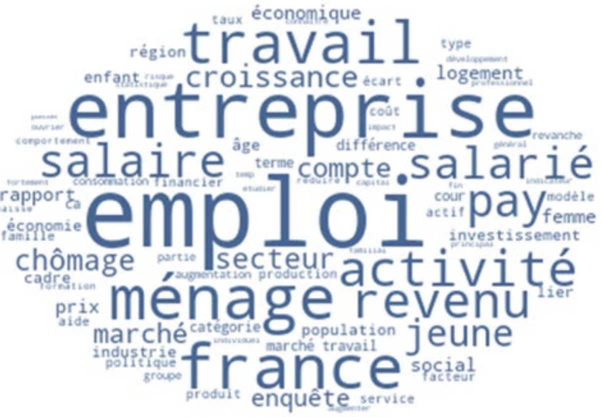
Fifty Years of Abstracts in the Journal Economie et Statistique

Julie Djiriguian* and François Sémécurbe*

Natural language processing, is nowadays a toolbox routinely used to explore the content of various texts. On the occasion of the 50th anniversary of the journal *Économie et Statistique* (then *Economie et Statistique / Economics and Statistics*), we propose in this short article an application to the abstracts of the 2,184 "academic" articles published in this journal since 1969 (see Box). Which words are most frequently used? What underlying topics do they suggest and have these topics changed over the years?

After preliminary treatments (Box), we obtain a set of 181,572 words for the 50 years. A representation in the form of a word cloud highlights the most frequent words (Figure I).

Figure I
Word cloud on the corpus of abstracts from 1969 to 2019



Note: Our apologies to those who do not read French, but the journal has published in French for most of its life, and it would not have made sense to translate the stock of words. Some translations will be provided below and in the rest of the article.
Reading Note: 'emploi' (employment) is the most frequent word in the entire corpus of Economic and Statistical abstracts (with 2,176 occurrences out of 181,572 words). The next most frequent words are 'entreprise' (companies), 'travail' (work), 'ménage' (household).
Sources: Abstracts of academic articles, *Économie et Statistique* (1969-2016) and *Economie et Statistique / Economics and Statistics* (2017-2019).

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

JEL Classification: C38, C63
Keywords: text analysis, natural language processing, topic modeling, Latent Dirichlet Allocation

** Insee, SSP Lab (julie.djiriguian@insee.fr; francois.semecurbe@insee.fr)*

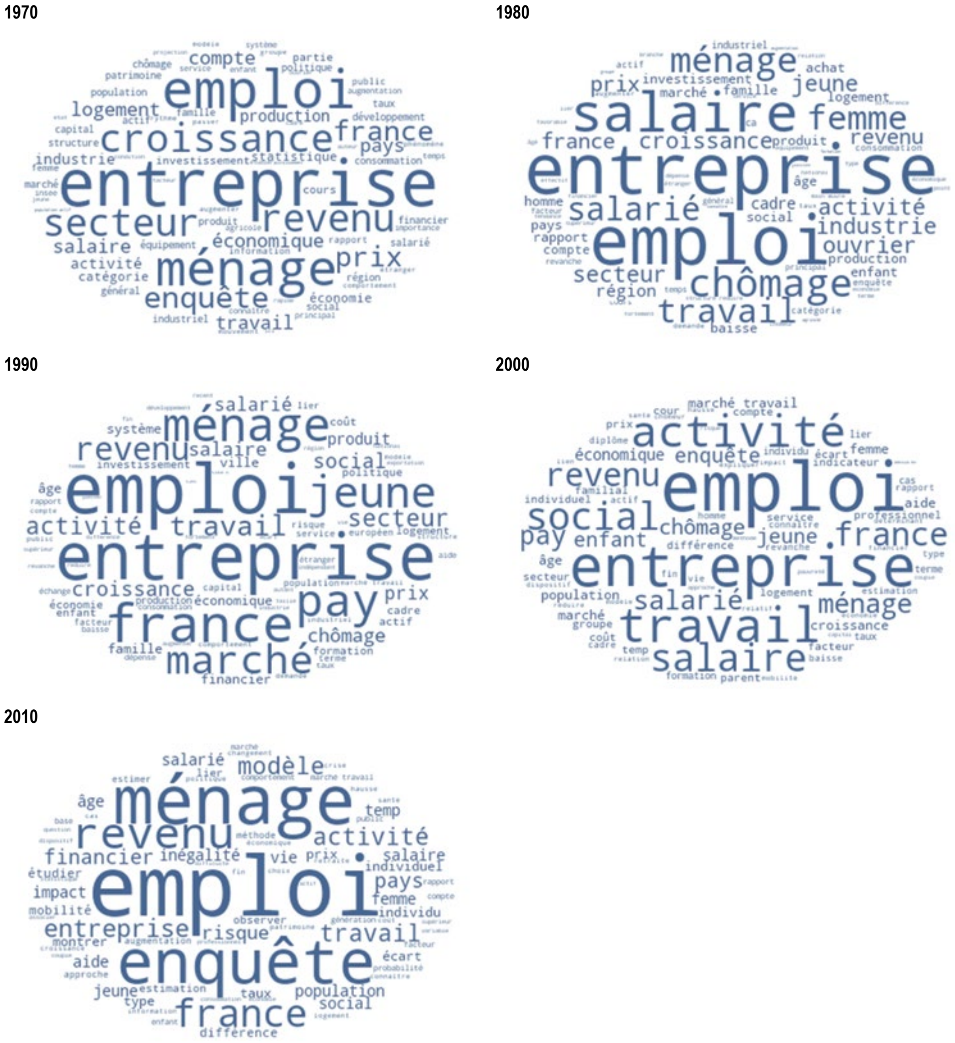
Translated from the original version: "Cinquante ans de résumés d'Economie et Statistique"

Citation: Djiriguian, J. & Sémécurbe, F. (2019). Fifty Years of Abstracts in the Journal Economie et Statistique. *Economie et Statistique / Economics and Statistics*, 510-511-512, 7–11. https://doi.org/10.24187/ecostat.2019.510t.1999

As a whole, this representation of the vocabulary of the abstracts illustrates first of all the generalist nature of the journal. The word with the highest relative frequency is ‘employment’ (emploi), then, by decreasing relative frequency, the words ‘company’ (entreprise) and ‘household’ (ménage).

The most frequent words over fifty years are, of course, also most frequent by decade, and the trilogy ‘employment’, ‘company’, ‘household’ is confirmed, even if in a variable order until the decade 2000 and with some eclipses: ‘household’ in the decade 2000, ‘company’ in the last decade (Figure II). Variability is much greater for words with a lower relative frequency.

Figure II
Word clouds by decade



Sources: Abstracts of academic articles, *Économie et Statistique* (1969-2016) and *Economie et Statistique / Economics and Statistics* (2017-2019).

The most constant is the word ‘employment’. However, it would be adventurous to interpret this dominance as a sign of a “specialization” of the articles published in the journal. Rather, it can be seen as a “hub”, around which many angles of economic analysis can be articulated, analysis of the activity at the macro level or, at the micro level, of the behaviour and situation of its actors, companies and households. If we pull a little on the thread, we can also see it as reflecting an almost permanent concern for employment since the late 1970s, which would make it either the subject of interest or the entry point for many articles; and if we draw a little further, we can recall that the *enquête Emploi* (the French LFS) is one of the oldest of Insee's surveys covering the working age population, and used in a large number of articles published in the journal.

Quantifying the most frequent words – even more since they are only "words" and not "keywords", and they are considered independently of each other – is obviously not enough to describe the contents of a set of texts. Topic modelling methods allow associations to be identified by simultaneously analysing all the words that are part of a text. To explore a little further, we use here Dirichlet's latent allocation (LDA, see Box), which is based on probabilistic modelling. This method is frequently used to interpret underlying topics based on the group of words that characterize them. It should be noted, however, that, like any textual analysis, this method is based on strong hypotheses and choices (in particular at the pre-processing step) that condition the result, and that the identification – or interpretation – of topics based solely on the words associated with them can be tricky.

This method requiring to fix a priori the number of topics, we have fixed it at three. After the various estimations made for all the abstracts, we obtain the following associations of words that we will, for convenience, call by their first most frequent word - which refers (necessarily, since at the base is the same "stock" of words) to one or the other of the three words that appeared most frequently in Figure I:

- a topic called "companies", which evokes the vocabulary of economic activity in the most standard, rather macroeconomic sense – including the words:

ecompany / growth / sector / france / industrial / market / country / production / industry / activity / employment / economic / price / investment / economy / labour / development / rate / productivity / product / price / demand / decline / structure / trade / trade / politics / foreign / small / foreign / term / account / capital / region / region / service / productive / explain / domestic / cost / equipment / high…

- a topic called "households" which is more about combining words from the vocabulary of income and living conditions:

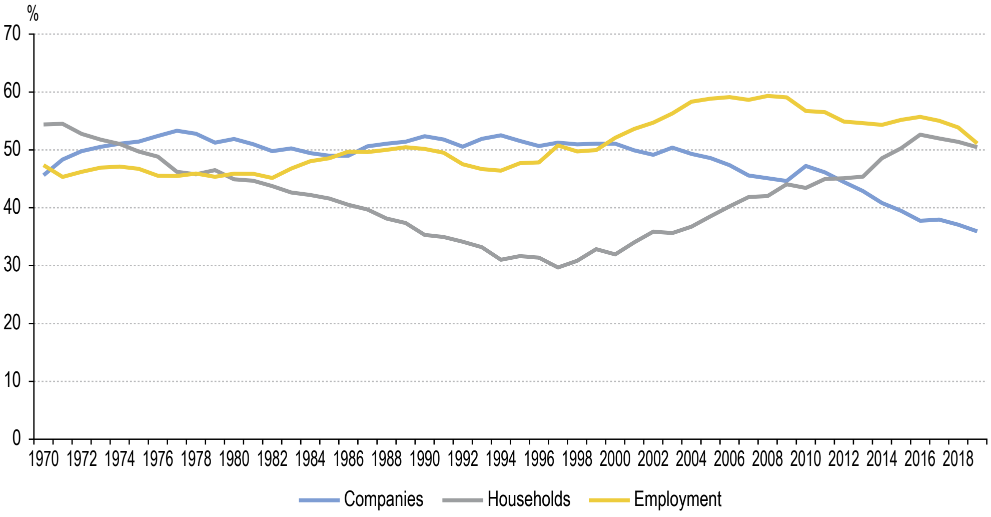
household / survey / income / account / economic / statistical / social / model / france / term / method / financial / housing / question / estimation / policy / work / information / consumption / rate / main / system / life / price / insee / help / national / public / approach / driving / cost / take / behaviour / base / individual / study / expenditure / population / evaluation / economy…

- a topic called "employment", where we find the categories of microeconomic analyses of the labour market:

employment / work / youth / woman / age / active / employee / unemployment / professional / activity / category / social / man / time / worker / life / child / wage / duration / manager / population / survey / family / higher / market / occupy / training / familial / old / sector / profession / increase / courses / gap / generation / work / diploma / unemployed / contrast / decrease…

To finish, we can represent the topics over the years in terms of their "weight", i.e. the proportion of abstracts containing at least three of the main words of each topic (Figure III). The presence in the abstracts of words associated with the topic "employment" tends to rise from the 1980s, then more markedly from the mid-1990s to the end of the 2000s. On the contrary, the number of words in the topic "companies" decreased from the 2000s onwards. Finally, the presence of the words of the topic "households" presents a more singular aspect, with a decline until the second half of the 1990s, then an increase of equivalent magnitude thereafter.

Figure III
Proportion of abstracts containing at least one of the three main words of each topic



Note: Adding the weights results in a total over 100%%, since a given word can appear in more than one topic.
Sources: Abstracts of academic articles, *Économie et Statistique* (1969-2016) and *Economie et Statistique / Economics and Statistics* (2017-2019).

* *
*

We are here at the limits of the exercise proposed in this short article, which had no other purpose than illustrative. To draw an interpretation would require a much larger investigation... which will have to wait, because if 2,184 articles and 181,572 words seem "a lot", it is a small corpus for the implementation of the techniques used here. □

Box – **Methodology**

Textual analysis gathers all the methods used to extract and analyse the information contained in texts. It can be used on data from a wide variety of sources (administrative texts, legal decisions, discussions on social networks, etc.) to reveal underlying topics, analyse feelings, or predict a variable (see Anzovino *et al.*, 2018; Wu *et al.*, 2018; Xing *et al.*, 2018). Text data is by nature unstructured, so any textual analysis process begins with a step of preparing the text to clean it up and transform it into usable digital data. These digital data are then used for the statistical analysis itself.

The statistical observations of a textual analysis are texts, called 'documents' (here the abstracts). Each document is divided into elements (tokens), words, punctuation, association of several words (n-grams) if necessary. Pre-processing consists first of all in removing non-informative elements - punctuation, numbers, and 'stopwords'. Stopwords are insignificant words that appear in the entire corpus studied; some are "obvious" (for example, conjunctions) but others require arbitration, necessarily subjective. All the terms remaining after this step will constitute the 'variables' of the analysis. These pre-treatment steps are often tedious and can involve arbitrary or ad hoc choices. The subsequent textual analysis is therefore very sensitive to these choices.

The informative words are then standardized to make them comparable: case harmonization, spelling correction, 'lemmatization'. Lemmatization consists in finding the neutral form of a word: for example, a conjugated verb is found, after this operation, at the infinitive. This operation is complex, because it requires in particular to clarify cases of homonymy. Documents and 'variables' can then be represented by a numerical matrix where each line measures for a given document the number of occurrences (or another measure: binary coding - presence/absence - is classic) of each word/variable of the whole vocabulary retained within each document. The resulting matrix is often large (there are more words/columns than documents/lines) and sparse (many 0). It can be analyzed using various statistical methods. ➔

Box (contd.)

The interpretation of a text results from the association of words (Hapke *et al.*, 2019). To examine these associations, we have implemented here an analysis related to topic modelling called Latent Dirichlet Allocation (LDA, cf. Blei *et al.*, 2003). It is a generative probabilistic model, which estimates by Bayesian inference methods (variational Bayesian, Gibbs sampling, etc.) from the words observed in the documents, the weight of the topics in each document and the distributions of the characteristic words of each theme. This method requires that the number of topics be determined *a priori*.

LDA is based on strong assumptions that need to be underlined. First, the estimation of the parameters of the word distributions for each topic and those of the topic within a document begins with a random initialization: two different initializations can generate two different thematic structures. Second, LDA, like a large part of textual analysis methods, is based on the assumption that the order of words has no impact (it is referred to as a 'bag-of-words' approach). Under this assumption, the documents are divided into unordered lists of words. As words are also decisive for the interpretation of themes, pre-treatment is also crucial here.

The analysis presented here covers the abstracts of all the academic articles published in the journal between 1969 and 2019. Prior to the pre-processing of the texts, we excluded 764 "non-academic" articles: until the 1970s, the journal published the presentation of survey results, territorial panoramas, or other short articles of information that then disappeared (or gave rise to publications in other Insee collections). General introductions to special issues were also not included. There remain 2,184 abstracts containing 432,000 words.

Pre-treatments have mainly consisted in removing the figures and stopwords contained in the abstracts and "lemmatizing" the words. To this end, we used the spaCy library with Python, which detects the grammatical function of words in a text, which is more effective than using a simple dictionary. In addition to the stopwords proposed by spaCy, we have excluded *ad hoc* words, on the basis of based on the results obtained in statistical analyses (for example, the word 'year', which produces insignificant links between summaries). At the end of these pre-processing operations, the database contains 2,184 abstracts and 181,572 words.

References

Anzovino, M., Fersini, E. & Rosso, P. (2018). Automatic identification and classification of misogynistic language on twitter. In: *International Conference on Applications of Natural Language to Information Systems,* pp. 57–64. Springer, Cham.

Blei, D & Ng, A. Y. & Jordan M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 3, 993–1022.

Hapke, H. M., Lane, H. & Howard, C. (2019). *Natural language processing in action*. Manning.

Wu, J. T., Dernoncourt, F., Gehrmann, S. ... & Celi, L. A. (2018). Behind the scenes: A medical natural language processing project. *International journal of medical informatics*, 112, 68–73.

Xing, F. Z., Cambria, E. & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1), 49–73.