NewsReader Italian and Spanish Guidelines for Annotation at Document Level

NWR-2014-6

Version DRAFT

Manuela Speranza, Ruben Urizar Anne-Lyse Minard
manspera@fbk.eu, ruben.urizar@ehu.eus, minard@fbk.eu

(1) Fondazione Bruno Kessler
Via Sommarive 18, 38123, Povo (Trento), Italy

Building structured event indexes of large volumes of financial and economic data for decision making
ICT 316404

Contents

1 Overview of annotation guidelines

For the annotation of the Spanish and Italian guidelines we adopted the NewsReader guidelines defined for English ([Tonelli *et al.*, NWR2014-2-2]).

In this document we describe only the extensions needed to adapt them to the specific morpho-syntactic features of Italian and Spanish. The revision and adaptation of the annotation guidelines for events is based on the It-TimeML guidelines ([Caselli *et al.*, 2011]) and on the Spanish TimeML guidelines ([Saurí *et al.*, 2009; Saurí *et al.*, 2010; Saurí, 2010]), while the revision and adaptation of the annotation guidelines for entities is based on the I-CAB guidelines ([Magnini *et al.*, 2006]).

2 Contractions of prepositions and definite articles (Italian's articulated prepositions)

In the annotation of entity mentions and time expressions in English, prepositions are excluded from the extent while articles are included (e.g. *to [the family]*, *in [the next months]*). This is problematic for Italian and Spanish which, unlike English, have contractions of simple prepositions and definite articles. This phenomenon, which is common to many prepositions in Italian (e.g. *di*, *a*, *da*, *in*, *su*) and includes both singular and plural (e.g. *al* vs. *agli*) and both masculine and feminine (e.g. *al* vs. *alla*), is limited in Spanish to two contractions, *al* and *del*, in which the prepositions *a* or *de* respectively merge with the masculine singular definite article *el*.
Ex-IT: ***al*** *governo degli Stati Uniti* '(to) the US government'
Ex-ES: ***al*** *gobierno de los Estados Unidos* '(to) the US government'
Ex-IT: ***dal*** *5 novembre* ***al*** *10 dicembre* 'from November 5 to December 10')
Ex-ES: ***del*** *5 de noviembre* ***al*** *10 de diciembre* 'from November 5 to December 10')

Based on the above mentioned related work, we decided that these contractions should not be split but treated as single units in the annotation process. In particular:

- ENTITY MENTIONs: following the I-CAB guidelines, they should be included in the extent;
- TIMEX3s: following It-TimeML and Spanish TimeML, they should not be included in the extent; when a time expression is introduced by a contraction, this is usually to be marked as temporal SIGNALs.

Ex-IT: ENTITY MENTION *[**al** governo degli Stati Uniti]* '(to) the US government'
Ex-ES: ENTITY MENTION *[**al** gobierno de los Estados Unidos]* '(to) the US government'
Ex-IT: SIGNAL+TIME EXPRESSION *[**dal**] [5 novembre] [**al**] [10 dicembre]* 'from November 5 to December 10'
Ex-ES: SIGNAL+TIME EXPRESSION *[**del**] [5 de noviembre] [**al**] [10 de diciembre]* 'from November 5 to December 10'

3 Modals

According to the NewsReader guidelines for English ([Tonelli *et al.*, NWR2014-2-2]), which are based on TimeML ([Pustejovsky *et al.*, 2003]), modal verbs are not annotated as events and the `modality` attribute is associated to the main verb (the value of the attribute is the token corresponding to the modal verb). On the other hand, the annotation of modals in NewsReader for Italian and Spanish follows It-TimeML and Spanish TimeML respectively: verbs expressing modality are themselves annotated as events (in particular, in the case of NewsReader, as events of type GRAMMATICAL); in addition, a GLINK (grammatical link) is created between the modal (source) and the main (target) verb (the `modality` attribute associated to the main verb is optional).
Ex-ES: *[podemos] [jugar]* 'we can play'
Ex-IT: *[possiamo] [giocare]* 'we can play'
Ex-ES: *[tendrán] que [mejorar]* 'they will have to improve'
Ex-IT: *[dovranno] [migliorare]* 'they will have to improve'
Ex-ES: *[podrías] [descansar]* 'you could / might take a rest'
Ex-ES: *[potresti] [descansar]* 'you could / might take a rest'

4 Clitics

For Spanish and Italian, we have devised specific guidelines to handle clitics, which do not exist in English.
Ex-IT: *Aveva gia' deciso di **parlargli*** 'He had decided to talk to him'
Ex-ES: *Había decidido **hablarle*** 'He had decided to talk to him'

As with contractions of prepositions and definite articles, we have decided

to leave the annotation at token level in the case of clitics. In particular, in the case of a token composed of a verb (i.e. an event) and a clitic (i.e. a pronominal mention of an entity), the whole token will be annotated both as an entity and as an event. As it is important to distinguish the two annotated elements, the `head` attribute of the entity mention (see NewsReader Guidelines, section 3.2) is not optional for clitics as it is for all other types of entity mentions, and the `pred` attribute of the event mention (see NewsReader Guidelines, section 5.2.1) is not optional either.

Ex-IT: EVENT MENTION: [parlargli], pred "parlare"
Ex-IT: ENTITY MENTION: [parlargli], head "gli"

Ex-ES: EVENT MENTION: [hablarle], pred "hablar"
Ex-ES: ENTITY MENTION: [hablarle], head "le"

As far as clitics in pronominal verbs are concerned, we have created specific guidelines for the different classes. Truly reflexive (the object of the action is the same as the subject) and reciprocal pronouns (expressing mutual action or relationship among the referents of a plural subject) are annotated as entities. In the case of benefactive (the focus refers to the person or thing an action is being done for) and pseudo-reflexive pronouns (which occur with intransitive pronominal verbs), we have no entity annotation.

Ex-IT: *[**Mi**] sono ferito in montagna* 'I hurt myself in the mountains'
Ex-ES: *[**Me**] lastimé en la montaña* 'I hurt myself in the mountains'
Ex-IT: *Quelle due persone [**si**] amano* 'Those two people love each other'
Ex-ES: *Esas dos personas [**se**] aman* 'Those two people love each other'
Ex-IT: ***Mi** sono lavato le mani* 'I washed my hands'
Ex-ES: ***Me** lavé las manos* 'I washed my hands'
Ex-IT: ***Mi** sono mosso troppo tardi* 'I acted too late'
Ex-ES: ***Se** movía demasiado deprisa* 'He moved too fast'

When the Spanish "se" and the Italian"si" are used as impersonal pronouns (which corresponds to 'one', 'you', 'we', or 'they' in English) and as passive pronouns, they are not annotated.

Ex-IT: ***Si** dice che sia molto intelligente* 'they/people say he is very smart'
Ex-ES: ***Se** dice que es muy inteligente* 'they/people say he is very smart'
Ex-IT: *Da qui **si** vede il lago* 'from here the lake can be seen'
Ex-ES: *Desde aquí **se** ve el lago* 'from here the lake can be seen'

5 Multiword events

For the annotation of English event mentions, the *minimal chunk* rule is applied. However multi-token entries such as phrasal verbs, idioms and prepositional phrases appearing in either the American or the British version of the Collins on-line dictionary are considered an exception to this minimal chunk rule (see section 5.1 in the annotation guidelines).

As for the annotation of Spanish event mentions, the online version of the *Diccionario de la Real Academia Española* (`http://lema.rae.es/drae/`) is used as a referent to decide which multiword expressions should be annotated (e.g. *dar a conocer* 'announce; make something known', *tener lugar* 'take place').

6 Misalignments

The corpora used to build the benchmark for Italian and Spanish consist of translations of the English corpus. The alignment between the source corpus in English and the corpora in the target languages was made at sentence level. We took advantage of this alignment to project the intra-document annotation to the other languages. The annotation of the Spanish and Italian texts was carried out taking as a starting point the English annotation. Markables in the translated texts were aligned with their corresponding English ones whenever this was possible so that the features of the markables and all the relations between them could be imported from the English corpora to the corpora in Spanish and Italian.

Although the texts in the Spanish and Italian corpora were translations of the English ones, obviously word-to-word translation was not always feasible. Due to this translation divergences, it was not always possible to make a one-to-one alignment of Spanish or Italian markables with their corresponding English ones.

The translation divergences are countless, but next we will give an account of some of the most frequent types of translation divergences that caused misalignments between English markables and the Spanish and Italian ones.

On the one hand, some English constructions may be translated into

Spanish and Italian with a single word. For instance, several light verb constructions and verb idioms may have a one-word equivalent in the Spanish or Italian.

Ex-ES: *has* ***given a boost*** ⇒ *ha* ***impulsado*** (lit. 'it has boosted')
Ex-IT: *has* ***given a boost*** ⇒ *ha* ***stimoltato*** (lit. 'it has boosted')
Ex-ES: *Apple* ***has plans*** *to import* ⇒ *Apple* ***planea*** *importar* (lit. 'Apple plans to import')
Ex-IT: *Apple* ***has plans*** *to import* ⇒ *Apple* ***intende*** *importare* (lit. 'Apple plans to import')
Ex-ES: ***bringing*** *production [...]* ***to an end*** ⇒ *y* ***acabó*** *la producción [...]* (lit. 'and production was finished')
Ex-IT: ***bringing*** *production [...]* ***to an end*** ⇒ *y* ***interruppe*** *la produzione [...]* (lit. 'and production was finished')

Also, many verb-particle construction (phrasal verbs) are translated with a single word.

Ex-ES: *they* ***gave up*** ⇒ *se* ***rindieron***
Ex-IT: *they* ***gave up*** ⇒ *se* ***rinunciarono***
Ex-ES: *he* ***put off*** *his appearance before the press for an hour* ⇒ ***retrasó*** *en una hora su comparecencia ante la prensa*
Ex-IT: *he* ***put off*** *his appearance before the press for an hour* ⇒ ***posticipó*** *di un'ora la sua apparizione di fronta alla stampa*

Other English constructions may also have a one-word equivalent.

Ex-ES: *it* ***is believed to be worth****... ⇒ se* ***estima*** *en... (lit. 'it is estimated in')*
Ex-IT: *it* ***is believed to be worth****... ⇒ si* ***stima*** *in... (lit. 'it is estimated in')*

On the other hand, sometimes the translation for some English one-word verbs is also a multiword construction, for example a light verb constructions.

Ex-ES: *Papermaster did not* ***comment*** *on the situation ⇒ Papermaster no* ***realizó*** *comentarios' al respecto (lit. Papermaster did not make comments on...')*
Ex-IT: *Papermaster did not* ***comment*** *on the situation ⇒ Papermaster non* ***fece*** *commenti riguardo a (lit. Papermaster did not make comments on...')*
Ex-ES: *they* ***used*** ⇒ ***hicieron uso*** *de (lit. 'they made use of')*

Ex-IT: *they* ***used*** ⇒ ***fecero uso*** *di (lit. 'they made use of')*

Furthermore, some multiword expressions may be translated through verbal periphrases in Spanish or Italian, for example, the phrasal verb *call back* may be translated in Spanish with the periphrasis *volver a llamar* (lit. 'return to call') comprising two event verbs while there was a single one in English. Equally, the adverb *again* may also be translated with the periphrasis '*volver a* + infinitive' (lit: 'return + infinitive).

Ex-ES: *Boeing allowed to* ***bid*** *for [...] contracts* ***again*** ⇒ *Boeing puede* ***volver a presentarse*** *a contratos de [...] (lit: 'Boeing can return to bid contracts [...]')*

In all these divergences, the alignment between source and target markables is made between the elements semantically closer (e.g. English noun *plans* with Spanish verb *planea*) rather than between elements belonging to the same part of speech (a verb with a verb).

Another frequent misalignment is due to the fact that Spanish and Italian, unlike English, are null-subject languages. That means the pronoun functioning as subject in English is often elided in the Spanish and Italian translations.

Ex-ES: ***She has been*** *with General Motors for 33 years* ⇒ *Ø* ***Lleva*** *33 años trabajando en General Motors*
Ex-IT: ***She has been*** *with General Motors for 33 years* ⇒ *Ø* ***Lavora*** *da 33 anni presso General Motors*

In these cases no alignment has been made with the English pronominal entity mention. However, even if the subject is not made implicit in Spanish and Italian, the verb holds the person information within it, e.g.:

Ex-ES: *lleva* = 'carry' + 3 SG PRES. ('he/she/it carries')
Ex-IT: *lavora* = 'has been working' + 3 SG PRES. ('he/she/it has been working')

Therefore, we plan to annotate these elided subjects as empty entity mentions in the near future. This way, the empty entity mention could be linked to the PERSON entity instance it refers to (through the REFERS_TO relation) as well as with the verbal event through the HAS_PARTICIPANT relation.

References

[Caselli *et al.*, 2011] Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151, 2011.

[Magnini *et al.*, 2006] Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*, 2006.

[Pustejovsky *et al.*, 2003] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34, 2003.

[Saurí *et al.*, 2009] R Saurí, O Batiukova, and J Pustejovsky. Annotating events in spanish. timeml annotation guidelines (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2009-01, 2009.

[Saurí *et al.*, 2010] R Saurí, E Saquete, and J Pustejovsky. Annotating time expressions in spanish. timeml annotation guidelines (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2010-02, 2010.

[Saurí, 2010] Roser Saurí. Annotating temporal relations in catalan and spanish. timeml annotation guidelines. (version tempeval-2010). Technical report, Barcelona Media. Technical Report BM 2010-04, 2010.

[Tonelli *et al.*, NWR2014-2-2] Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. NewsReader Guidelines for Annotation at Document Level. Technical report, Fondazione Bruno Kessler, NWR2014-2-2.