# T2K$^2$: a System for Automatically Extracting and Organizing Knowledge from Texts

**Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, Simonetta Montemagni**

Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab – www.italianlp.it
via G. Moruzzi, 1 – Pisa (Italy)
name.surname@ilc.cnr.it

**Abstract**

In this paper, we present T2K$^2$, a suite of tools for automatically extracting domain–specific knowledge from collections of Italian and English texts. T2K$^2$ (Text–To–Knowledge v2) relies on a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine learning which are dynamically integrated to provide an accurate and incremental representation of the content of vast repositories of unstructured documents. Extracted knowledge ranges from domain–specific entities and named entities to the relations connecting them and can be used for indexing document collections with respect to different information types. T2K$^2$ also includes "linguistic profiling" functionalities aimed at supporting the user in constructing the acquisition corpus, e.g. in selecting texts belonging to the same genre or characterized by the same degree of specialization or in monitoring the "added value" of newly inserted documents. T2K$^2$ is a web application which can be accessed from any browser through a personal account which has been tested in a wide range of domains.

**Keywords**: Natural Language Processing, Information Extraction, Knowledge Management

## 1. Introduction

T2K$^2$ (Text–To–Knowledge v2) extracts domain–specific information from texts, provides a structured organisation of extracted knowledge and indexes document collections with respect to the automatically acquired information. It relies on a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine learning which are dynamically integrated to provide an accurate representation of the domain–specific content of text corpora in different domains.

T2K$^2$ originates from the the ontology learning system named T2K (Text–to–Knowledge, Dell'Orletta et al. (2008), Lenci et al. (2009)) with a number of main novelties: i) it can be used to extract information from both Italian and English texts; ii) the linguistic pre–processing of texts is carried out by state–of–the–art NLP tools; iii) the range of performed knowledge extraction and organization tasks is wider; iv) it permits manual revision of results before more complex tasks are carried out.

T2K$^2$ can be accessed from any web browser through a personal account: it allows storing and managing uploaded corpora in a personal repository. By relying on a battery of linguistic pre–processing tools in charge of linguistically annotating uploaded corpora, T2K$^2$ performs three main knowledge extraction steps focusing on the extraction of domain–specific entities and of Named Entities as well as on the construction of a relation graph connecting them.

The main features of T2K$^2$ can be summarised as follows. First, it presents itself as a *self–service* system for information extraction which is configurable by following a simple and user–friendly configuration procedure where end–users can choose the domain–specific information to be extracted matching his/her own information needs. Secondly, it supports the user in the construction of corpora which are representative of a given domain and/or are characterized by a homogeneous distribution of features representative of a specific textual genre or readability level. Thirdly, in T2K$^2$ the different levels of extracted information interact resulting in a *multi–dimensional* knowledge representation graph creating the prerequisites for sophisticated text mining processes that would be difficult to carry out if the levels of information were dealt with separately. To this end, T2K$^2$ uses a battery of tools and algorithms meant to visualizing, surfing and managing the knowledge graph. Last but not least, acquired knowledge can also be used for indexing document collections on the basis of extracted domain–specific entities, Named Entities and the relations connecting them. The index can be either used by T2K$^2$ to retrieve and visualize the text spans containing the information searched for by the user or downloaded and used as an input for external tools.

The paper focuses on the linguistic pre–processing and information extraction steps and on the automatic construction of the knowledge graph.

## 2. T2K$^2$

Figure 1 provides an overall picture of the T2K$^2$ workflow. As it can be seen, T2K$^2$ encompasses two main sets of modules, respectively devoted to carry out the linguistic pre–processing of the acquisition corpus and to extract domain knowledge from the linguistically annotated texts. Once the user has uploaded a collection of documents, multi–level linguistic pre–processing is carried out. The linguistically analyzed corpus is used by the linguistic profiling module to support the user in evaluating its homogeneity and representativeness: at this level of analysis, the user can exploit the linguistic profiling results to validate the internal composition of the corpus while extending it with new texts. Once the acquisition corpus has been defined, knowledge

extraction is performed in three different steps, aimed at acquiring domain–specific entities, Named Entities and the relations linking them. T2K$^2$ exploits acquired information to create a structured representation of the input text and to index it with respect to the extracted information: the different information types extracted from the text are organized in a knowledge graph that can be visualized and surfed. The results of the linguistic pre–processing and knowledge extraction steps can be downloaded for inspection by the user who can correct and upload them back into the system in order to proceed in the workflow on the basis of the manually checked results.

### 2.1. Linguistic Pre–processing of Corpora

#### 2.1.1. Linguistic Analysis

In T2K$^2$ linguistic pre–processing of texts is performed by a battery of annotation tools developed by the ItaliaNLP Lab and the Department of Computer Science of the University of Pisa. Each uploaded text is linguistically annotated at increasingly complex levels of analysis, represented by sentence splitting, tokenization, Part–Of–Speech tagging and dependency parsing. In particular, morpho–syntactic tagging is carried out by the POS tagger described in Dell'Orletta (2009) and dependency parsing by the DeSR parser (Attardi, 2006) using Support Vector Machine and Multilayer Perceptron as learning algorithms: as reported in Dell'Orletta (2009) and Attardi and Dell'Orletta (2009), both of them represent state–of–the–art tools for Italian and English. The results of the part–of–speech tagging and of the dependency parsing steps can be downloaded by the user in CoNLL format (Nivre et al., 2007), where *a)* sentences are separated by a blank line and *b)* each token starts on a new line and it is annotated with the following information types: lemma, coarse and fine grained parts of speech, morphological features and syntactic dependency information.

#### 2.1.2. Linguistic Profiling

T2K$^2$ carries out the "linguistic profiling" (van Halteren, 2004) of the collection of texts to be used as acquisition corpus. Linguistic profiling consists in gathering statistics for a wide range of features spanning across different levels of linguistic description (i.e. lexical, morpho–syntactic and syntactic) which can be reliably extracted from automatically analyzed texts with the final aim of reconstructing the text profile. In particular, this pre–processing step is based on the "linguistic profiling" methodology described in Dell'Orletta et al. (2013) and within T2K$^2$ it is meant to support the user in the construction of his/her domain–specific corpus by investigating its underlying linguistic features. It can be usefully exploited, for example, to build corpora containing a homogeneous distribution of linguistic characteristics typically associated with a specific textual genre, since it is a widely acknowledged fact that the extraction of domain entities is affected by the degree of specialization of domain corpora (see, among others, Cabrè (1999)). Linguistic profiling results can also be exploited to monitor the "added value" of newly inserted documents: for this purpose, measures of vocabulary variation such as e.g. the type/token ratio (TTR) can be exploited, as demonstrated by Caruso et al. (2014).

### 2.2. Extraction of Domain–Specific Knowledge

The first two steps of the Knowledge Extraction module are devoted to extract domain–specific entities denoting domain–specific concepts (see Section 2.2.1.) as well as Named Entities specific to the domain under analysis (see Section 2.2.2.). The input corpora are annotated and indexed with respect to the extracted information.

#### 2.2.1. Domain–Specific Entities

The extraction of domain–specific terms denoting domain entities follows the methodology described in Bonin et al. (2010). By default, the automatically POS–tagged and lemmatized text is searched for candidate domain–specific terms, expressed by either single nominal terms or complex nominal structures with modifiers (typically, adjectival and prepositional modifiers), where the latter are retrieved on the basis of a set of POS patterns encoding morpho–syntactic templates for multi–word terms. According to the default configuration, T2K$^2$ extracts domain–specific entities typically expressed through nominal (either single or complex) terms. The domain relevance of multi–word terms is weighted on the basis of the C–NC value (Frantzi and Ananiadou, 1999) score, currently considered as a state–of–the–art method for terminology extraction.

Once a shortlist of well–formed and relevant candidate terms denoting domain entities is extracted from a given target corpus, the user can decide whether to apply the term extraction contrastive method, introduced in Bonin et al. (2010): candidate terms are first searched for in an automatically POS–tagged and lemmatized corpus representative of a different domain and are then weighted on the basis of the C–NC Value; the list of extracted terms is revised afterwards and re–ranked with respect to the associated contrastive score, reflecting their domain relevance which was computed on the basis of the inter–domain distribution of terms. The corpora to be used for the contrastive analysis step are selected among those previously uploaded by the user.

Besides nominal terms, T2K$^2$ also allows the identification of domain–specific properties and events which are expressed in the text through different sequences of POS patterns, headed e.g. by adjectives or verbs: the user can enforce POS–restrictions concerning the start–, internal– and final–token of the POSs sequences to be used for identifying candidate terms, and can also define thresholds for what concerns the length of multi–word terms as well as the amount of single and multi–word terms to be extracted.

At the end of this extraction step, the user can download the list of extracted terms denoting domain–specific entities with an associated domain–relevance score calculated with respect to the whole acquisition corpus. The user can correct the list by pruning the erroneous or simply not relevant terms and re–upload the revised list into the system to proceed with the following analysis steps which will be carried out on the basis of the manually revised data.

The (possibly revised) list of extracted terms denoting domain–specific entities is then used to perform *entity indexing*: the results of this step are represented by a glossary

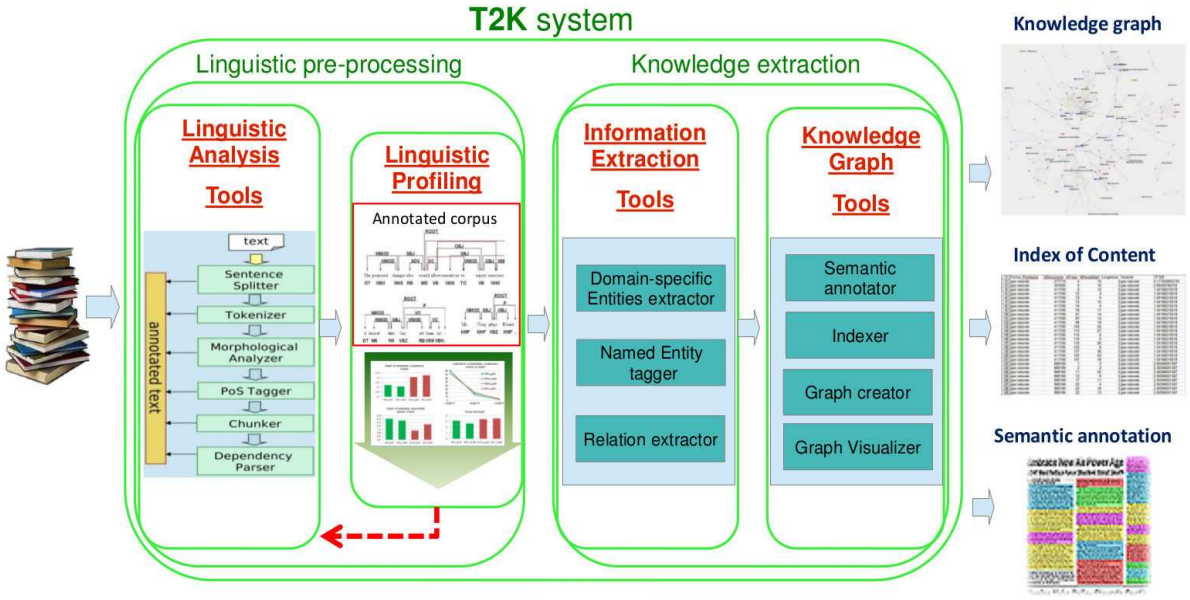



Figure 1: T2K$^2$ workflow.

of the acquired domain–specific terms and an index of the input documents with respect to the extracted terminology. For each term in the glossary, the following information types are provided: the prototypical form, corresponding to the term form most frequently attested in the acquisition corpus, the lemmatized form and the frequency of occurrence of the term (in all attested forms) within the whole document collection. As pointed out in Lenci et al. (2009), the choice of representing a domain term through its prototypical form rather than the lemma (as typically done in ordinary dictionaries) follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts: in fact, domain–specific meanings are often associated with a particular morphological form of a given term (e.g. plural). The index is in tabular format where each column contains the following information types:

- the prototypical form of the indexed entity;
- the identifiers of the document and of the sentence containing the entity mention as well as of the start–token;
- the length of the term denoting the entity (calculated in terms of tokens);
- the variants of the indexed entity attested within the input corpus. Following Nenadic et al. (2004) and Lenci et al. (2009), different types of term variation are considered, i.e. orthographic, morphological as well as structural.
- the TF/IDF score (Salton and Buckley, 1988) measuring the relevance of the indexed entity with respect to the specific document.

Both the glossary and the index can be downloaded by the user for them to be used as input for external tools. In addition, T2K$^2$ allows downloading the input corpora in CoNLL format where the start–, internal– and final–token of the POSs sequences representing each domain term are annotated using the IOB labeling format (Ramshaw and Marcus, 1995), where the start–token is annotated with "B" (marking the begin of the entity mention), the internal– and final–tokens with "I" (marking the internal tokens) and all other tokens with "O".

#### 2.2.2. Named Entities

T2K$^2$ includes a multi–lingual module to extract Named Entities called ItaliaNLP NER. This module is a classifier based on Support Vector Machines using LIBSVM (Chang and Lin, 2001) that assigns a named entity tag to a token or a sequence of tokens. ItaliaNLP NER uses 5 kinds of features:

- orthographic features, i.e. the orthographic characteristics of the analyzed token, e.g. capitalized letters, presence of non–alphabetical characters, etc.;
- linguistic features, i.e. lemma, Part–Of–Speech, prefix and suffix of the analyzed token;
- dictionary look–up features, marking whether the analyzed token is part of an entity name occurring in one of the "People", "Organization" and "Geo–political" gazetteers. The "Organizations" and "People" gazetteers were automatically generated from Wikipedia[1], while the "Geo–political" gazetteer was downloaded from the United Nations Economic Commission for Europe web site [2].
- contextual features, referring to orthographic, linguistic and dictionary look–up features of the context words of the analyzed token(s);

[1] http://en.wikipedia.org/wiki/Category:Lists_of_organizations and http://en.wikipedia.org/wiki/Lists_of_people

[2] http://www.unece.org/cefact/locode/welcome.html

- non–local features: starting from the assumption that identical tokens in the same document very likely refer to the same entity, we used this feature type to exploit previous label assignments to predict the label for the token being analysed. The effectiveness of these *history* features is corroborated by what observed in Ratinov and Roth (2009), where the authors affirm that the named entities in the beginning of the documents tend to be more easily identifiable.

For Italian, ItaliaNLP NER is trained on I-CAB (Italian Content Annotation Treebank) (Magnini et al., 2006), the dataset used in the NER Task at EVALITA 2009 (Speranza, 2009) including four standard named entity tags, i.e. Person, Organization, Location and Geo–political entity. For English, the system is trained on the REUTERS corpus (Lewis et al., 2004), the dataset used in the CoNLL–2003 shared task focusing on language–independent named entity recognition (Sang and Meulder, 2003) and including four types of named entities: Persons, Locations, Organizations and Miscellaneous entities, i.e. that do not belong to the previous three groups. The results obtained for the two languages are in line with the state of the art when compared with the systems that participated to the EVALITA and CoNLL–2003 shared tasks.
At the end of this extraction step, the user can download the list of extracted named entities with associated classification, absolute and relative frequency; relative frequency information is reported with respect to both all classes and the assigned class. As in the case of domain–entities, T2K$^2$ allows downloading the input corpora in CoNLL format where the start–, internal– and final–tokens of the sequences representing each named entity are annotated using the IOB labeling format.
The list of extracted named entities is used to carry out *Named Entity indexing*, which results in an index in tabular format where each column contains the following information types:

- the indexed named entity;
- the identifiers of the document and of the sentence in which the named entity is mentioned as well as of the start–token;
- the length of the named entity calculated in terms of tokens;
- the semantic class of the named entity;
- the TF/IDF score (Salton and Buckley, 1988) measuring the relevance of the indexed named entity with respect to the specific document.

The user can download the index and use it as input for external tools.

### 2.3. Knowledge Organization and Knowledge Graph Construction

In T2K$^2$ extracted knowledge is organized at different levels. As in the previous T2K version, the extracted domain–specific terms are organized into fragments of taxonomical chains, grouping terms which share the semantic head (e.g. *health research*, *international research*, *cancer research* are classified as hyponyms of the more general term *research*). Other types of organization are currently being investigated: for instance, terms are grouped on the basis of shared modifiers defining their scope (e.g. *research projects*, *research excellence*, *research infrastructure*, *research results* where *research* is the shared modifier).
The extracted domain–specific entities and named entities are also organized in a knowledge graph where the arcs linking the entities correspond to relations extracted from the analyzed corpus. Currently, in T2K$^2$ two different types of relations can be extracted: co–occurrence and similarity relations. The former is the case of relations holding between entity mentions co–occurring within the same context. To this end, different types of contexts can be selected, ranging from the whole document to the sentence or a span defined on the basis of a given number of tokens: in the future, a more linguistically oriented notion of context will also be used. The resulting knowledge graph is weighted with respect to the frequency of occurrence or using the log–likelihood metric for binomial distributions as defined in (Dunning, 1993). This notion of *context* makes T2K$^2$ robust with respect to non–canonical input, such as e.g. the language of social media, microblogs, etc., and enables the system to capture both inter– and intra–sentence relations. The risk that the system may also capture not relevant relations holding between entities casually co–occurring within the same context is limited by the fact that T2K$^2$ is meant to be used on big corpora where the statistics is supposed to highly reduce the impact of these events.
On the other hand, similarity relations are computed on the basis of the amount of contexts shared by the same entity mentions. This kind of relations is weighted on the basis of the cosine similarity between the entity context vectors. The components of each vector contain the association strength (computed in terms of log–likelihood) between the considered entity and its context entities.
T2K$^2$ allows visualizing the whole knowledge graph or sub–graphs created by selecting a sub–set of entities. The knowledge graph can also be filtered on the basis of relations' weight, where the relations are weighted according to the their frequency, log–likelihood or cosine similarity, and on the basis of the entities' frequency. The resulting graph can be used in a number of different graph mining analysis, such as e.g. extraction of all relations involving a given entity, or extraction of sub–graphs containing a given entity, or extraction of relations shared by two or more entities, or extraction of entities which share the same relations.
The input corpus is indexed afterwards on the basis of extracted relations and entities. Accordingly, T2K$^2$ allows visualizing the text spans of the input corpus mentioning the extracted entity and/or relation.

### 2.4. Output Examples

In this section, we exemplify the output of the different information extraction steps described in Sections 2.2. and 2.3., using as acquisition corpus a collection of educational textbooks for high school on Italian art history.[3]

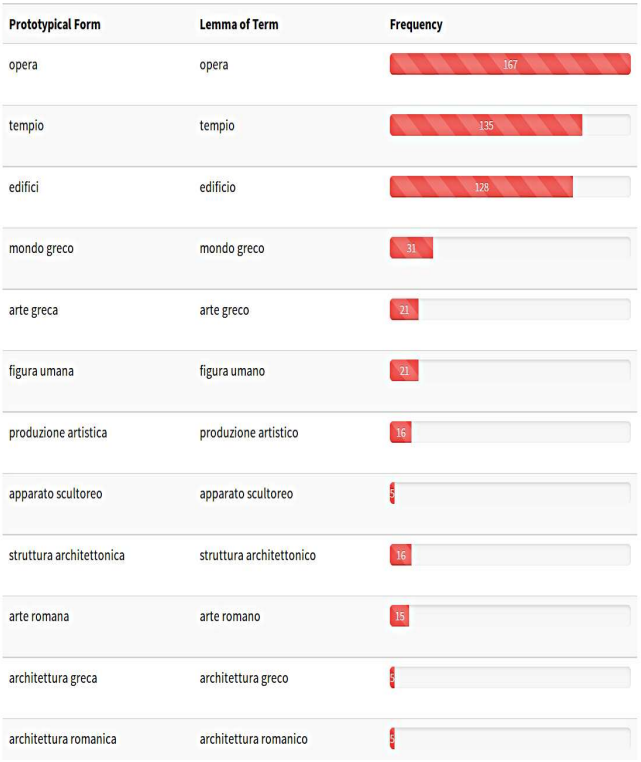
[3]The corpus was built in the framework of the *iSLe – intelligent Semantic Liquid eBook*, http://www.progettoisle.it/il-

Figure 2 exemplifies the output of the term extraction step, where terms are ordered on the basis of their domain relevance within the input corpus: for each term the absolute frequency is also reported. Figure 3 shows a fragment of the index. The terms are alphabetically ordered: each line refers to a single occurrence of a given term within a document. For example, the multi–word term *architettura greca* 'Greek architecture' occurs in five lines, four times in the *Storia dell'Arte 4* volume and one time in the *Storia dell'Arte 5* volume. The column reporting the $TF/IDF$ score shows that this entity is more relevant in the *Storia dell'Arte 4* volume, with a score of 0.00704 (against 0.0046 for the other volume).

In Figure 4 the domain–specific terms are organized into fragments of taxonomical chains, for example *architettura greca* 'Greek architecture', *architettura longobarda* 'Longobard architecture', *architettura romana* 'Roman architecture', *architettura romanica* 'Romanesque architecture' share the same semantic head *architettura* 'architecture'.

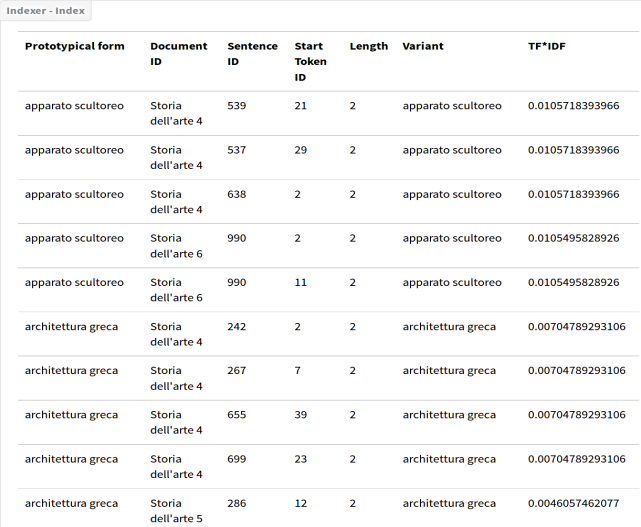
A short list of extracted named entities (ordered by decreasing frequency) is reported in Figure 5, with each line containing a single entity mention.

Figure 6 shows the visualization of the Knowledge graph, where domain–specific terms and named entities are represented by nodes, and the relations linking them by arcs. The different entity types are marked by different colors, whereas the size of nodes is proportional to the frequency of the corresponding entity. T2K$^2$ allows the user to create graph views focusing on specific relation types (i.e. co-occurrence or similarity) and using different weight measures (i.e. frequency or log-likelihood). As described in Section 2.3., T2K$^2$ permits the visualization of sub–graphs of selected entities and/or relations: this is exemplified in Figure 7 showing the sub–graph of the *Person* (*PER*) *Giotto*, while Figure 8 reports the relations shared by two named entities, i.e. *Giotto* and *Cimabue*. Figure 9 reports a T2K$^2$ screenshot with a ternary relation linking *Giotto* and *Cimabue* with the domain specific entity *basilica superiore* 'superior basilica'. The textual box reports the text span mentioning the extracted relation: it is said that *Giotto* and *Cimabue* decorated the superior basilica of St. Francesco in Assisi in different periods.



| Prototypical Form | Lemma of Term | Frequency |
|---|---|---|
| opera | opera | 167 |
| tempio | tempio | 135 |
| edifici | edificio | 128 |
| mondo greco | mondo greco | 31 |
| arte greca | arte greco | 21 |
| figura umana | figura umano | 21 |
| produzione artistica | produzione artistico | 16 |
| apparato scultoreo | apparato scultoreo | 5 |
| struttura architettonica | struttura architettonico | 16 |
| arte romana | arte romano | 15 |
| architettura greca | architettura greco | 5 |
| architettura romanica | architettura romanico | 4 |

Figure 2: An excerpt of the glossary with extracted domain–specific terms.



Indexer - Index

| Prototypical form | Document ID | Sentence ID | Start Token ID | Length | Variant | TF*IDF |
|---|---|---|---|---|---|---|
| apparato scultoreo | Storia dell'arte 4 | 539 | 21 | 2 | apparato scultoreo | 0.0105718393966 |
| apparato scultoreo | Storia dell'arte 4 | 537 | 29 | 2 | apparato scultoreo | 0.0105718393966 |
| apparato scultoreo | Storia dell'arte 4 | 638 | 2 | 2 | apparato scultoreo | 0.0105718393966 |
| apparato scultoreo | Storia dell'arte 6 | 990 | 2 | 2 | apparato scultoreo | 0.0105495828926 |
| apparato scultoreo | Storia dell'arte 6 | 990 | 11 | 2 | apparato scultoreo | 0.0105495828926 |
| architettura greca | Storia dell'arte 4 | 242 | 2 | 2 | architettura greca | 0.00704789293106 |
| architettura greca | Storia dell'arte 4 | 267 | 7 | 2 | architettura greca | 0.00704789293106 |
| architettura greca | Storia dell'arte 4 | 655 | 39 | 2 | architettura greca | 0.00704789293106 |
| architettura greca | Storia dell'arte 4 | 699 | 23 | 2 | architettura greca | 0.00704789293106 |
| architettura greca | Storia dell'arte 5 | 286 | 12 | 2 | architettura greca | 0.0046057462077 |

Figure 3: An excerpt of the term index.

## 3. Ongoing Applications

T2K$^2$ is currently being tested and specialized in the framework of different ongoing projects aimed at extracting and organizing knowledge from different Italian domain–specific corpora:

- *Legal Text Mining: building semantic networks to support advanced queries in legal textual corpora (JURNET)*, a project funded by the Tuscany Region and aimed at accessing the knowledge contained in case law corpora;
- *iSLe – intelligent Semantic Liquid eBook*, a project funded by the Tuscany Region and aimed at i) developing an innovative software platform for digital educational publishing augmented with NLP-based functionalities for knowledge management and at ii) supporting authors during the creation of educational textbooks using the knowledge graph extracted from domain–specific corpora;
- *INMOTO: INformation and MObility for Tourism*, a national Italian project where T2K$^2$ is used to create an ontology of tourism supporting the writing of travel guides on the basis of the knowledge extracted from web touristic sites.

progetto-isle/.

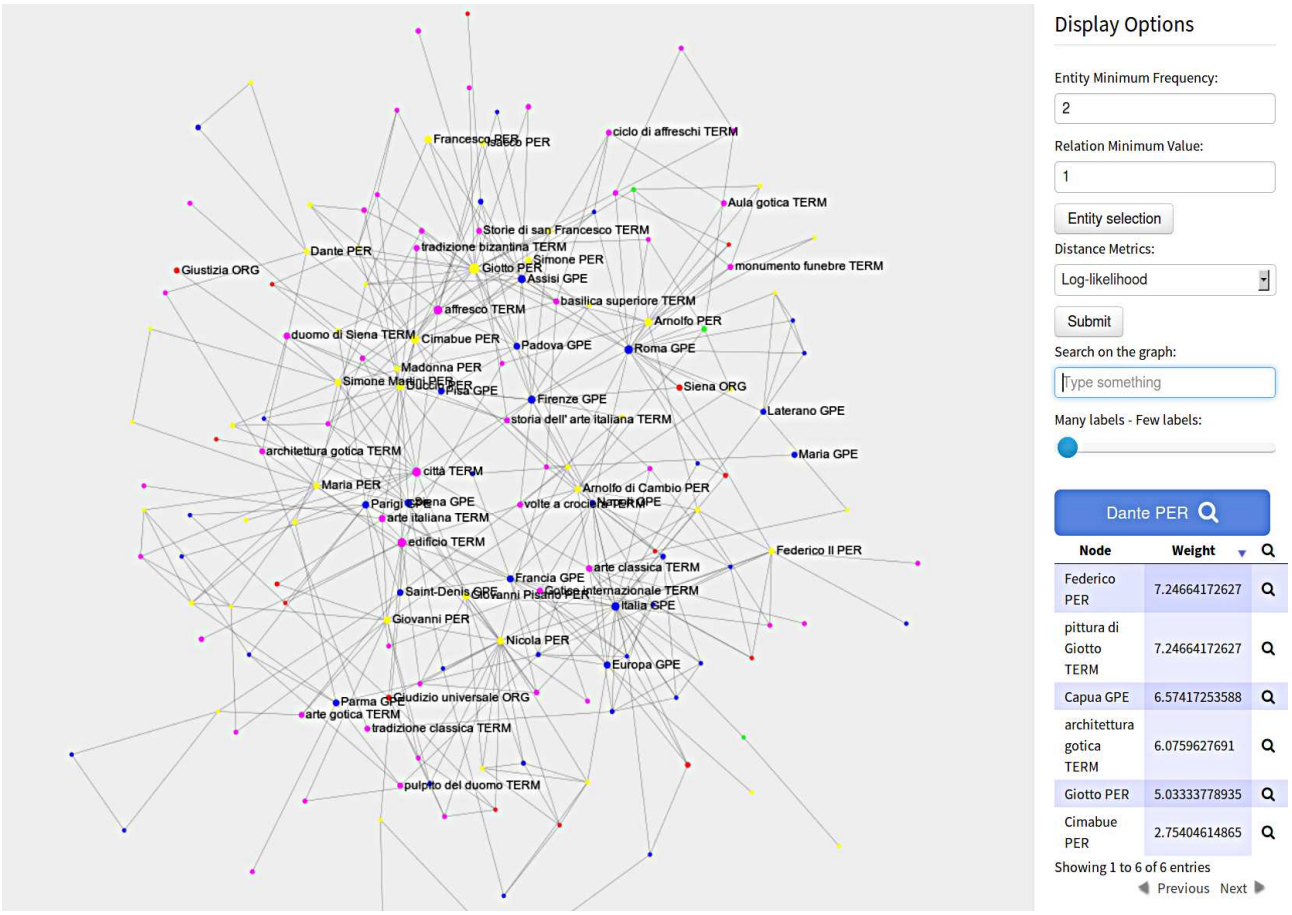



Figure 6: A screenshot of the extracted knowledge graph.

T2K$^2$ is also being tested on English texts in studies focused on the analysis of technical documents such as patents and system requirements with the final aim of helping engineers to discover relevant concepts and relations:

- The contrastive analysis performed by T2K$^2$ was used by Ferrari et al. (2013) to extract domain–specific entities from brochures. Identifying and comparing the features provided by the other vendors might greatly help during the market analysis. However, mining common and variant features of from the publicly available documents of the competitors is a time consuming and error–prone task. In this scenario, they used the domain specific entities extracted by T2K$^2$ to perform commonality and variability mining. A case study was carried out to qualitatively evaluate the approach in the metro systems domain: the proposed approach demonstrated its expected time–effectiveness.

- T2K$^2$ was used to detect and extract information about the functions, the physical behaviour and the states of the system directly from patents (Fantoni et al. (2013)). In the field of automatic patent analysis a number of domain specific ontologies have been successfully proposed in a variety of research projects, mostly focusing on upper level concepts hand–crafted by domain experts. However, realistically large knowledge–based applications need comprehensive ontologies that should be continuously updated. To overcome this problem, they used the terminology extraction approach developed by T2K$^2$ on automatically crawling patents belonging to specified patent classes and subclasses for automatically acquiring a collection of entities characterizing the analized domain. Acquired entities were used to create a knowledge base representing a key ingredient in the proposed approach to disambiguate, gather, select and organize information from technical documents.

- Ferrari et al. (2014) applied T2K$^2$ on system requirements specifications. They proposed two metrics that take into account the relevant terms of the input documents, and the relevant relationships among terms to measure and improve the completeness of the requirements with respect to the input documents of the requirements definition phase, such as preliminary specifications, transcripts of meetings with the customers, etc. They use T2K$^2$ within their system named Completeness Assistant for Requirements (CAR) to extract relevant concepts and relations mentioned in the input documents.

## 4. Further Directions of Research

Current lines of research and development cover different areas. First, we are working towards the definition of innovative methods to extract and classify domain–specific


Figure 4: An excerpt of extracted taxonomical chains.

Name Entity Extraction

Show 50 entries Search:

| Entity | Class | Frequency | Frequency (%) | Class Frequency (%) |
|---|---|---|---|---|
| Giotto | Person | 66.0 | 7.61 | 15.64 |
| Cimabue | Person | 26.0 | 3.00 | 6.16 |
| Roma | GeoPolitical Entity | 21.0 | 2.42 | 7.02 |
| Duccio | Person | 21.0 | 2.42 | 4.98 |
| Arnolfo | Person | 18.0 | 2.08 | 4.27 |
| Italia | GeoPolitical Entity | 18.0 | 2.08 | 6.02 |
| Assisi | GeoPolitical Entity | 17.0 | 1.96 | 5.69 |
| Francesco | Person | 15.0 | 1.73 | 3.55 |
| Siena | GeoPolitical Entity | 15.0 | 1.73 | 5.02 |
| Firenze | GeoPolitical Entity | 14.0 | 1.61 | 4.68 |
| Nicola | Person | 12.0 | 1.38 | 2.84 |
| Giovanni | Person | 12.0 | 1.38 | 2.84 |
| Giovanni Pisano | Person | 10.0 | 1.15 | 2.37 |
| Parigi | GeoPolitical Entity | 10.0 | 1.15 | 3.34 |
| Francia | GeoPolitical Entity | 9.0 | 1.04 | 3.01 |
| Simone Martini | Person | 9.0 | 1.04 | 2.13 |

Figure 5: An excerpt of extracted Named Entities.

named entities; we are currently developing methods for clustering and labeling domain specific entities and for indexing the input document collection on the basis of them. For what concerns relation extraction, we are implementing linguistically driven methods relying on the syntactic structure of the text: T2K$^2$ will exploit linguistic information both to automatically extract domain–specific relations and to allow the user to define syntactic patterns conveying them. Last but not least, we are also devising methods for graph mining: in particular, we are integrating in T2K$^2$ algorithms for frequent subgraph mining, graph clustering and community discovery to detect graph areas that are thematically homogeneous as well as algorithms for detecting hub nodes in order to identify relevant topics of the input corpus.

## 5. References


G. Attardi and F. Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL–HLT 2009*, pages 261–264.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of CoNLL-X 2006*, pages 166–170.

F. Bonin, F. Dell'Orletta, G. Venturi, and S. Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 3222–3229.

M.T. Cabrè. 1999. *The terminology. Theory, methods and applications*. John Benjamins Publishing Company.

A. Caruso, A. Folino, F. Parisi, and R. Trunfio. 2014. A statistical method for minimum corpus size determination. In *Proceedings of JADT2014 (Journées Internationales d'Analyse des Données Textuelles), Paris, 3-6 Juin 2014*.

C. Chang and C. Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

F. Dell'Orletta, A. Lenci, S. Marchi, S. Montemagni, V. Pirrelli, and G. Venturi. 2008. Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio. *AIDA Informazioni, Proceedings of the National Conference Ass.I.Term I-TerAnDo*, AIDA, n. 1-2/20085:185–206.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2013. Linguistic profiling of texts across textual genre and readability level. an exploratory study on italian fictional prose. In *Proceedings of RANLP-2013*, pages 189–197.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of Evalita'09*.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

G. Fantoni, R. Apreda, F. Dell'Orletta, and M. Monge. 2013. Automatic extraction of function-behaviour-state information from patents. *Journal of Advanced Engineering Informatic*.

A. Ferrari, G. Oronzo Spagnolo, and F. Dell'Orletta. 2013. Mining commonalities and variabilities from natural language documents. In *Proceedings of the 17th International Software Product Line Conference (SPLC-2013)*.

A. Ferrari, F. Dell'Orletta, G. Oronzo Spagnolo, and S. Gnesi. 2014. Measuring and improving the complete-

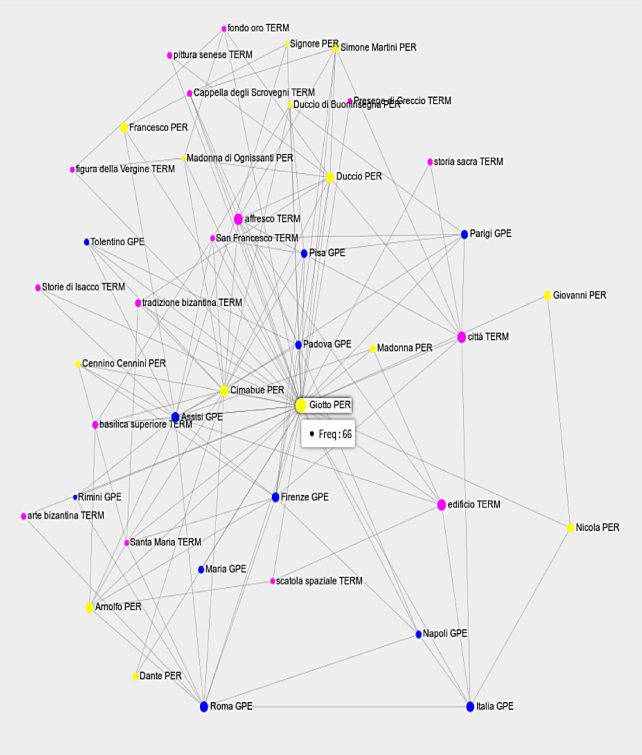



Figure 7: The sub–graph of *Person* (PER) **Giotto**.

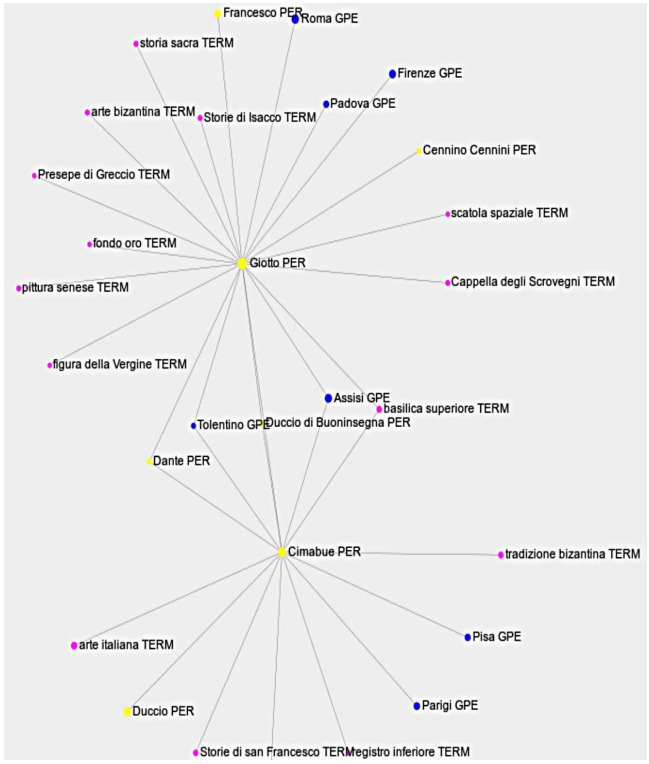



Figure 8: The relations involving the two named entities **Giotto** and **Cimabue**.

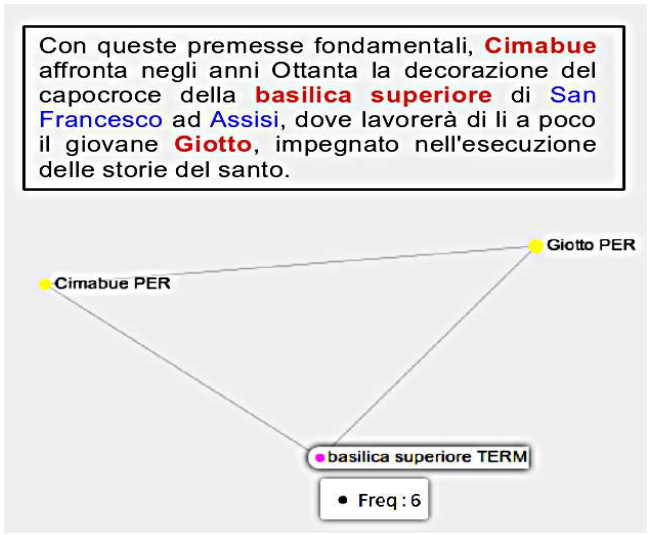



Figure 9: The ternary relation linking *Giotto* and *Cimabue* with *basilica superiore* 'superior basilica'.

ness of natural language requirements. In *Proceedings of the 20th International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ 2014)*.

K. Frantzi and S. Ananiadou. 1999. The c-value / nc value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi. 2009. Ontology learning from italian legal texts. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web*, pages 75–94, Amsterdam, The Netherlands. IOS Press.

D.D. Lewis, Y. Yang, T. Rose, and F. Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, and R. Sprugnoli. 2006. I-cab: the italian content annotation bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

G. Nenadic, S. Ananiadou, and J. McNaught. 2004. Enhancing automatic term recognition through term variatio. In *Proceedings of Coling 2004*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL 2007*, pages 915–932.

L.A. Ramshaw and M.P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, pages 82–94.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155.

G. Salton and C. Buckley. 1988. Term-weighing approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

E.F.T.K. Sang and F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

M. Speranza. 2009. The named entity recognition task at evalita 2009. In *Proceedings of Evalita'09*.

H. van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of ACL 2004*, pages 200–207.