CATEx: German-Italian parallel corpus of legal and administrative texts

In the late 90ies, the Institute for Specialised Communication and Multilingualism\(^1\) of the European Academy Bolzano/Bozen carried out a research project about (semi-)automatic terminology acquisition, the CATEx (Computer Assisted Terminology Extraction) project. The project had emerged from the need to provide a systematic terminology of legal and administrative language in German and Italian, as due to the equal status of the Italian and the German language in South Tyrol, legal and administrative documents have to be written in both languages.

In order to support the elaboration of a consistent and comprehensive bilingual terminology by computer assisted terminology extraction tools the bilingual parallel corpus CATEx was built. The corpus consists of Italian and German legal and administrative texts, which are translations of one another. It contains about 5 million words and is composed of the following Italian law codes:

- Leggi emanate dalla Provincia Autonoma di Bolzano - Alto Adige / Gesetzbuch des Landes Südtirol
- Codice Civile / Italienisches Zivilgesetzbuch
- Leggi complementari al Codice Civile / Nebengesetze zum italienischen Zivilgesetzbuch
- Codice di Procedura Civile / Italianische Zivilprozeßordnung
- Codice di Procedura Penale / Italianische Strafprozeßordnung
- Ordinamento del Notariato / Italienische Notariatsordnung
- Codice del fallimento ed altre procedure concorsuali / Italienisches Konkursrecht und andere Insolvenzverfahren
- Testo Unico delle Imposte sui Redditi / Einheitstext der Steuern auf das Einkommen
- Processo amministrativo / Der italienische Verwaltungsprozeß

The corpus is encoded with the CES (Corpus Encoding Standard). It includes annotations of part-of-speech tags, and parallel texts are aligned on the sentence level.

For further information on the corpus and its availability please contact us at:  
communication.multilingualism@eurac.edu


\(^{1}\) At the time ‘language and law’.