Two Decades of Terminology: European Framework Programmes Titles

Gabriella Pardelli*, Sara Goggi*, Silvia Giannini *, Stefania Biagioni *

*CNR, Istituto di Linguistica Computazionale, “Antonio Zampolli”, Italy
* CNR, Istituto di Scienza e Tecnologie dell’Informazione “A. Faedo”, Italy
[gabriella.pardelli, sara.goggi]@ilc.cnr.it, [silvia.giannini, stefania.biagioni]@isti.cnr.it

Abstract

This work analyses a corpus made of the titles of research projects belonging to the last four European Commission Framework Programmes (FP4, FP5, FP6, FP7) during a time span of nearly two decades (1994-2012). The starting point is the idea of creating a corpus of titles which would constitute a terminological niche, a sort of “cluster map” offering an overall vision on the terms used and the links between them. Moreover, by performing a terminological comparison over a period of time it is possible to trace the presence of obsolete words in outdated research areas as well as of neologisms in the most recent fields. Within this scenario, the minimal purpose is to build a corpus of titles of European projects belonging to the several Framework Programmes in order to obtain a terminological mapping of relevant words in the various research areas: particularly significant would be those terms spread across different domains or those extremely tied to a specific domain. A term could actually be found in many fields and being able to acknowledge and retrieve this cross-presence means being able to linking those different domains by means of a process of terminological mapping.

Keywords: terminology extraction; natural language processing; terminological comparison

1. Introduction and background

This task of automatic extraction enabled the authors to undertake a diachronic linguistic survey over the language used for the titles of projects belonging to European Framework Programmes (FP): in this period of twenty years taken into account (1994-2012), the terminology adopted in the various fields is continuously evolving thus revealing its dynamic nature. “When we read the articles or papers of a particular domain, we can recognise some lexical items in the texts as technical terms. In a domain where new knowledge is generated, new terms are constantly created to fulfil the needs of the domain, while others become obsolete. In addition, existing terms may undergo changes of meaning…” (Kageura K., 1999). The starting point of this research dates back to a study of 2013 which was focusing on the terminology adopted within the ICT¹ domain: more specifically, a map of terms distributed over the Seventh Framework Programme (FP7) project titles was modelled. The first results – published in 2015 – encouraged us to plan further works: “As future work we are evaluating the possibility of extracting topic information from all the […] FP titles for designing an automatic process of extracting obsolete and new words in order to identify old and most recent fields of European research in the time span 1994-2013” (Goggi et al., 2015). The present study above all investigates the use of terms in FP4, FP5, FP6, FP7 corpora. “… corpora used to extract terminological units can be further investigated to find semantic and conceptual information on terms or to represent conceptual relationships between terms. (Bourigault D. et al., 2001). From this survey specialised terms comes to light, playing the essential role of conveying the semantic meaning thus becoming identifiers and information retrieval anchors for the respective scientific domains of the several Framework Programmes. It has been especially difficult to identify the adequate criteria for detecting and classifying the most significant candidate terms: “Candidate terms are defined as noun phrases satisfying a restricted set of part-of-speech patterns and are classified by their length, i.e. the number of noun, adjectives, verbs and participles…” (Blank I., 2000). The requirement was to retrieve the terms in the various fields and verify their tendency over the time span 1994-2012; the starting point has been the analysis and the mapping of the terms of each Framework while subsequently the four mappings have been compared. This method – described in detail in paragraph 2.4 – has permitted to recognize the following types of terms: Terms recurring in the four frameworks [permanent terms - PT], Terms belonging to one framework only [hapax terms - HT], Terms belonging to consecutive frameworks [distributed terms - DT]; Terms belonging to non-consecutive frameworks [irregular terms - IT].

So the focus of this work consists in creating a map of PT, HT, DT and IT terms.

2. Terminology extraction and methodology

The analysis on the 63,377 titles has been carried out by using the new version of the software named Text-to-Knowledge (T2K) (Dell’Orletta et al., 2014) developed by the Italian NLP Laboratory of the Institute of Computational Linguistics “Antonio Zampolli” of National Research Council, (CNR)². T2K is a platform developed in order to extract knowledge from texts: it is based both on supervised and unsupervised statistical algorithms and is able to automatically index texts by using

¹ ICT stands for Information and Communications Technology

² www.iltalianlp.it. (last accessed on January 15, 2016).
the extracted domain terms as well as create semantic relations among the entities retrieved. In short, this is a hybrid system for knowledge extraction and document indexing that combines Natural Language Processing, (NLP) technologies and statistical techniques. The tool performs three levels of automatic extraction: at the first level, it extracts a list of single and multi-word terms ordered by relevance with respect to the context; at the second one it creates fragments of taxonomical chains and at the last one clusters of related terms. The following paragraph is split up in four sections:

1. downloading of data (titles of projects within the four FPs) from the CORDIS website;
2. data processing using T2K;
3. terminological analysis;
4. terminological comparison and results

### 2.1. Downloading

Data has been downloaded from CORDIS - European Union Open Data Portal: “Extraction of all FP4, FP5, FP6, FP7 projects from the CORDIS public repository, containing the following fields: RCN (CORDIS record number), Reference (grant agreement number), Acronym, Title, Start Date, End Date, Total Cost, EC Contribution, Programme (Acronym), Sub-programme (Activity Code), Funding Scheme (contract type), Subject (SIC codes), Keywords, Coordinator country (code), Participant countries (comma-separated code), Objectives”.

### 2.2 Data processing

The first level of extraction shows the glossary of each Framework Programme: the structure of the glossary envisages the presence of the prototypical form of each term, of the lemma and the number of occurrences (frequency) (Table 1). The system returns a list of words ordered by the relevance of each single term within the context. This list has then been re-ordered by the frequency of terms while the graphic presentation remained the original one.

<table>
<thead>
<tr>
<th>Prototypical_form</th>
<th>Lemma_of_term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development</td>
<td>development</td>
<td>1129</td>
</tr>
<tr>
<td>applications</td>
<td>application</td>
<td>332</td>
</tr>
<tr>
<td>Europe</td>
<td>europe</td>
<td>224</td>
</tr>
<tr>
<td>production</td>
<td>production</td>
<td>197</td>
</tr>
<tr>
<td>quality</td>
<td>quality</td>
<td>184</td>
</tr>
<tr>
<td>technology</td>
<td>technology</td>
<td>175</td>
</tr>
<tr>
<td>software</td>
<td>software</td>
<td>170</td>
</tr>
<tr>
<td>water</td>
<td>water</td>
<td>155</td>
</tr>
<tr>
<td>research</td>
<td>research</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 1: FP4 glossary

The second level of extraction (Table 2) creates the taxonomic chains: the list which comes out first of all shows all the head-based terms, those that “…share the same semantic head…” and afterwards the modifier-based ones, which are those “…grouped on the basis of shared modifiers defining their scope…” (Dell’Orletta et al., 2014).

### 2.3 Terminological analysis

The analysis starts with the process of identification of the terms of high, medium and low frequency within the glossaries provided by the first level of extraction. For frequency segment of vocabularies we mean the terms co-occurring with it (ordered in a decreasing way); the second (#Likelihood) reveals the associative strength of a term, that is its “attitude” to associate to other terms and consequently its own degree of specificity; the third (#Cosine similarity) states the similarity between the two terms by measuring the number of common entities (Table 3).

<table>
<thead>
<tr>
<th>#Method frequency</th>
<th># FORMAT</th>
<th>#</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>entityAfrequency(entityA)</td>
<td>entityBfrequency (entityA,entityB)</td>
<td>entityCfrequency (entityA,entityC)</td>
<td></td>
</tr>
<tr>
<td>reduction 37</td>
<td>Development 4</td>
<td>generation 2</td>
<td></td>
</tr>
<tr>
<td>computational methods 4</td>
<td>Development 1</td>
<td>data 1</td>
<td></td>
</tr>
<tr>
<td>simulation 53</td>
<td>environment 5</td>
<td>industrial 3</td>
<td></td>
</tr>
<tr>
<td>susceptibility genes 5</td>
<td>identification 2</td>
<td>European 1</td>
<td></td>
</tr>
<tr>
<td>process management 3</td>
<td>engineering 1</td>
<td>integration 1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: FP4 clusters of related terms

---

3 For a complete description of this tool, please see the references section and visit www.italianlp.it.
4 The expression “term” includes both single and multi-words.
5 “CORDIS is the European Commission’s primary public repository and portal to disseminate information on all EU-funded research projects and their results in the broadest sense” http://cordis.europa.eu/home_en.html (last accessed on January 15, 2016).
the high and medium frequencies is placed immediately above the first parity, that is, the first pair of words that occur the same number of times. To determine the freqmin segment and separate it from the mid-range, it is necessary to start from the bottom, i.e. from the hapax, and consider the first gap in the consecutive number of increasing occurrences. After having organised the terms, it results that the highest percentage of terms is to be found in the lowest frequency segment: this applies to all Frameworks. The FP7 glossary stands out for the substantial amount of terms in the highest segment while the medium segment can be allocated to FP5.

More generally, there is a massive use of terms with a very wide semantic content: for instance, in the high segment the terms “development”, “applications”, “research”, “social” can be detected but there are also – to a smaller extent – more specific terms such as “gene”, “food”, “disease”, “quantum”, “carbon” or terms connected with specific FP actions, i.e. “short-term”, “short term scientific mission”, “cost action”. The survey on fragments of taxonomic chains and co-occurrences sheds light on the usage of specific terms and defines their scope within the several research areas in the various Frameworks. Consider the example of the noun *development*, which is always in the first position for overall amount of occurrences in the project titles of each FP: it occurs both as a “head-based” and a “modifier-based” term. The taxonomic chains created at the second level of extraction report about the link of this noun with European research on rural or urban areas (rural development, urban development), industry (industrial development), technology (software development) rather than with social development. There are also connections with research on health and medicine (cancer development, brain development, lymphocyte development, vaccine development). The list of co-occurrences shows that in FP4 the term *development* is associated with the highest frequency to the noun *technology* (42); to follow, terms such as *production* (34), *components* (26), *techniques* (22), *industrial, materials* and *water* (18) and so on up to terms with which *development* is rarely associated - such as *transport* (3), *environmental impact* (2), *optical fibre* (1).

The #Likelihood and #Cosine similarity methods calculate the similarity between two terms by always using values between 0 and +1: when the value is 0 the two are completely different, while when it is 1 the words are identical; and a value around 0.8 indicates a high degree of similarity. In particular, the #Likelihood method (329 occurrences) measures the associative strength of terms thus enhancing the exclusive links with other terms, while the #Cosine similarity method (934 occurrences) analyses the similarity on the basis of the occurrences they have in common. For instance, by decreasingly ordering the cluster of the word *development* obtained by the Likelihood method, it can be noted that the term mainly relates with words like *components, techniques, materials*. The lowest values belong to terms such as *land, ship, digital* and *high precision*.

### 2.4 Terminological comparison

The terms processed by T2K have been further processed and compared in order to monitor the evolution of the terminological flow and indicate the resulting lexical trend within the research fields of the EU projects.

“A single concept may be represented in different way in text – for example, as misspellings or abbreviations” (Park, Byrd and Boguraev, 2002). By applying this concept to the corpus of European projects titles, the following illustrative results are obtained:

a) Symbolic variants such as:
- Pharmacokinetic/Pharmacodynamic > Pharmacokinetic-pharmacodynamic;
- Insulin dependent diabetes mellitus (IDDM) > insulin-dependent diabetes mellitus;
- Re-engineering > re engineering;
- Co-ordination action > Coordination Action

b) Compounding variants such as:
- nano-structured materials > nanostructured materials.
- Wave energy device: broadband seapower energy recovery buoy > Wave energy device - broad band seapower energy recovery buoy.

In a) and b) there are variations in the title of the same project: “broadband” - single term (1997) uses the colon for separating the subheading: ‘broad band’. two terms (1999) uses the dash for separating the subheading. [PROJECT] Wave energy device - broad band seapower energy recovery buoy

Ref.: JOR3987026
Start date: 1999-01-01, End date: 2001-02-28
Programme: FP4-NNE-JOULE C
Record Number: 47414
Last updated on: 2002-04-19

[PROJECT] Wave energy device : broadband seapower energy recovery buoy
Ref.: JOR3971004
Start date: 1997-07-22, End date: 1998-03-31
Programme: FP4-NNE-JOULE C
Record Number: 40344
Last updated on: 1998-03-16

c) Uppercase/lowercase variants: the same terms are written in a different way:

Free-Electron > (FREE ELECTRON) > free electron;
DNA Damage Response > DNA damage response;
Knowledge Management > knowledge management

d) Formulas, initials and acronyms, abbreviations:

The need to express a scientific concept with a few signs is solved with the creation of acronyms (initials of words which together assemble a brand new word):

- ‘CO2’: chemical formula of carbon dioxide
- ‘3D Digital’: abbreviation of dimension > 3-dimensional
- ‘EURAXESS’: name of a pan-European electronic gateway
- ‘RNA’: stands for ribonucleic acid, a biological macromolecule
- ‘ICT’: Information and Communication Technology
- ‘ES cells’: Embryonic Stem cells
- ‘NK cells’: Natural Killer cells
- ‘DNA’: Deoxyribonucleic Acid
- ‘SME’: European Monetary System
- ‘HIV’: Human Immunodeficiency Virus
- ‘IST’: Information Society Technologies
- ‘LHC’: Large Hadron Collider

More specifically the following four sets have been taken...
into consideration as a sub-corpus (C1, C2, C3; C4):

2.4.1. C1. Terms recurring in all the four frameworks [permanent terms - PT]


Air: ACTIVITY_AREA FP7: a) ICT for environmental management and energy efficiency, b) Nanotechnology-based sensors for environmental monitoring, Mobilising environmental knowledge for policy and society, c) Marie-Curie Action: "Intra-European fellowships for career development" (Life Sciences), d) Design and optimisation of locally reacting acoustic material (Active noise control of turbomachinery);

Air: ACTIVITY_AREA FP6: a) Climate change: the Kyoto protocol and beyond; b) Safety of road and air transports; Co-operative Research (all areas of science and technology); c) Innovative air traffic management research; d) Quality of life issues relating to handicapped/disabled people;

Air: ACTIVITY_AREA FP5: a) Key action Economic and Efficient Energy for a Competitive Europe; b) Key Action Innovative Products, Processes and Organisation; c) Key action Environment and Health, d) Key Action New Perspectives in Aeronautics;

Air: ACTIVITY_AREA FP4: a) Tropospheric physics and chemistry, b) Measurement and testing for the control of production, c) Microsystems, d) Risks to human health.


2.4.2. C2. Terms belonging to one framework only [hapax terms - HT]

The HT is constituted by terms belonging to just one Framework which therefore stand for the specific research activities carried out in that programme. Just to make a few examples: in FP4, the adjective ‘ceramic’ is involved in taxonomic chains such as ‘ceramic material’, ‘ceramic matrix composite’, ‘ceramic tile’; the noun ‘transport’ is combined with another noun thus forming the pairs ‘transport industry’, ‘transport policy’ and the trio ‘vertical ozone transport’; the noun ‘water’ is coupled both with a verb describing its quality – ‘polluted water’ – and with other nouns as ‘water industry’7 and ‘water Column’ (a technical term of the oceanographic domain).

As an example, the triple ‘spinal cord injury’, dealing with domains such as biotechnology, life sciences, medicine and health, is recorded in FP6 only, thus defining its temporal time span.

2.4.3. C3. Terms belonging to consecutive frameworks [distributed terms - DT]

This set is made up of terms recurring in consecutive frameworks (e.g. FP4-FP5 or FP5-FP6 or FP5-FP6-FP7). Amongst the terms present in FP4-FP5 only, there are expressions like ‘social integration’ used just for the period from 1994 to 2002 and then abandoned. ‘Social integration’ deals with themes such as social disparity, economic growth, instruction, assessment of the social effect of the integration politics and so on; the reason for its disappearance from the project titles is to be found in the (at least partial) dropping of these specific contents from the European research.

The words occurring only in FP5 and FP6 are forty-five: an example is given by ‘molecular epidemiology’ which represents research about monitoring infectious diseases and providing basic tools for genomics. In this case, its absence in FP4 indicates the simultaneous lack of research in this area while in FP7 the study of genomes has a different connotation which stands for a different tendency (e.g. ‘genome regulation’, ‘genome stability’, ‘comparative genomics’…). Another example can be ‘cystic fibrosis’ which is clearly connected to research on genetic diseases and more specifically to co-ordinated actions against that particular disease at an European level.

Also the term ‘multimedia’ appears in FP4 and FP5 only: in FP4 is found in the activity area Evaluation of the inter-relations between short and medium-term needs and socio-economic changes and new scientific and technological developments; in FP5 is spread over the following three areas: Information Processing, Information Systems; Information, Media; Life Sciences.

2.4.4. C4. Terms belonging to non-consecutive frameworks [Irregular Terms - IR]

The terms belonging to this group are used in frameworks which are chronologically far: the first one (1994-1998) and the last one (2007-2012). Amongst the terms recurring in FP4 and FP7 only, there is the pair ‘welfare state’ which concerns the problems regarding migration and multicultural societies; and the noun ‘child’ appears in fourteen titles of FP7 research projects dealing with important themes like vaccination, technologies for children road safety, obesity and educational tools. Moreover, in the HEALTH area the EC supports the creation of an European multilingual platform channelling all research related to children health.

3. Results

This comparative analysis has highlighted some aspects which can easily be summarized by observing some classes

---

of adjectives: for instance, origin adjectives have a high frequency in every Programme:


The same adjective can also be paired with other types of nouns thus representing the efforts for the future of the EU: ‘European construction’, ‘European development’, ‘European perspective’. Finally, when the adjective ‘European’ is linked to specific domain terms, it represents the most significant research fields of the Union: ‘European cancer’, ‘European climate’, ‘European cultural heritage’, ‘European fishery’, ‘European library’, ‘European leather’, ‘European plant’, ‘European textile industry’.

By means of samplings, it has been possible to monitor the terminological flow in a diachronic perspective: over these twenty years the lexicon of the several research fields is continuously evolving and disclosing its dynamic nature, thus contributing to a deeper knowing of the EU research areas.

By adopting a diachronic point of view, some terms have been pointed out as obsolete while others emerged as very up-to-date: the latter are those chosen for assembling the latest titles in the same discipline or even for labeling emerging fields of knowledge.

5. Acknowledgements

The authors would like to thank Felice Delli’Orletta, researcher of the Institute of Computational Linguistics “Antonio Zampolli” (ILC) of the National Research Council for his support in the data processing performed by the T2K2 system.

6. Bibliographical References


4. Conclusion

This survey on the results of the information extraction process performed by the T2K tool has been a sort of linguistic path in the past and present of the terminology adopted in the titles of FP projects.

8 ICT stands for Information and Communication Technology.


