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performing a terminological comparison over a period of time it is possible to trace the presence of obsolete words in outdated research areas as well as of neologisms in the most recent fields.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 ICT[1] 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 efficient recognition process of 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[3].

The tool performs three levels of automatic extraction: at the first level, it extracts a list of single and multi-word terms[4] 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[5] 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 codes), Objectives"[6].

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.

Prototypical_form	Lemma_of_term	Frequency
Development	development	1129
applications	application	332
Europe	europe	224
production	production	197
quality	quality	184
technology	technology	175
software	software	170
water	water	155
research	research	150

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).

#Broader-Narrower Terms	
#Head based	#Modifier based
Development	Development
--> Efficient Development	-->development environment
--> Further development	--> development process
--> INDUSTRIAL DEVELOPMENT	
--> New developments	
--> application development	

Table 2: FP4 level of extraction

The third level of extraction shows the clusters of related terms by using different approaches: the first (#Broader-Narrower Terms) repeats the information about the term in the glossary and its frequency, indicating as well all 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).

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

Table 3: FP4 clusters of related terms

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 organisation of the words for decreasing frequencies, starting from the word with freqmax and coming to those with freqmin, usually with only one occurrence (hapax). The occurrences can be organised into groups: high, medium and low. In the high-frequency segment each word has a different number of occurrences. The limit between

---

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 impac*t (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]**

The set is made up of 227 terms. The distribution of nouns, adjectives and acronyms shows the steadiness in the usage of these terms during the time span 1994-2012 and in some cases indicates the main research tendencies within the European projects. The following nouns can be retrieved: 'air', 'brain', 'cancer', 'climate', 'drug', 'energy', 'food', 'gene', 'health', 'laser', 'life', 'plant', 'safety', 'soil', 'water'…; as for acronyms: 'dna' [Deoxyribonucleic Acid], 'laser' [Light Amplification by Stimulated Emission of Radiation], 'SMEs' ['Micro, small and medium-sized enterprises]. The term 'air', for instance, is used in the following 'Activity Areas' of the four Frameworks:

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)eSafety 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.

Among the various types of adjectives, 'new' tells us how old something is: 'new concept', 'new material', 'new method', 'new methodology', 'new process', 'new technique', 'new technology', 'new tool'. The synonym 'novel' is used only once combined with 'approach'.

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

---

[7] From Wikipedia: the water industry provides drinking water and wastewater services (including sewage treatment). Last accessed on January 15, 2016.

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

a) 'European': is often coupled with nouns with a peculiar semantic connotation (e.g. a group of people who gather and operate for a specific common purpose) such as 'European community', 'European consortium', 'European country', 'European network', 'European union' 'European association', 'European catchment', 'European commission', 'European co-ordination', 'European doctoral training', 'European group', 'European society'. 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'.

b) 'Mediterranean' - paired in order to outline portions of territories of different size like 'Mediterranean area' 'Mediterranean country', 'Mediterranean region'.

On the reverse, in FP7 adjectives like 'smart' and 'open' build up conceptually novel taxonomic chains which are not present in the other frameworks. The first one denotes a specific trend in ICT[8] research and pairs with the nouns 'city', 'energy', 'grid', 'technology'; the adjective 'open' creates taxonomic chains as 'open access', 'open data' and 'open source', mainly multi-word expressions (MWE) which introduce new paradigms the scientific literature. The peculiarity of FP7 research can be also inferred from the hapax term 'big data' which is used for the first time in the last Framework Programme for indicating the management of a considerable amount of heterogeneous information.

Other adjectives such as 'human', 'mental' and 'public' indicate the care of the Union for the wellness of citizens when they are coupled with nouns like 'health'; and yet, adjectives associated with domain terms return multi-word expressions like 'cultural heritage', 'information society', 'multiple sclerosis', 'renewable energy', 'rheumatoid arthritis', 'sustainable development'. These words recur in the four FPs and therefore confirm the continuum of the European research during the time span 1994-2012.

From the glossaries of each Framework, the first use of the multi-word expression 'semantic web', 'new economy', 'cultural heritage' is found in the titles of projects of the Fifth Framework Programme (1998-2002); 'open source' is retrieved in FP6 (2002-2007) and 'big data' in FP7 (2007-2012).

A curiosity: the first trace of the term 'internet' dates back to FP5 (1998-2002).

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.

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.

6. Bibliographical References

Blank I. (2000). Terminology extraction from parallel technical texts, in J. Veronis (Ed.), *Parallel Text Processing*, Kluwer Academic Publishers. Pages 237-252.

Bourigault D., Jacquemin C., L'Homme Marie C. (2001) Introduction, in Didier Bourigault, Christian Jacquemin, Marie-Claude L' Homme (Eds.), *Recent Advances in Computational Terminology*. Amsterdam, John Benjamins.

Dell'Orletta F., Venturi G., Cimino A., Montemagni S. (2014) "T2K: a System for Automatically Extracting and Organizing Knowledge from Texts". In *Proceedings of 9th edition of International Conference on Language Resources and Evaluation* (LREC 2014), 26-31 May, Reykjavik, Iceland. Paris, European Language Resources Association (ELRA). Pages 2062-2070.

Fellbaum C. (2014). Large-scale Lexicography in the Digital Age (2015). *International Journal of Lexicography*, Volume 27, number 4. Pages 378-395.

Grefenstette G. (1994). Semantic extraction, in Gregory Grefenstette, *Exploratios in automatic automatic thesaurus discovery* . Kluwer. Pages 7-32.

Grefenstette G. (1992). Sextant: Exploring Unexplored Contexts for Semantic Extraction from Syntactic Analysis, *30th Annual Meeting of the Association for Computational Linguistics*, Proceedings of the Conference. Association for Computational Linguistics. Pages 324-326.

Goggi S., Pardelli G., Sassi M., Giannini S., Biagioni S. (2015) A Terminological Survey on the Titles of the Seventh Framework Programme (FP7). In Leonel Ruiz Miyares, María Rosa Álvarez Silva y Alex Muñoz Alvarado (Eds.), *Fourteenth International Symposium on Comunicación Social: retos y perspectivas*, Proceedings, (Santiago de Cuba, 19-23 de Enero 2015). Centro de Lingüística Aplicada, Ministerio de Ciencia, Tecnología y Medio Ambiente. Volume I, Pages 223-227.

---

[8] ICT stands for Information and Communication Technology.

Kageura K. (1999). Theories 'of' terminology: A quest for a framework for the study of term formation, *Terminology*, Volume 5, number 1, (1998/1999). Pages 21-40.

Megerdoomian K. (2003). Text Mining, Corpus building, and testing, in Ali Farghaly*, Handbook for Language Engineers.* CSLI Publications, Stanford. Pages 213-268.

Park Y., Byrd Roy J., Boguraev Branimir K. (2002). Automatic Glossary Extraction: Beyond Terminology Identification, COLING '02, *Proceedings of the 19th international conference on Computational linguistics*. Taipei, Taiwan, August 24 - September 1 Association for Computational Linguistics , Volume 1, Pages 1-7.

Pazienza M.T, Virdigni M.(2003). Agent based ontological mediation in IE systems. In: *Information Extraction in the WEB Era. Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents*, edited by M.T. Pazienza (Lecture Notes in Artificial Intelligence, v. 2700), Heildelberg, Springer.

Temmerman R. (1997). Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology. *Journal of Linguistics* n° 18 - 1997. Pages 51-90

<http://cordis.europa.eu/home_en.html> last accessed on January 15, 2016.

<http://www.maldura.unipd.it/romanistica/cortelazzo/mlat/tc.html> last accessed on January 15, 2016.

<http://ec.europa.eu/research/health/6th-framework-programme_en.html> last accessed on January 15, 2016.