Annotating modality cross-linguistically: theory, practice, problems

Session: Main Session
Topics: theoretical perspectives, typological perspectives, experimental approaches, mood/modality

In the current contribution, we present a multi-lingual annotation scheme for modality and its implementation to a corpus of parallel/comparable texts (see List of Corpora below for details). The scheme shows some innovative features over state-of-the-art annotation proposals: i) hierarchical and layered structure; ii) primacy of the functional level; iii) identification, characterisation, and linking of modality triggers.

Recently, the Computational Linguistics and Natural Language Processing communities have shown interest in automating the recognition of extra-propositional components of meaning in general and modality in particular. The first step towards the development of systems that can automatically deal with the interpretation of modality is the creation of appropriate, annotated resources. The last few years have witnessed the development of annotation schemes and annotated corpora for different aspects of modality in different languages (Nierenburg and McShane (2004), Hendrickx et al. (2012), Baker et al (2012) , Wiebe et al (2005), Szarvas et al. (2008), Sauri and Pustejovsky (2009), among others). While important contributions, these remain mainly separate efforts, and no shared standards for converting modality-related issues into annotation categories are found. Under this respect, linguistic typology has already gone a long way in the study of modality across languages.

We promote (i) a cross-linguistic annotation model of modality which relies on a wide, typologically motivated approach, and (ii) a hierarchical, layered model accounting for both factuality and speakers attitude (modality in the tool), while modelling these two aspects through separate annotation schemes. We also take care of characterising the linguistic triggers of these two aspects. Working in a multilingual environment eases the task of leaving the layer of functional categories distinct from the actual linguistic realisation, while making it possible to observe how each language encodes with its own means what is specified at the functional level.

We have implemented our scheme using the MMAX2 annotation tool (Müller and Strube 2006), which allows for customised categories to be organised hierarchically, and typed links between annotated entities. This way we can code and visualise separate links between triggers of modality and triggers of factuality and let each trigger have its own specific features. This is crucial as features might differ when triggering factuality or modality, even if the linguistic item is exactly the same. For example, "permettono" (en: they permit) is a trigger of factuality on a macro-level due to its being in the indicative form – a morphological feature (Figure 1). However, the same string also triggers both the factuality and the modality of what follows thanks to its semantics – a lexical feature (Figure 2). The tool allows to specify this through separate annotations of the same string and separate typed links (Figure 3 and Figure 4).

After discussing some crucial issues to achieve shared standards in modality annotation, we will illustrate our scheme and its implementation to a small corpus of comparable/parallel texts. Eventually, we will discuss annotation-related issues and how they affect the notion of modality from a typological, theoretical, and practical perspective. We will also offer a demo of the annotation tool.

Figure 1: Marking morphological features for the trigger “permettono”.

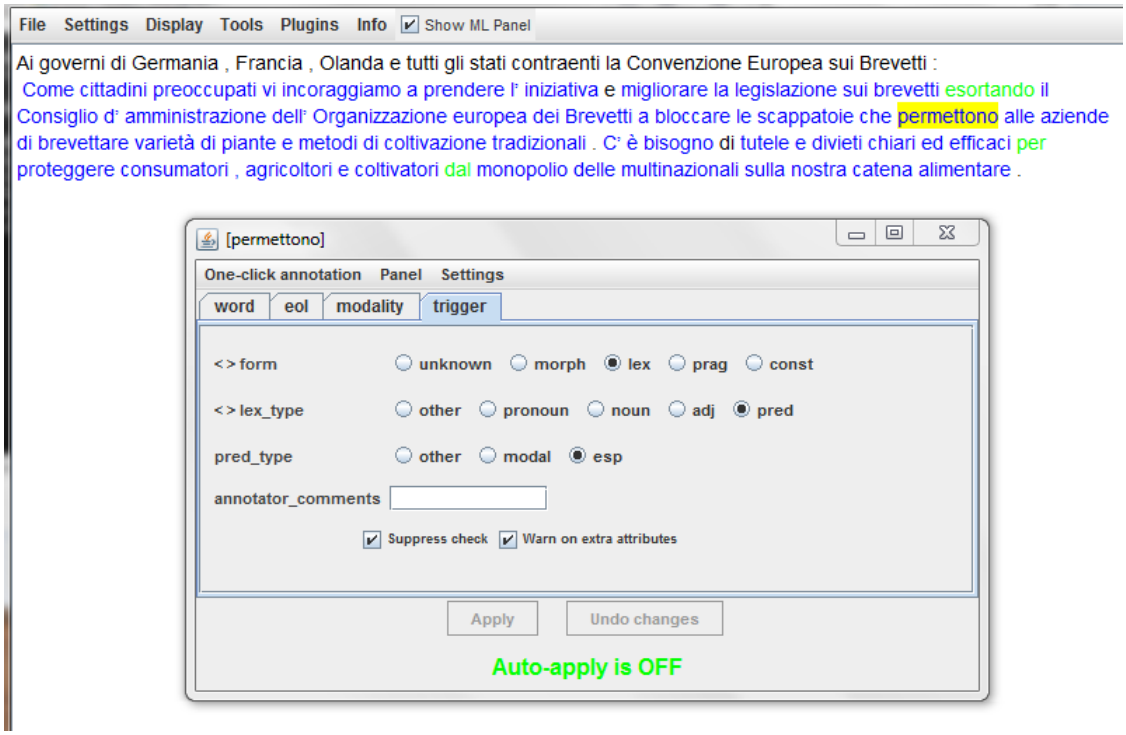

Figure 2: Marking lexical features for the trigger “permettono”.

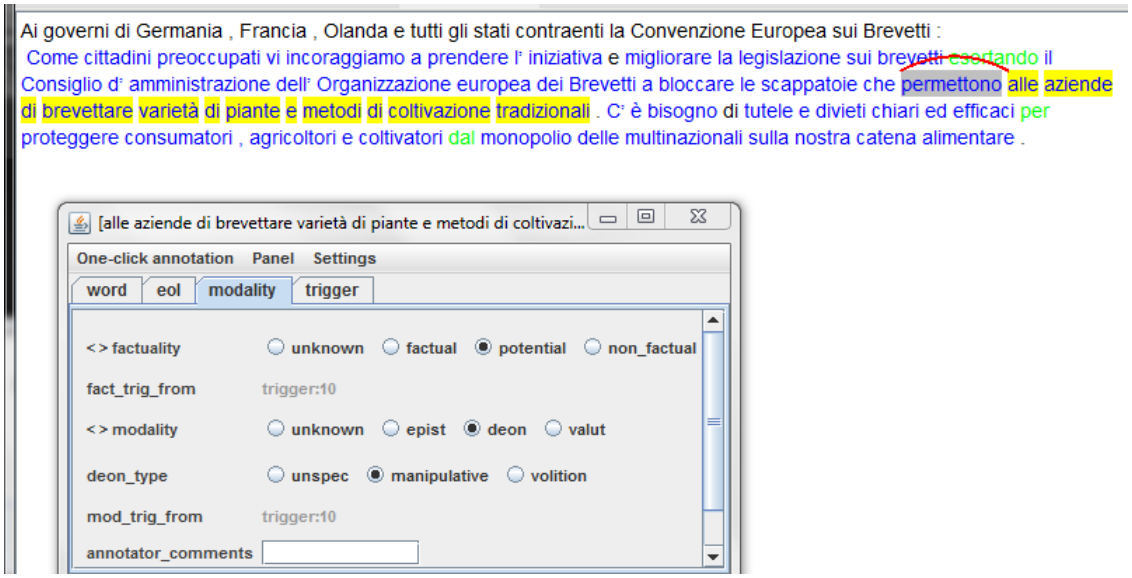

Figure 3: Marking "permettono" as a factuality trigger (red link) and specifying factuality features as well as speaker's attitude ("modality") features for current markable (highlighted).

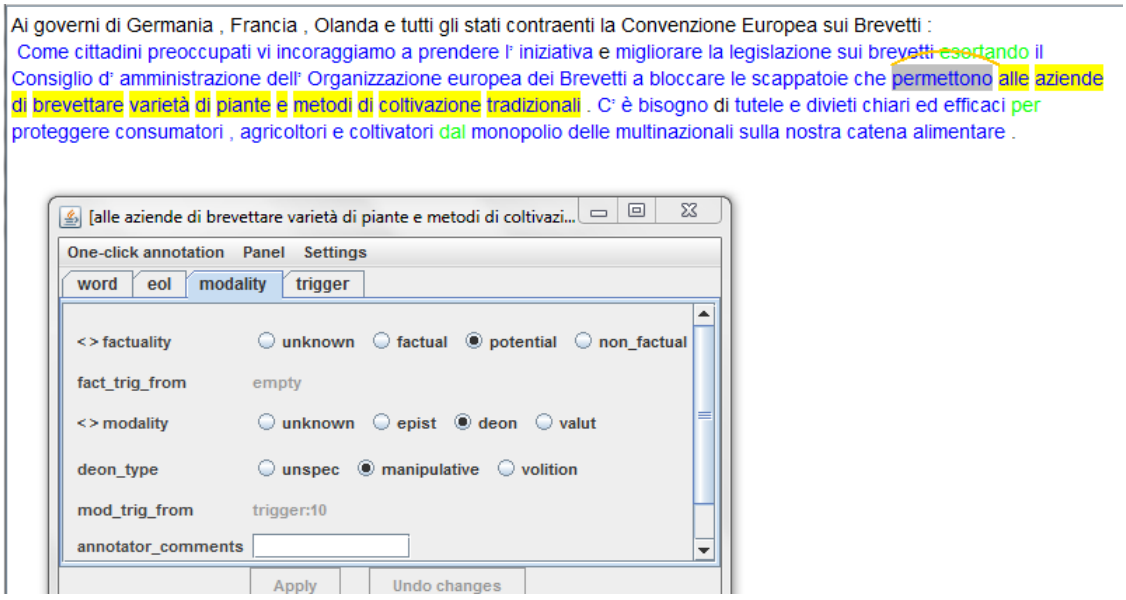

Figure 4: Marking "permettono" as a modality trigger (yellow link) and specifying factuality features as well as speaker's attitude ("modality") features for current markable (highlighted).

References

Baker, K., B. Dorr, M. Bloodgood, C. Callison-Burch, N. Filardo, C. Piatko, L. Levin, and S. Miller (2012). Use of modality and negation in semantically-informed syntactic MT. Computational Linguistics 38(2).

Hendrickx, I., A. Mendes, and S. Mencarelli (2012). Modality in text: a proposal for corpus annotation. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proceedings: the Tenth Machine Translation Summit, Phuket, Thailand, pp. 79-86. AAMT.

Müller, C. and M. Strube (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In: S. Braun, K. Kohn, J. Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214.

Nirenburg, S. and M. McShane (2008). Annotating modality. Technical report, University of Maryland, Baltimore County.

Sauri, R. and J. Pustejovsky (2009). Factbank: a corpus annotated with event factuality. Language Resources and Evaluation 43(3), 227-268.

Szarvas, G., V. Vincze, R. Farkas, and J. Csirik (2008). The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP'08, Stroudsburg, PA, pp. 38-45.

Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2-3), 165-210.

Pilot corpus collection (ca. 6500 words)

Parallel component:

- Declaration of Human Rights (lang: de, en, fr, it, nl)
- Europarl (lang: en, fr, it)
- Petitions (lang: en, es, it)
- Press (lang: en, fr, it)
- Petit Prince (lang: en, fr, it)
- Pavia Corpus of Film Dialogue (lang: en, it)

Comparable component:

- Map Task (lang: en, it)