore, Ch. (1968). The case for case. In Bach and Harms (Eds.), *Universals in Linguistic Theory*. Holt, Rinehart, and Winston Ed., New York.

Husarciuc, Maria, Trandabăţ, Diana, Lupu. Monica. (2005). Inferring Rules In Importing Semantic Frames From English FrameNet Onto Romanian FrameNet. In *Proceedings of 1st ROMANCE FrameNet Workshop*, held at EUROLAN 2005 Summer School. Cluj-Napoca, Romania, pp.44-49.

Moszczynski, Radoslaw. (2007). A practical Classification of Multiword Expressions, In *Proceedings of the 45th Annual Meeting of the ACL. Companion Volume: Proceedings of the Student Research Workshop.* Prague, Czech Republic, pp. 19-24.

Ruppenhofer, Josef, Baker, C.F., Fillmore, C.J. (2002). Collocational Information in the FrameNet Database. In *Proceedings of the Tenth Euralex International Congress*. Copenhagen, Denmark, Vol. I, pp. 359-369.

Ruppenhofer, Josef, Ellsworth, Michael, Petruck, Miriam, Johnson, Christopher and Scheffxzyk, Jan. (2006). FrameNet II: Extended Theory and Practice, International Computer Science Institute, University of California, Berkeley.

Trandabăţ, Diana. (2007). Semantic Frames in Romanian Natural Language Processing. In *Proceedings of the NAACL-HLT 2007 Companion Volume: Doctoral Consortium*. Association for Computational Linguistics, Rochester, New York, USA, pp. 29-32.

Tufiş, D., Ion R., Ceauşu, Al., Stefănescu, D. (2005). Combined Aligners. In *Proceeding of the ACL2005 Workshop on «Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond»*, Ann Arbor, Michigan.