The making of the TIP@50

Methodological choices underlying the choice of associated words method – based on Cortext

Antoine Schoen & Philippe Larédo

OECD meeting, March 12th, 2018

RISIS
Research infrastructure for research
and innovation policy studies

Input materials used

- 330 out of 460 documents of the OECD TIP corpus received - we have excluded agendas and minutes of TIP meeting (98 documents together); and documents that are an analysis of the situation of an individual country (34 documents).
- We have used the executive summaries or have rebuilt them (when missing or too short).
- Note : an exploratory test with full text of reports produced very similar results

Various methodological approaches

Typology of models for the analysis of textual corpuses in Social Sciences

Approach	Main software	Underlying Maths	Type of data and ouput	Foundation / sociological model
Associated words	Leximappe, Calliope, VosViewer, CiteSpace, Cortext	Analysis of co-occurence, community detection,	Strategic diagrams, network maps	Sociology of translation, actor-network theory (Michel Callon)
Political lexicography	Lexico, Hyperbase	FCA	Lists	Analysis of political discourse through frequencies
Correspondence analysis	FactoMiner, Prince	FCA, MCA	Factorial space	Proposed by Benzecri, popularised by Bourdieu
Alceste	Alceste, Iramutek, TXM, Tlab.it	Frequential	Classification of lexical worlds	Developed by Reinert / focus : internal organisation of discourse
Topic modelling	Gensim, topic models R	Bayesian generative models	Probabilistic classifications	LSA
Word embedding	Glove, Gensim..	Neuronal networks	Continuous Semantic space	Mikolov (Google & Facebook)

Source: Jean-Philippe Cointet, (2017), Cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales, Mémoire de synthèse en vue de l'obtention de l'Habilitation à Diriger des Recherches

Our methodological approach

- The approach chosen is part of the family of ‘associated words’ which analyses groups of documents together, identifying within them the most ‘distinctive’ ‘multi-words’ (noun phrase - not innovation alone but innovation systems) and analysing their clustering bottom-up: it provides a view of groupings (what is called community detection), of their internal consistency and of the linkages between groupings

Software decisions

- We use CORTEXT*, for 2 main reasons:
- (a) it is a freely accessible software that is relatively user friendly though it offers for advanced users all the possibilities of other software in the same family;
- (b) it has a fast growing community of users which help in its evolution.
- In addition, the Cortext platform offers several other scripts : topic modelling, geo-coding, NER…)

* https://managerv2.cortext.net

Methodological insight 1

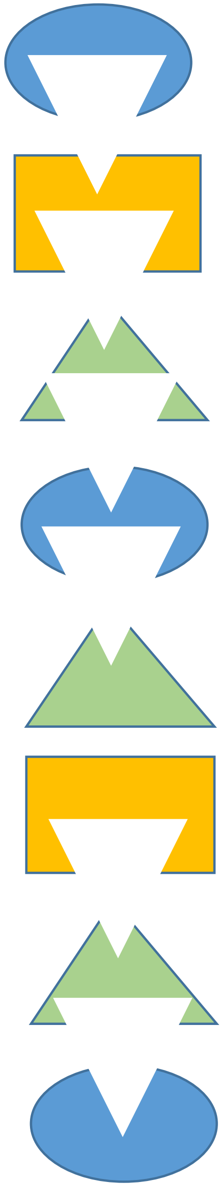
- We want to underline that whatever the software, the machine does not replace theory, and that without an underlying approach to the analysis made, results have all the chances to be meaningless, or not interpretable.

Methodological insight 2

- Human expertise is required all along the process :

1. For selecting and organising the documents
2. For choosing properly the adequate method(s)
3. For pruning and grouping the automatically selected noun phrases.

An expert – data “dialog”



Experts delineate the corpus

Codified Corpus and Cortext-structured DB

Cortext extracts terms

Experts edit the terms

Cortext indexes the corpus with edited terms

Cortext-structured DB with edited terms

Cortext processes relational maps

Experts analyse relational maps