OECD

CSTP-TIP Workshop

Semantic Analysis for Innovation Policy

12-13 March 2018 – Paris, OECD Conference Centre

Textual Data Analysis

lexicometric approach, explorative methods and application on TIP Corpus

Pasquale Pavone

www.textminer.it

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

CAPP – Centro di Analisi delle Politiche Pubbliche

Automatic processing for text analysis, **enables us to not read a text** but **different representations** of the **information contained** in it.

Text Mining is a multidisciplinary research field that combines with equal importance, instruments from ***Computational Linguistic***, ***Information Retrieval*** and ***Statistics,*** with the purpose of extracting **information of interest** from a collection of documents (Corpus)

Usually a ***Text Mining (TM)*** strategy includes the following steps:
Pre-processing
Lexical processing
Information Extraction:

Taltac2 Software **-** consists of a series of instruments which allow for the study of any kind of linguistic data - collected in a Corpus - by employing the techniques of "**textual statistics**". The automatic processing, based on a **lexicometric approach**, allows us to find certain constants in the text, a sort of **DNA** of the **Corpus**

The Role of Meta Data in the Automatic Analysis

meta-data *generated during the analysis, consists in* ***annotation*** *of* ***lexical units*** *and* ***categorization*** *of* ***textual units***

Lexical Automatic Processing: *object of study is the* ***Vocabulary***

Annotations are made on Vocabulary DB (types table)

elementary lexical unit is the ***lexeme (graphical forms)*** (type = atom of meaning)

Textual Automatic Processing: *object of study is the* ***corpus***

Like a collection of texts to categorise on Documents DB

textual unit is the ***document*** (or fragment of text)

In **Taltac2** the unstructured textual information is structured in two main **DB**

Vocabulary DB

Annotation Meta Data

- Linguistic:
    - Grammatical Tagging
    - Semantic Tagging (by dictionaries - glossary)

- Numerical Statistic:
    - Peculiar (dev. from a model)
    - Specific (dev. from eq.categ.)
    - Relevant (TFIDF on Vocabulary)

Recognition of Multiwords

- Analysis of repeted segments
- Lexical-Textual **Algoritm**

Matrix Words x Categories

Textual DB

Categorization – Meta Data

- Regular Expression
- TFIDF (textual query - dictionaries)

Selection of Documents:

- Concordances Analysis
- Co-Occurence Analysis

Matrix Documents x Words

Multidimensional Analysis, is an **explorative method** and consists in a reproduction of dimensions (factors) through which we can **simplify**, **synthesise** and **represent** the **Studied Phenomenon**.

It is an **overall analysis** that intends a **Corpus** as a system and puts in evidence the relations exisisting in the whole system, based on the **euclidean logic**.

The method of **Correspondence Analysis** allows us to represent the **information in terms of similarities** among the elements of of **row profiles** vs. **column profiles** in each **factorial plan**

By ***Cluster Analysis***, on a **factor analysis** (on a **matrix «Documents x Words»**), we are able to **group documents** based on their similarity in **terms of words**. The **dictionaries characteristic** of each **cluster** represent the **topic** (or **topics**) of each one.

What Strategy to use?

- Which are the **criteria** (among all of those we possess) to identify and select the lexical units of analysis? (word – multiword – entities)
- Which is the **definition** of **textual units** of analysis?

It Depends each time on the Corpus under analysis

From its **main characteristics** and the **objectives** we want to reach.

- A **Corpus** could be a Collection of: web page texts – newspaper articles – political speeches – scientific papers (title, Abstract, Full paper) – Post or Comments in Social Networks – **collection of technical documents** – open ended questions…
- Common language vs specialistic-technical language
- Sectorial **Homogeneity** vs **Dishomogeneity** of the language of the various documents
- Dimensional **Homogeneity** vs **Dishomogeneity** of the documents
  - if homogeneous, they can be either very short or very long

Corpus of WP-TIP Activities

It is a **Technical Specialistic Corpus** and it is composed of **274 documents** published from 1993 to 2017

Vocabulary consists of **58.592 Types** for a total of **2.772.128 occurrences**

Objectives:

- **Identification of terminology**
- **Identification of the activity topics over time**

Starting points:

- **No initial Hypothesis**
- **No previous knowledge of the TIP activity**

1) Terminology Identification

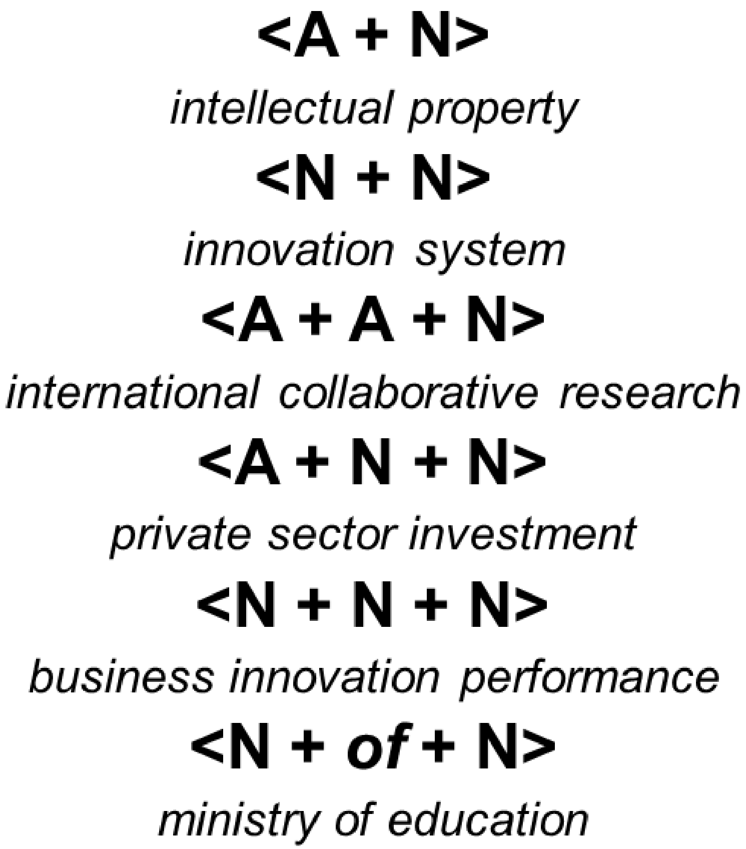
We focused on identifying **nominal idioms**, **collocations** and **complex lexemes** to obtain a dictionary of non ambiuguos semantical terms

A **Collocation** is a sequence of **two or more words** characterized by a **strong reciprocal link (Sinclair, 1991)**

Complex Lexemes, particularly existent **in a technical specialistic Corpus**, represent an important part of the terminology of the corresponding sector. **Although they are not nominal idioms**, they **represent a technical specialistic expression (De Mauro 1999-2003).**

1) Terminology Identification

Definition of the syntactical structures of the most common collocations



<A + N>

intellectual property

<N + N>

innovation system

<A + A + N>

international collaborative research

<A + N + N>

private sector investment

<N + N + N>

business innovation performance

<N + *of* + N>

ministry of education

<A + N> + *of* + N>

present value of depreciation

<N + *of* + <A + N>

mobility of human resource

<A + N> + *of* + <A + N>

national bureau of economic research

The added value of the structure **<N of N>** is given by the preposition "***of***", which introduces the second nouns as a property of the first **(Rouget, 2000)**

Each structure is considered as a Regular Expression

The meta-query composed of all Regular Expressions identifies in the Corpus all the possible sequences. By **Lexicalization** of the entities with more than 10 occurrences we obtain **3.337** new types in the Vocabulary

MultiWord Expression Type	Occ	MultiWord Expression Type	Occ	MultiWord Expression Type	Occ
innovation policy	1033	***policy maker***	441	*research activities*	131
case study	844	*european commission*	429	***business model***	131
technology policy	737	*national innovation system*	424	[…]	[…]
intellectual property	664	***system innovation***	413	***productivity growth***	129
innovation system	640	*research organisation*	400	***critical mass***	129
tax credit	638	***technology transfer***	363	[…]	[…]
working group	634	*focus group*	346	***innovation activities***	119
private sector	603	[…]	[…]	***government policy***	119
working party	543	***policy measure***	141	***business angel***	119
tax incentive	538	***innovation performance***	141	***market failure***	118
research institute	524	***technology development***	141	[…]	[…]
innovation process	457	[…]	[…]	***behavioural additionality***	116
research institution	454	***knowledge transfer***	135	*foreign firm*	114

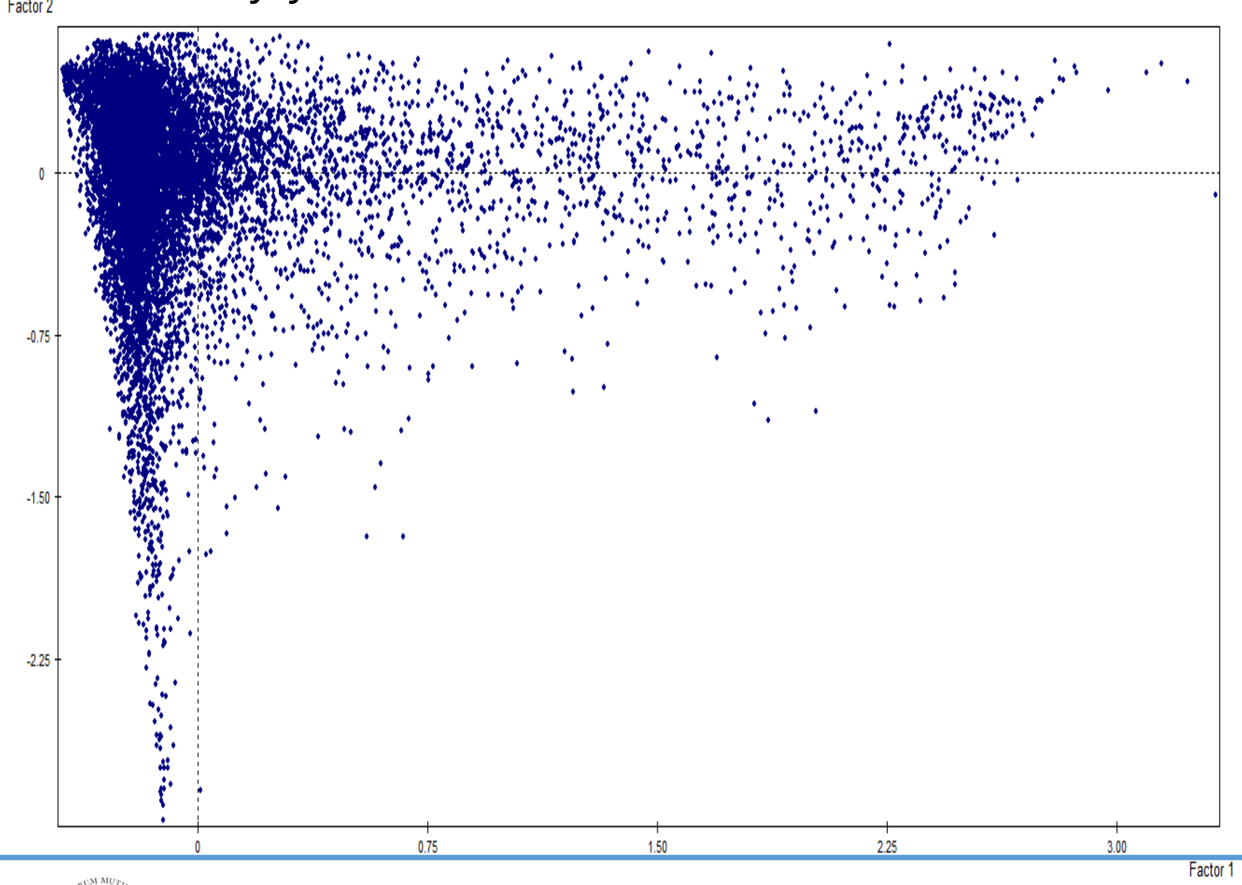
2) Exploration of Activities

Factor Analysis - selection of all nouns, words and multiwords, (that represent the subjects and objects of a discourse) and adjectives **with minimum 5 occurrences**.

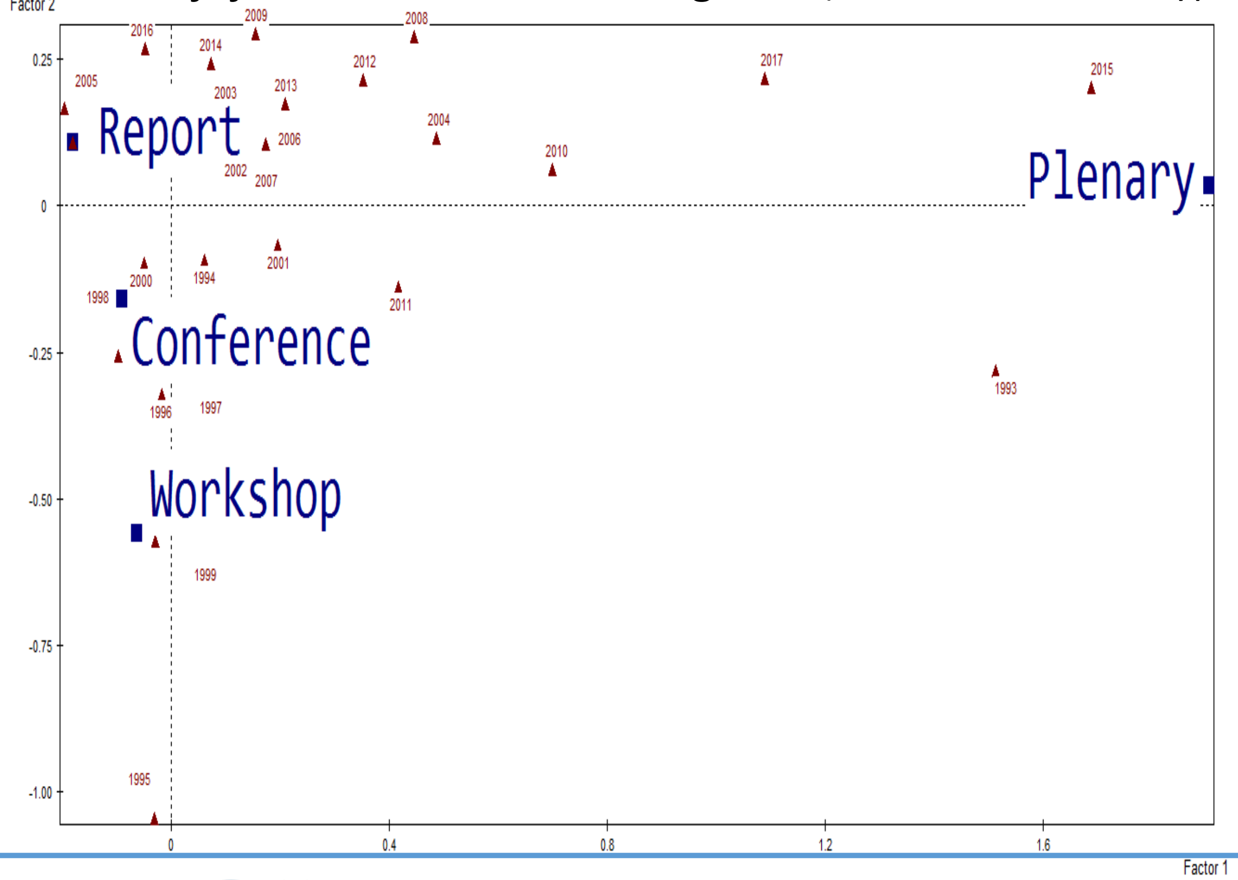
12.119 selected terms, in which **3.337** are **multiword expressions**

Matrix **Selected Terms × Categories (3285×29)**

Factor Plan *f1f2* - Distribution of Lexicon



Factor Plan *f1f2* - Distribution of Categories (25 Years – 4 Event Types)

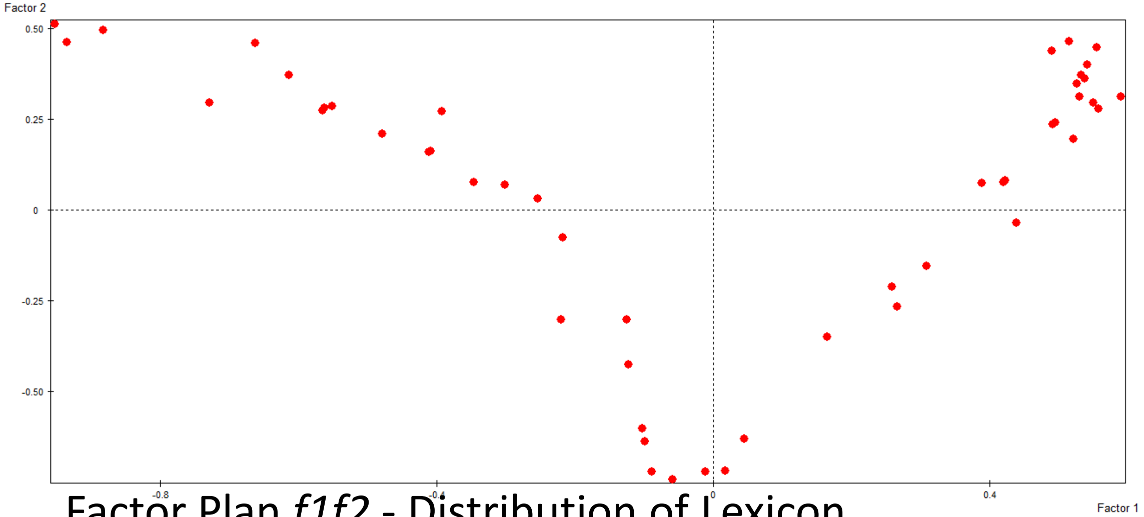


Plenary Analysis

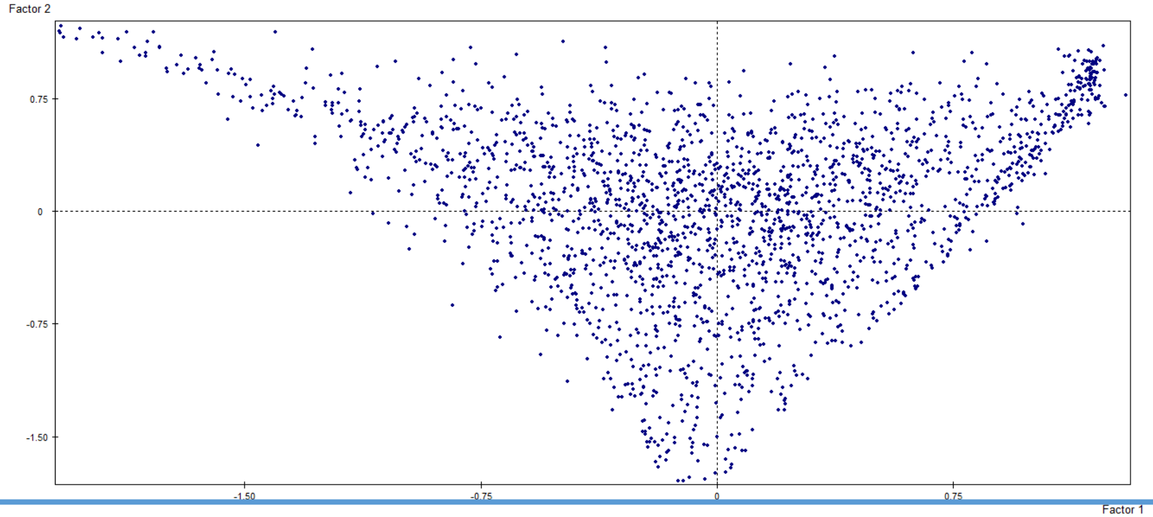
Factor Analysis selection of all nouns (words and multiwords) and adjectives **with minimum 5 occurrences**. **2.253** **selected words**, in which **777** are **multiword expressions**

Matrix **Events** × **Selected Terms** **(49 × 2.253)**

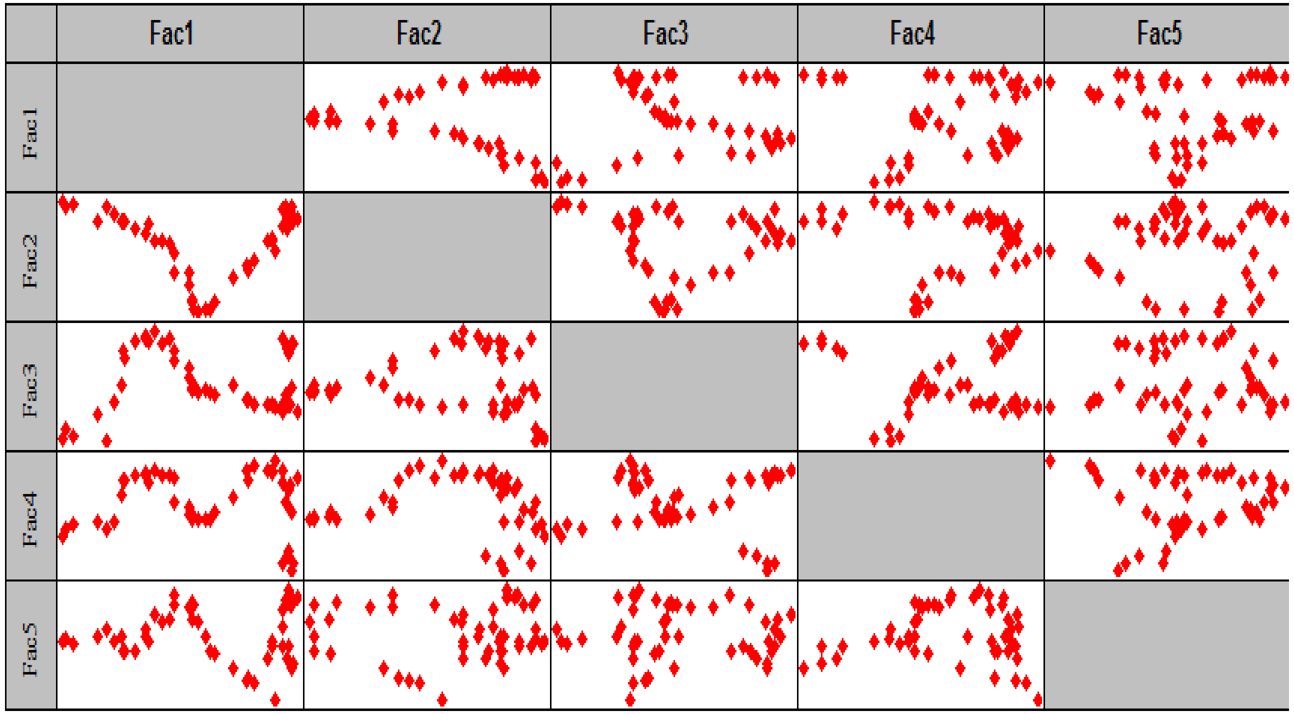
Factor Plan *f1f2* - Distribution of Documents



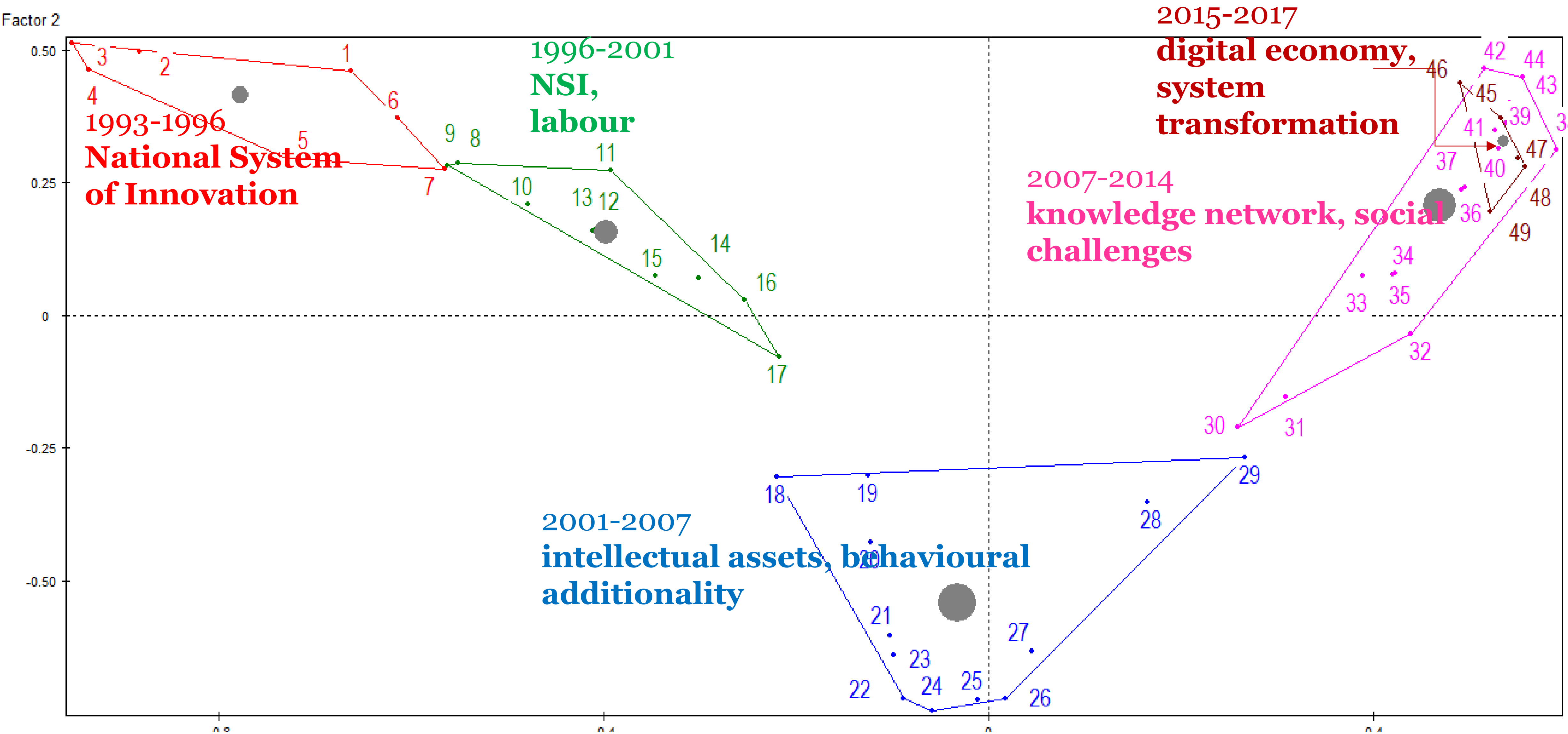
Factor Plan *f1f2* - Distribution of Lexicon



Distribution of Documents on the combination first 5 factors



Factorial Plan *f1-f2* – Distribution Documents – Convex Hulls - **5 Temporal Clusters**



1993-1996
National System of Innovation

1996-2001
NSI, labour

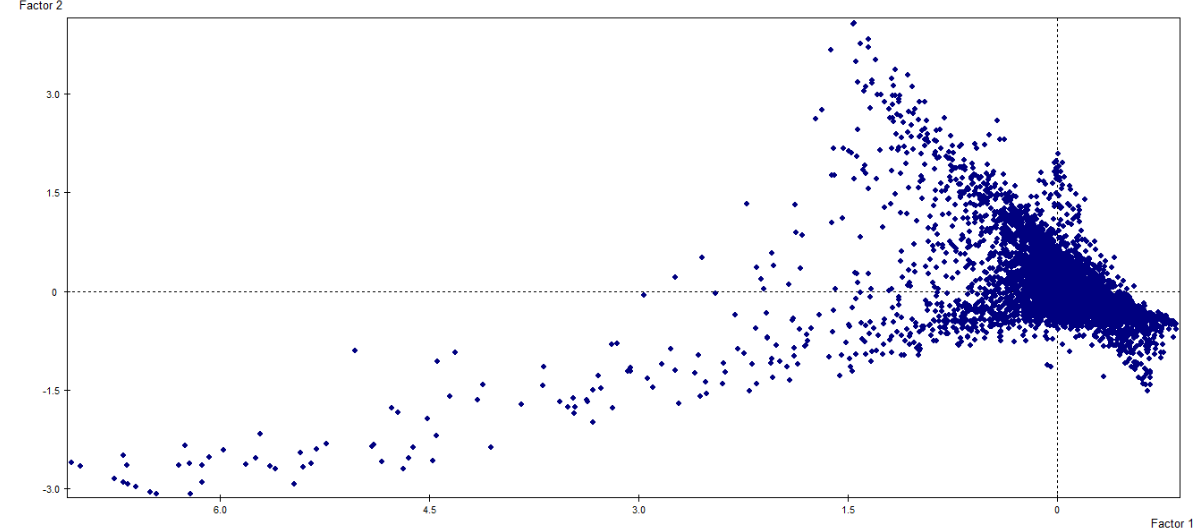
2001-2007
intellectual assets, behavioural additionality

2007-2014
knowledge network, social challenges

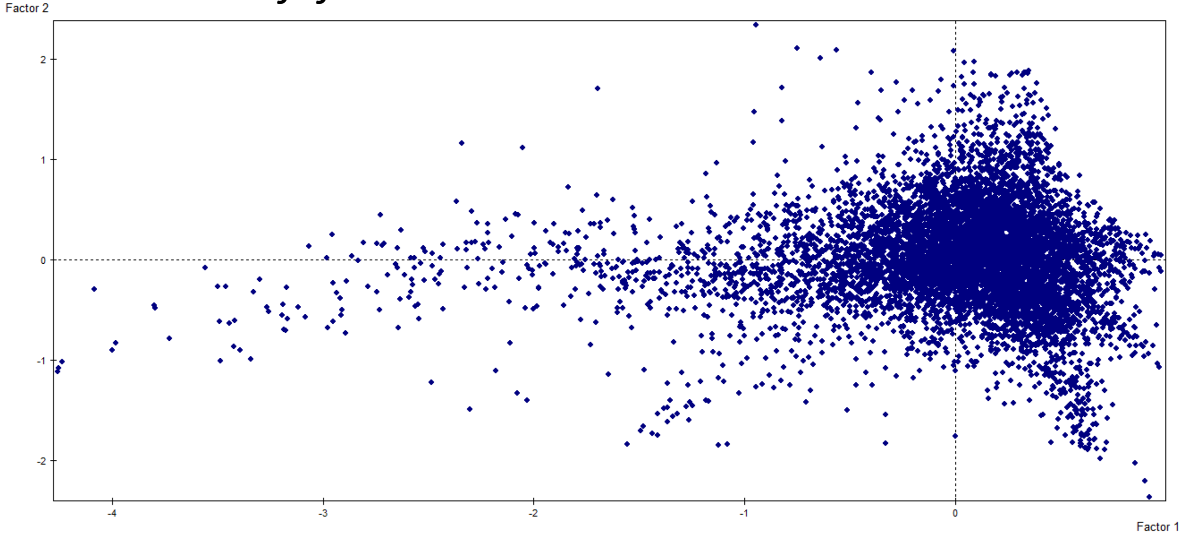
2015-2017
digital economy, system transformation

Report Analysis

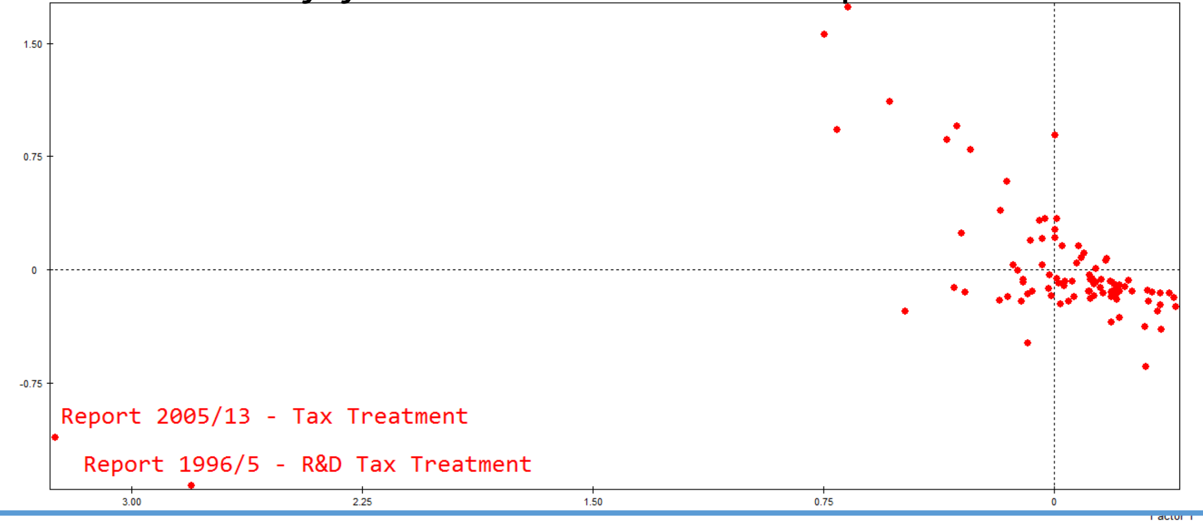
Factor Plan *f1f2* - Distribution of the Lexicon



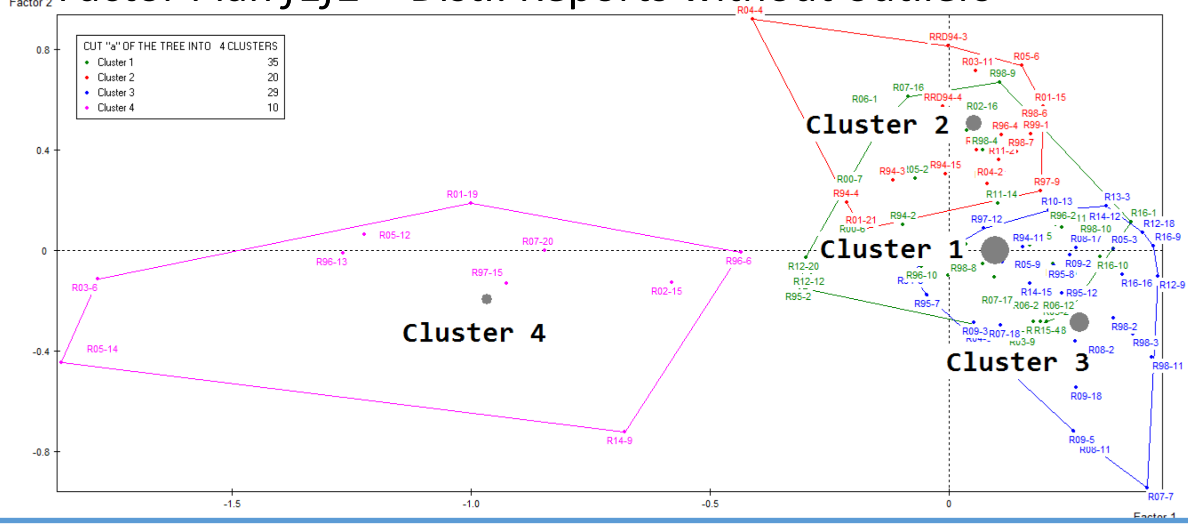
Factor Plan *f1f2* – Distr. Lexicon without outliers



Factor Plan *f1f2* - Distribution of the Reports



Factor Plan *f1f2* – Distr. Reports without outliers



In Conclusion Explorative Analysis:

Does not need

- **initial hypotheses**
- **previous knowledge of the domain under analysis**

Needs

- **Homogeneity**
  - **Language**
  - **Textual Units Dimension**

Preliminary steps

Purpose of the analysis:

- **It is NOT to model the Data based on a previous Hypothesis**
- **It is To Explore the Data to represent the Information contained in it**

Costs of the analysis:

- Data Analysis Competencies +
- Software (0 - 3.000€)