Interlingual annotation of the Catalan Dataset

Irene Castellón
Universitat de Barcelona
Glòria Vàzquez
Universitat de Lleida
Ana Fernández, Elisabeth Comelles
Universitat Autònoma de Barcelona
Victoria Arranz, David Farwell
Universitat Politècnica de Catalunya

0. Introduction

The following is a summary of the issues that arose during the annotation of the Catalan dataset. After briefly presenting the background for the discussion in section 1, we turn to a presentation of the questions that arose related to format in Section 2, to segmentation in Section 3 and to annotation in Section 4.

1. The task

Briefly, the task included:

- translating the Catalan version of an article in the UNESCO Currier into English*;
- segmenting three versions of the same article (the Catalan version, an original English version and the translated English version) into contiguous minimal translation units;
- annotating each unit of each version for conceptual type (as an entity, as an event, as a state or as something else);
- annotating each segment of each version for semantic class based on EuroWordNet (EWN).

* The English translation was done by Michael Younkman of Learningworks in Barcelona, Spain.

Each annotation consisted of:

- an index,
- an expression,
- a conceptual type (ENTITY, EVENT or STATE) with co-indexing to indicate co-reference,
- a semantic class (corresponding to a EWN synset identifier) with pointers to arguments where relevant,
- comments.

For example, in:

```
15 | currently provides technical assistance to
          | EVENT | ASSIST(01442355v)[ENTITY-2,ENTITY-16]
```

"15" is the index, "*currently provides technical assistance to*" is the expression being annotated, "EVENT" is the conceptual type and "ASSIST(01442355v)[ENTITY-2,ENTITY-6]" is the semantic class, "01442355v" being the EWN synset identifier and "ENTITY-2" and "ENTITY-6" pointing to the arguments of the assist event.

Two people were assigned to annotate each version. They each annotated the text independently and then they reviewed their annotations together so as to settle on a single common annotation. Then, all six people met and compared annotations with a eye toward standardizing the form of the annotations and the methodology for assigning them. Issues of content were also discussed such as the choice of conceptual type and of semantic class (i.e., synset) as well as whether or not differences in annotation across versions were appropriate and, if so, why.

In what follows, we present a number of issues that arose during the process along with our suggestions for resolving them.

2. Format

As mentioned, each annotation consisted of five fields:

- an index,
- an expression,
- a concept type,
- a semantic class (EWN synset identifier),
- comments.

The second field contained the expression being annotated. We used, "[ ]" to indicate segments which are implicitly understood but are not explicitly present in the text and which are in one way or another referred to (e.g., ellipted elements such as implicit subject pronouns, implicit elements in coordinate constructions or implicit heads in correlative constructions – e.g., (Catalan) *la* [*condició*] *de*… (the [condition] that…).

The third field indicates the conceptual type of the expression. These may be indexed in order to show a co-reference relation.

The fourth field contains the semantic class or EWN synset identifier. In some cases, especially for Catalan, the relevant EWN synset could not be found for the language in question. In such cases, our strategy was to first see if an equivalent could be found is one of the other language versions of EWN and, if not, to make up a dummy synset identifier. In regard to events, references to the actors are found in square brackets using the head of the relevant referring expression or, if that expression has been ellipted, a variable as a pointer.

Thus, for example, for:

Bancosol ha esdenvigut un banc comercial

the annotation of the second segment, *ha esdenvigut* (has transformed itself), has the following form:

```
24 | ha esdevingut
          | EVENT | esdevenir(00089026v)[BancoSol,banc]
```

The comments field was used on some occasions to list alternative possible semantic classes wherever the choice was unclear.

3. Segmentation

Segmentation was done intuitively the first time through and then revised on the basis of a consensus agreement among the members of the group. In fact, segmentation often depended on decisions related to the annotation task and to the treatment of reference.

One question that arose was how to best deal with copular complements (especially predicate nominals). On the one hand, they might be treated as states, essentially merging the copular verb *be* (or its equivalents) with the following nominal and viewing the whole as a unified predicate as in the following example:

```
1 | ACCION International | ENTITY | ORGANIZATION(05149489n)
2 | is a U.S.-based private non-profit organization | STATE
           | BE(01472320v)_ORGANIZATION(05149489n)[ACCION]
```

On the other hand, they might be treated as generic noun phrases and, therefore, not referring. Or, again, they might be treated as referring noun phrases and, therefore, co-referring with the subject. We decided to treat them following the first strategy, that is, as a single segment.

A second question was related to differing treatments of certain verb+object constructions. For instance, for one annotator *currently provides technical assistance to* was treated as unified predicate akin to *currently assists* (*technically*) as in:

```
4 | currently provides technical assistance to
          | EVENT| ASSIST(01442355v)[ENTITY-1, ENTITY-5]
```

A second annotator chose to analyse the construction more compositionally as in:

```
4 | currently provides
          | EVENT | PROVIDE(04025567v)[ENTITY-1, EVENT-5]
5 | technical assistance
          | EVENT | ASSIST(01442355v) [ENTITY-1, ENTITY-5]
```

Neither of these analyses offers obvious advantages and each has certain disadvantages, but the consensus for the mark-up was to provide a unified analysis where plausible.

Yet another question arose that concerned the treatment of possessive adjectives. Although there was little disagreement once the issue was discussed, initially some annotators did not deal with possessive adjectives (e.g., *its*) or the Saxon genitive (e.g., *BancoSol's*) as separate referring expressions having their own index. It was agreed, however, that they should be treated as separate indexed segments. Thus they were marked as below:

```
16 | Its | ENTITY-1 | ORGANIZATION(05149489n)
```

```
17 | network | ENTITY | NETWORK(05354409n)
```

rather than as:

```
16 | Its network | ENTITY | NETWORK(05354409n)
```

Other segmentation procedures agreed to included:

- grouping determiners and adjectives with their head,
- grouping pre-modifying partitives with the following head, e.g., ***some of*** *the X*, unless they enter into a co-reference relation, e.g., ***tres de*** *las institucios* (three of the institutions),
- treating post-modifying adjective phrases (i.e., adjectives with complements) as separate segments, e.g., *préstecs* ***concedits cada any*** (loans granted each year) consists of three segments and *ensenyaments* ***adquirits de l'experiència*** (lessons learned through experience) consists of four,
- treating certain fixed or periphrastic phrases as a single segment when they behave as a single semantic unit, e.g. *tenir accés* (to have access), *obrir les portes* (to open its doors) or *sense finalitat de lucre* (non-profit),
- treating compound tenses as single segments.

4. Annotation

4.1. Entities, Events and States

There were often doubts as to when to characterize something as an entity as opposed to an event. Sometimes the interpretation could be disambiguated by context. For example, in *esforç d'informació* (effort of information), *informació* is annotated as an event because it is implied to be provided. In *una cartera de préstecs* (portfolio of loans), *préstecs* is annotated as an entity because portfolios contain documents. In *servei d'estalvi* (service of savings), *estalvi* is annotated as an event because the service provided is that of opening (or holding) a savings account. In *banc de dipòsit*, (bank of deposit) *dipòsit* is annotated as an event because, again, it is understood that it is a place where people deposit money. Finally, in *pèrdues del banc* (losses of the bank), *pèrdues* is annotated as an entity because the losses here are sums of money and sums are entities.

As discussed above, we also decided to treat copulas along with their complements as single unified attributes or states. If the complement was a nominal, an entity marker was added to the state marker, e.g., STATE(ENTITY).

```
12 | és una organització | STATE(ENTITY)
          | SER1(01472320v)_ORGANISME1(05355086n) [BANC1]
```

Noun phrases which are typically entities but which have an event interpretation, e.g., *esforç d'informació*, are treated as if they were complements of the ellipted verb, and they are annotated as EVENT(ENTITY). The corresponding semantic class consists of the relevant synsets merged into a single complex predicate.

```
Esforç    ...
d'
[donar] informació | EVENT(ENTITY)
```

```
      | DONAR1(xxxxxxxxv)_INFORMACIO1(xxxxxxxxn) [BANC1]
```

There was also a question as to whether or not generic references, such as that made by the use of *non-profit organizations* below, should be indexed. The argument is that they are not actually being used to refer but rather to classify or describe some entity and, therefore, they cannot play a role in any co-reference relation.

```
23 | as
24 | non-profit organizations
                    | ENTITY-SET | INSTITUTION(05152219n)
```

The problem is that, if this strategy is followed, annotators will have to be sensitive to the difference between generic and non-generic reference.

4.2. Semantic Classes

To support the annotation task, we used EuroWordNet to determine they basic representational vocabulary. Specifically, the semantic class is annotated with the number of the corresponding synset. The advantage of using EuroWordNet is that it is developed for both Catalan and English (as well as a number of other languages, including Spanish). Thus, a given synset not only contains entries in, say, Catalan, but also entries in English. This then provides a mechanism for establishing equivalence (or non-equivalence) between annotations across languages. Also, in some cases, if in one language there are no entries in EWN, one can still identify an appropriate synset by looking for the membership of an equivalent term in a related language (e.g., Spanish for Catalan).

One issue that arose in regard to semantic classes, concerned whether or not to include multiple possibilities (i.e., multiple EWN synset indices) or to allow only one. The problem is that it is not always clear which synset a particular usage belongs to and thus it might be easier to simply list the various possibilities. The decision here was to select one and list the other possible analyses in the comments field.

A second issue, here, concerned participial forms. It was agreed that if a given participial were categorized as an adjective in EWN, then it was considered to be a property and, therefore, not annotated. However, if it were categorized as a verb, then it was assumed to refer to an event (or state) and, therefore, annotated. An example of the former situation is *regulated* in:

```
45 - regulated financial institutions
                    | ENTITY-SET | INSTITUTION(05152219n)
```

An example of the latter is *created* in:

```
54 - a non-profit joint venture
                    | ENTITY | JOINT_VENTURE(00437417n)
55 - [] | ENTITY | JOINT_VENTURE(00437417n)
56 - created | EVENT-10
          | ESTABLISH(00942370v) [ENTITY-SET-51, ENTITY-54]
```

```
57 - in
58 - 1986 | ENTITY | YEAR(09127492n)
```

This issue affects the mark-up of the Catalan text to a greater degree than the English texts but none the less arises in all of them. Eventually, of course, adjectives will also be annotated and will quite possibly refer to events or states or even entities and will not all be associated to properties.

A third issue arose in dealing with complex expressions such as compounds verbal expressions (e.g., auxiliary plus main verb). We choose to annotate the main verb.

A fourth issue concerned the annotation of metaphors or fixed expressions such as:

```
1  - ACCION | ENTITY | ORGANIZATION(05149489n)
2  - SPEAKS LOUDER THAN WORDS | EVENT |
      DOES(01449587v)
                [ENTITY-1, ENTITY-SET | THINGS( 00019561n)]
```

Here one annotator provided an annotation corresponding to an interpretation of the punned adage as a whole while the other provided an annotation corresponding to the analytic semantics of the internal expressions of the adage. During the discussion of this divergence it became clear that if the former strategy is adapted (of annotating the interpretation), there is likely to be greater variation in interpretation (interpretations vary to a greater degree than semantic representations) and information about the surface form of the metaphor, which might be useful to translation, is lost. On the other hand, if the latter strategy is used (annotating the analytic semantics of the metaphor), it is quite possible the translation will not be based on the interpretation of the metaphor (this problem is related to the full range of non-literal or semi-literal expressions).

Here we have opted to annotate the interpretation but, alternatively, there appear to be two options: annotate both the semantics of the expression and its interpretation, i.e., provide two ILs for the text, or annotate the semantics and assume that there is a process that takes such an IL as input and produces an alternative IL as output which represents the interpretation.

A fifth question was related to the treatment of proper names. Many proper names are included in EWN and where present, the corresponding synset identifier was used to indicate the PN´s semantic class. But, if the name was not included in EWN, the question then was whether or not they should be provided with some other semantic classification as in:

3 | *MARÍA OTERO* | ENTITY | PERSON(00004865n)

Initially some annotators had simply identified them as entities without searching for a semantic classification but after some discussion it was agreed that semantic classes should be included. One disadvantage of this choice is that there may be variation in the class chosen (e.g., *MARÍA OTERO* could be classified as PERSON, AUTHOR, WOMAN (if this is not a property), EXECUTIVE, or no doubt as something else).

Finally, we also had doubts on occasion as whether or not to annotate a modifier that contributed in such a way to the meaning of the expression as to warrant inclusion in the annotation along with the head of the expression. For instance, in *sistema bancari* (banking system) or *estructures financeres* (financial structure) the modifiers actually seem to determine the meaning of the expression as much, if not more, than the actual head of the expression. But, unless the compound were found in EWN, the semantic contribution of such modifiers was not annotated.

There were other controversial issues and discussions but they generally resolved around the question of whether to annotate the interpretation or to annotate the semantics (e.g., *loans* in *loans averaging ...* as an AMOUNT vs a LOAN or *opened its doors* as BEGIN vs OPEN-DOOR, etc.). In general, the more "fixed" or formulaic the expression, the more likely we opted for an "interpretation" but many cases are simply not obvious.

4.3 Reference

Reference was indicated through the use of indexes. The types of phrases that at one point or another were identified as co-referring included:

- pronouns,
- definite NPs,
- proper names,
- possessive pronouns.

We have considered two techniques for indexing expressions. In one case only referring expressions are indexed or co-indexed if co-referent with a prior expression. In the other, every expressions is given an index as in:

```
1 | ACCION International | ENTITY | ORGANIZATION(05149489n)
2 | is a U.S.-based private non-profit organization | STATE
          | BE(01472320v)_ORGANIZATION(05149489n)[ENTITY-1]
3 | that | ENTITY | ORGANIZATION(05149489n) | ENTITY-1
4 | currently provides technical assistance to
          | EVENT| ASSIST(01442355v) [ENTITY-1, ENTITY-5]
5 | a network | ENTITY | NETWORK(05354409n)
6 | of
7 | institutions | ENTITY-SET | INSTITUTION(05152219n)
8 | in
9 | thirteen countries | ENTITY-SET | COUNTRY(05400698n)
10 | in
11 | Latin America | ENTITY | LATIN_AMERICA(05449837n)
12 | and
13 | six cities | ENTITY-SET | CITY(05390395n)
14 | in
15 | the United States. | ENTITY | UNITED_STATES(05400698n)
```

We opted for this latter approach so as to make co-references to events reported by sentences, clauses or complex nominalizations easier to indicate (i.e., co-refers with expressions I through J).

One issue that arose in regard to reference was whether or not to allow "dummy elements" to be inserted into the annotated text in order to represent ellipted expressions, especially pronouns. We have decided to include them whenever they are not automatically recoverable by virtue of the syntactic contexts (e.g., holes in relative clauses with overt complementizer). This is exemplified by expressions 36 and 49 in the following segment:

```
34 | in
35 | loans | ENTITY-SET | AMOUNT(08180701n)
36 | [] | ENTITY-SET | AMOUNT(08180701n)
                                        | co-refers with 35
37 | averaging | EVENT
     | AVERAGE(01497019v) [ENTITY-SET-35, ENTITY-SET-38]
38 | less than $500. | ENTITY-SET | DOLLAR(08358605n)
…
48 | and
49 | [] | ENTITY-SET | INSTITUTION(05152219n)
                                        | co-refers with 42
50 | have,
51 | in
52 | the last five years, | ENTITY-SET | YEAR(09127492n)
53 | converted into | EVENT
     | CONVERT(00228189v) [ENTITY-SET-49, ENTITY-SET-54]
54 | regulated financial institutions
                    | ENTITY-SET | INSTITUTION(05152219n)
```

Finally, we also considered the question of whether references to groups of entities should be represented differently from references to single entities as exemplified by:

```
14 | this non-profit institution
                    | ENTITY | PROGRAM(03985827n)
```

versus:

```
23 | eighteen independent organizations
                    | ENTITY-SET | INSTITUTION(05152219n)
```

The ENTITY-SET notation makes co-reference to elements of the set more transparent although it means that the annotator will have to be sensitive to the distinction between referring to individuals and to groups of individuals.

5. Conclusion

We have described a number of issues that arose during the annotation exercise. These issues are related to both the form and the content of the annotations as well as the methodology for assigning them. The discussion was divided into three broad areas of concern: format, segmentation and annotation. In regard to annotation, we looked at problems related to the assignment of conceptual type, semantic class and reference relations.