Insights into the French socio-ecological research network through Natural Language Processing

Ingrid Falk, Isabelle Charpentier

Insights into the French socio-ecological research network through Natural Language Processing

Ingrid Falk[a], Isabelle Charpentier[a,b]

[a] *LTSER-FR Zone Atelier Environnementale Urbaine, 5 allée du Général Rouvillois, 67000 Strasbourg, France*
[b] *ICUBE (UMR7357), CNRS&Unistra, 2 rue Boussaingault, 67000 Strasbourg, France*

Abstract

The French long term socio-ecological research network RZA (*Réseau des Zones Atelier*) promotes inter- and trans-disciplinary research on essential socio-ecological systems (SESs). Scientific research on SESs has grown exponentially since the seventies, but because of the heterogeneity in the actors, the disciplines and the collected data some efforts are still necessary to build a common language and, in the next future, a thesaurus for indexing data.

We use Natural Language Processing (NLP) methods to analyse a French corpus derived from the *5th* colloquium of the RZA which marked the network's 20th anniversary. Using the topic modelling technique we show how these methods allowed us to gain insights about the current status of research within the RZA community. We investigate the involved vocabulary to cross reference the subjects of interest and explore how well these automatically extracted topics are related to the ambitions of the RZA community in terms of inter- and trans-disciplinary environmental research.

Keywords

Natural language processing (NLP), topic modelling, socio-ecological system (SES), *Réseau des Zones Ateliers* (RZA)

1. Introduction

The French long term socio-ecological research network (LTSER), namely *Réseau des Zones Ateliers* (RZA) [1], is a distributed inter-agency research network conducting action-oriented studies on essential socio-ecological systems (SESs) including river basins, rural, urban or mountainous territories, and protected areas.

With a special emphasis on human-nature interactions and sustainability [2], this inter- and trans-disciplinary research network promotes the observation of environmental variables and practices through *in situ* experimentation, the definition of key environmental indicators and the design of data-based decision support tools.

Therein, researchers from various disciplines of Natural Sciences and Engineering, Humanities and Social Sciences, as well as local administrations and citizens have the opportunity to share their knowledge and co-build research activities around the adaptation of SESs to internal changes and external constraints (climate change, globalisation, biodiversity loss...) and their transformation into more sustainable ones. The scientific research on SES has grown exponentially since the seventies [3]. However, the heterogeneity in both the actors and the collected data still requires some efforts to build a common language and, in the next future, a thesaurus for indexing metadata and data [4].

ORCID: 0000-0002-7351-6628 (I. Falk); 0000-0001-7511-2910 (I. Charpentier)

For its 20-year anniversary in 2020, the RZA organised its $5th$ colloquium focusing on the analysis of SES trajectories, the resilience and environmental perspectives, and the co-construction of research with local stakeholders. Contributions were selected on the basis of 500-word abstracts. These form the RZA corpus we analyse with Natural Language Processing (NLP) tools with the objective to better understand the French SES research community, its characteristics and activities.

More specifically we use the topic modelling NLP technique to assess the thematic structure of this corpus by automatically extracting main topics. We analyse how well these topics reflect the subjects of interest to the RZA community and the inter- and trans-disciplinarity of the environmental research conducted within this network.

The layout of the paper is as follows. Corpus and methodology are presented in Sections 2 and 3. Topic models with higher semantic coherence scores are computed and analysed in Section 4 with respect to their plausibility and their findings and we discuss the limits of corpus and methodology. Section 5 provides a conclusion and raises some research opportunities.

2. Data - the corpus

For our linguistic analysis we compiled the French abstracts of the 5^{th} colloquium of the Réseau des Zones Ateliers into a digital corpus on socio-ecological research.

We hypothesise that this corpus may reveal newly emerging, more or less, inter- and trans-disciplinary research subjects by exploring how well these transcend disciplinary spheres (atmo-, bio-, geo-, hydro-, socio-) still in use to represent the environment and research on it [5].

The submission guidelines and the three main themes were chosen collectively by the conference's organisers, independently from this study:

1. analyse trajectories of socio-ecosystems through human-nature interactions, water management or environmental risk,
2. consider socio-ecosystems from the perspective of resilience or environmental health,
3. co-build an applied research with local stakeholders and mobilise trans-disciplinarity in research (collective actions, governance...).

The corpus consists of 81 abstracts of on average 412 words, amounting to about 35000 words in total. This means the corpus is small but well balanced, with documents of about the same length, suggesting a certain structural consistency.

To assess the semantic similarity in this corpus we needed to compute a word embedding (Section 3). As our corpus is small, the pre-processing was crucial since it allowed to keep only words which are potentially meaningful for our analysis.[1]

Figure 1 gives a frequency based overview of the corpus content.

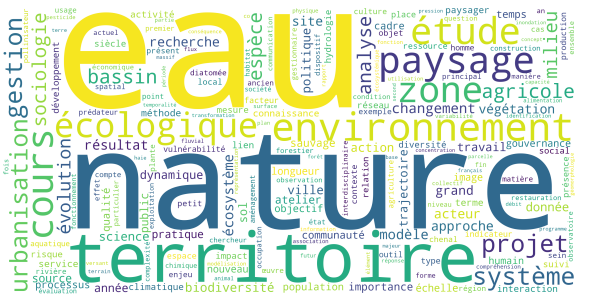

Figure 1: Frequency based word cloud for the pre-processed RZA corpus. The most frequently words are *eau (water)*, *nature*, *territoire*.

[1] We used SpaCy for removing stop-words and punctuation, perform lemmatisation, and only kept nouns, adjectives, verbs, adverbs and proper nouns.

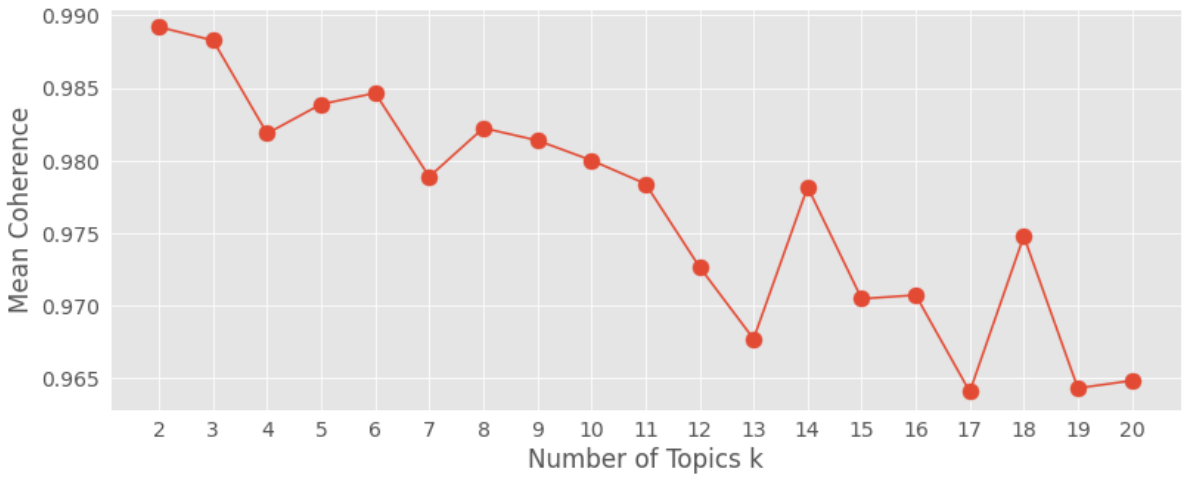

Figure 2: Semantic coherence scores for topic models with 2 to 20 topics.

3. Methods - Topic Modelling

A topic model (in machine learning and NLP) is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. It can be used to explore hidden semantic structures in a corpus of texts. It takes a collection of texts as input and identifies a set of "latent" topics in the form of lists of words characterising each topic. These collections of words are then considered the main subjects addressed in the corpus. The model assigns to each document a finite number of topics with a specific (probability) weight.

Two algorithms are mainly used to build topic models: *Latent Dirichlet Allocation* (LDA, [6]) and *Non-negative Matrix Factorisation* (NMF, [7]). LDA is a probabilistic graphical model while NMF is based on multivariate analysis and linear algebra. As for LDA, solutions are found by numerical approximations.

When applying topic modelling, with both LDA and NMF, one is confronted with two major questions. Firstly, the number of topics of the method is essential and should be chosen with caution. Secondly, an important issue is how to evaluate whether the resulting topics are plausible and meaningful. In practice, topics are given as a weighted list of terms for which the quality and interpretability may be difficult to assess objectively.

Since after first experiments we found that the topics produced with NMF were more coherent and meaningful, we decided to use NMF for our analysis.

The suitable numbers of topics was determined by building many topic models and computing for each the average coherence of the extracted topics (Fig. 2). Intuitively the coherence of a topic reflects how semantically homogeneous or similar in meaning the top ranking words in the topic are. To assess the semantic similarity of the words we used the *word2vec* word embedding technique [8]. Therein, a neural network-based algorithm represents words from a corpus as real-valued multi-dimensional vectors that explicitly encode the linguistic regularities inherent in the texts.

The best semantic coherence is achieved with 2, 3 and 6 topics. These models are analysed in Section 4.

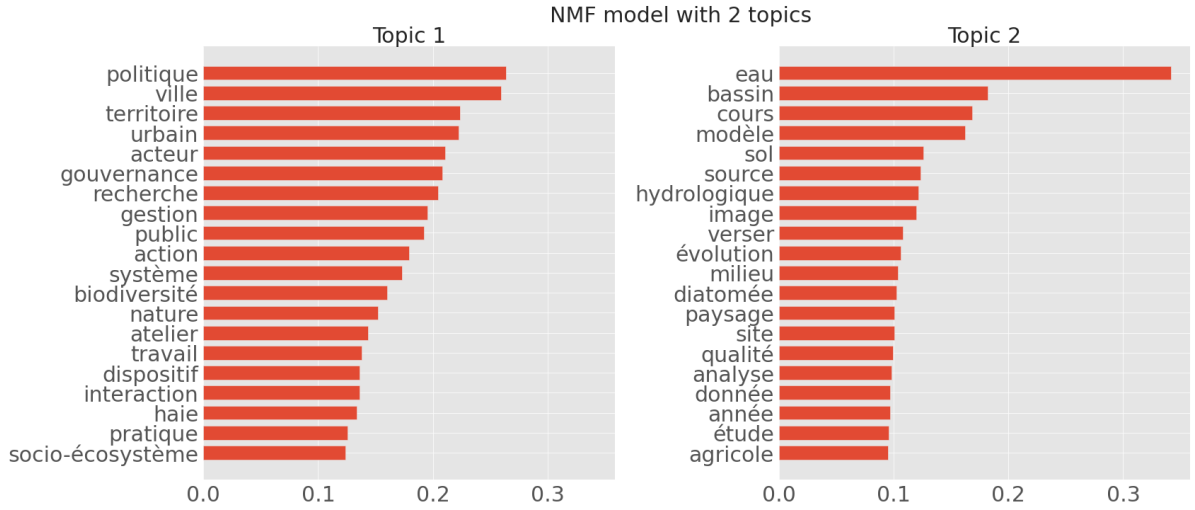

Figure 3: Top weighted terms for topic model with 2 topics.

4. Results and discussion

The 20 terms with the highest weights were computed for the topic models involving 2, 3 and 6 topics. In the discussion, to facilitate readability, key French words were translated to English when roots differ across languages.

Subjects addressed in the conference. The **2-topic model** is presented in Fig. 3. It exhibited the best overall semantic coherence score, suggesting that the corpus may be described with two topics.

Intuitively pairs of terms from each of these topics do not look semantically very similar. The first topic, with *politique*, *ville (city)* and *territoire* as top weighted terms, contains many non-physical terms highlighting social and management aspects that relate to the sociosphere [5], see Section 2. The second topic, with *eau (water)*, *bassin* and *cours (stream)* as top terms, contains many physical terms evoking the hydrological sphere. This was an expected result since half of the Zones Ateliers are named from a river basin, a sea or an ocean, and develop activities in relation with a hydrosystem.

With regards to the literature, the 2-topic model can be considered reminiscent of the social and biophysical templates as conceptualised in the SES framework proposed in Bretagnolle et al. [1]. It also illustrates the dichotomy in the socio-ecological research between natural sciences, and social and political sciences [9] as observed in the European LTSER network.

Figure 4 depicts the top weighted terms for the **3-topic model** which showed the second best overall semantic coherence score. Comparing it with the 2-topic model, we observe that the topic dealing with the hydrosphere (Topic 1) is stable, with almost the same top weighted terms.

Topic 1 in the 2-topic-model (Fig. 3) was split into a topic where societal and political terms are more prominent (Topic 2) and another one (Topic 3) containing terms evoking both an urban environment (*ville*, *urbain*) and some ecology terms (*espèce (species)*, *prédateur*, *biodiversité*), respectively. Their allocation to the hydrosphere, the sociosphere and the biosphere is plausible, although less clear for the last topic.

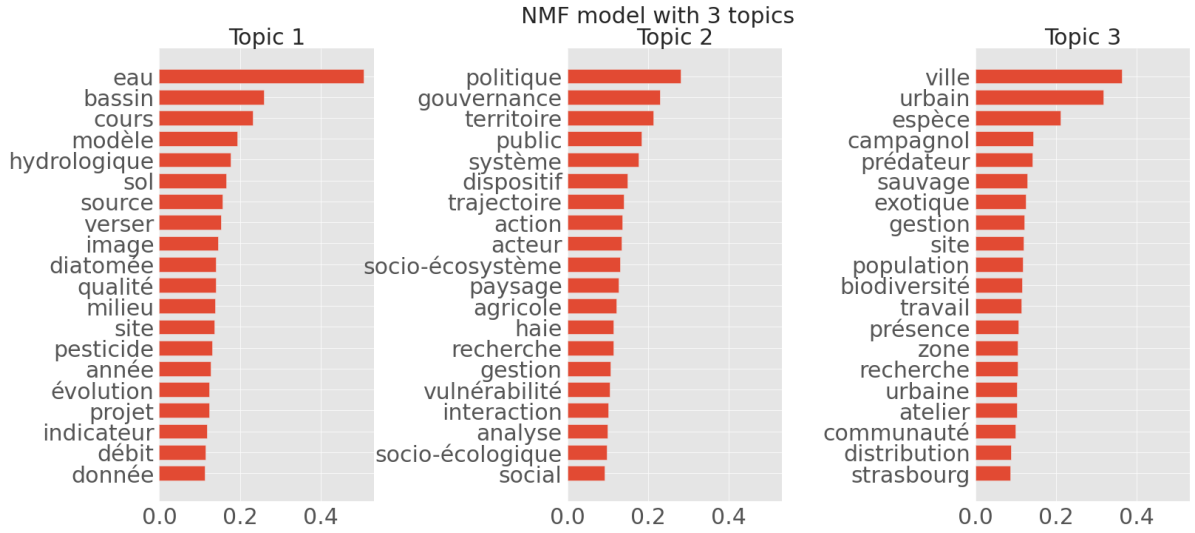

Figure 4: Top weighted terms for topic model with 3 topics.

Finally, Figure 5 shows the top weighted words in the **6-topic model**. For the purpose of this analysis, we denominate its topics based on the top ranking words. Naming topics 1, 2 and 3 is straightforward:

1. **Topic socio-sys** – *politique, gouvernance, territoire, public* – clearly reflects social and management aspects,
2. **Topic hydro-sys** – *eau (water), bassin, modèle, hydrologique* – is about water,
3. **Topic urban-sys** – *ville (city), urbain, espèce* – is concerned mainly with the urban environment,

We found that Topics 4, 5 and 6 deal with various biological or geographical aspects:

4. **Topic land-dyn** – *image, paysage (landscape), evolution*, *dynamique* – is a mixture of image analysis and landscape dynamics or evolution. To name this particular topic we consider the top 10 words since its first three words do not present sufficiently predominant weights. This topic is partly related to the first theme of the conference, the main key word of which is trajectory,
5. **Topic biodiv-pred** – *campagnol (vole), prédateur, population* – exemplifies predation,
6. **Topic agro-eco** – *haie (hedge), bocager, paysage (landscape)* – points out agro-ecology.

In brief, the abstracts of the conference are clearly and mainly about the bio-, hydro- and socio-spheres, and combinations thereof. Some of them could be related to the geosphere through the **Topic 4 (land-dyn)**. None of them really addresses atmospheric research which is mostly considered a "universe science", rather than a natural science by the French National Centre for Scientific Research (CNRS).

To refine the analysis we more closely inspected the abstracts best characterised by the topics using the heat map of Fig. 6. This shows how the topic weights are distributed across the 81 abstracts. The darker a cell is coloured, the larger the weight, and the more clearly the abstract can be associated with the respective topic.

As plausibility checks, we see from the heat map Fig. 6 that the most unequivocal mappings are those abstracts with ids *329437* and *329872* assigned to **Topic 6**. These abstracts stem

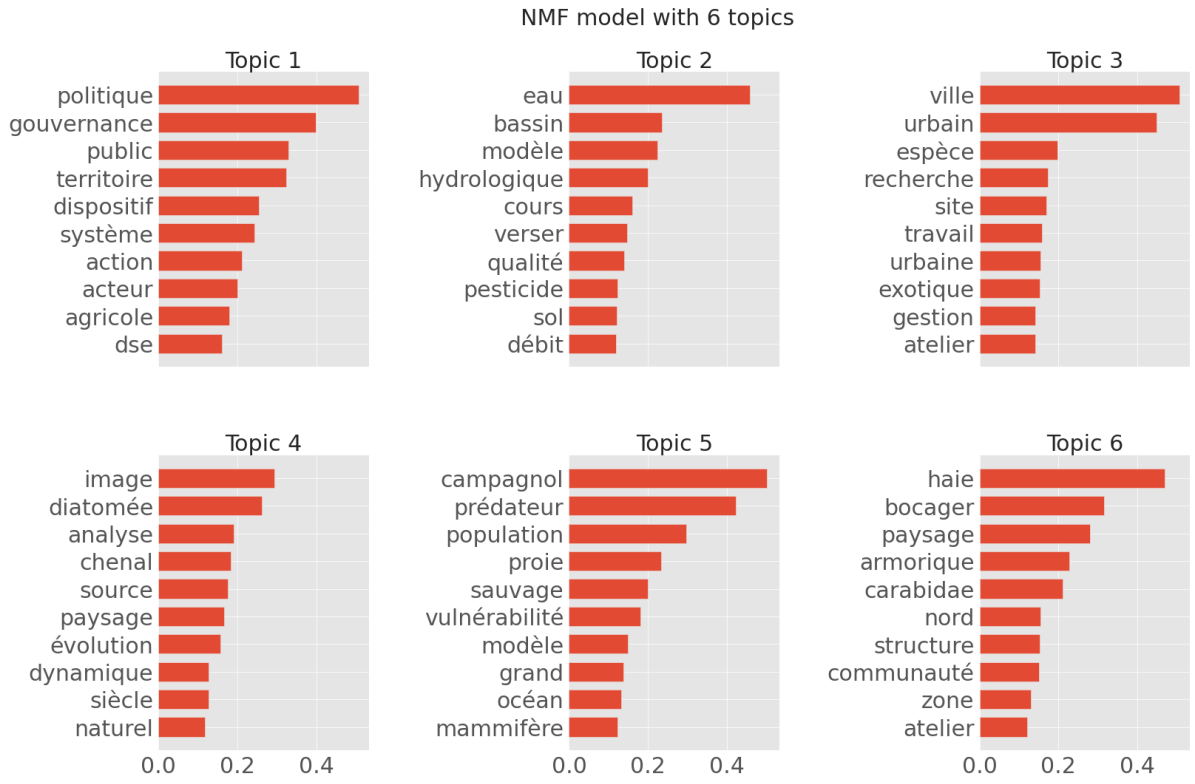

Figure 5: Top 10 weighted terms for topic model with 6 topics.

from the *ZA Armorique* whose main research topic is agro-ecology, thus this association is plausible. The abstract *329872* appears to also address, albeit to a much lesser extent, issues relating to **Topic 1** which evokes social and management topics. Another document where **Topic 6** is a secondary topic is *329857*, which deals indeed with grassland and field ecosystems but is mainly focusing on the influence of public policies. This is consistent with the darker colour for **Topic 1** in our heat map.

Overall, these observations suggest that the 6-topic model is plausible and coherent. We can therefore be confident that it accurately reflects the contents of the conference abstracts.

Inter/Trans-disciplinarity. We use the 6-topic model to analyse to what extent the abstracts presented at the conference were inter- or trans-disciplinary. Indeed, one of the claimed research characteristics in the RZA is that it not only involves different scientific disciplines, like natural and social sciences, but also seeks to be pro-active in the collaboration with the local communities and stakeholders [1] that may be affected by this research.

To get an overview of the topics addressed in each of the abstracts we use the topic weights assigned in an automated manner. A darker colour (*i.e.* a higher weight) in Figure 6 suggests that the corresponding document deals with this topic. Thus for example, the document *331578*'s colour for **Topic 2** (**hydro-sys**) and **Topic 3** (**urban-sys**) is darker, so we consider that it is concerned with these topics.

We assume that a topic is irrelevant for a document if the assigned weight is less than 0.03, moderately relevant between 0.03 and 0.25 and very relevant above 0.25.

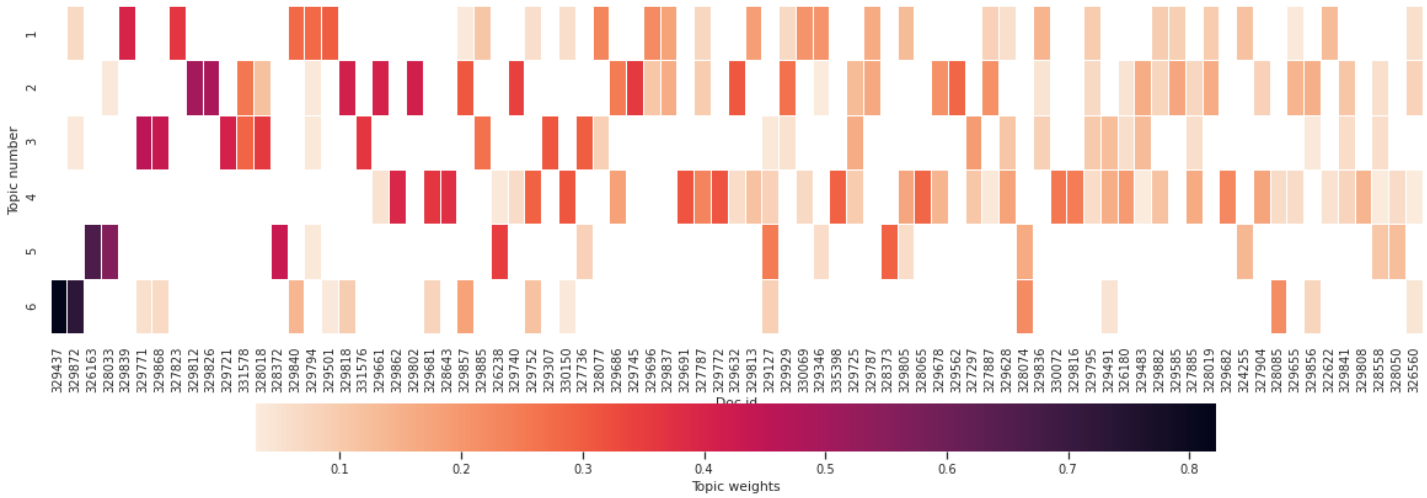

Figure 6: Topic weights computed for documents (81 abstracts). The darker the colour, the larger the weight, the more predominantly the topic characterises the document. For weights < 0.03 topic-document cells are coloured white, in this case the topic is not considered relevant for the document.
Documents (their identifiers) are sorted by increasing residuals. The lower the residual, the stronger the correlation between documents and topics.

As illustrated in Fig. 7, the **socio-sys** subject was addressed in most abstracts (50 out of 81), it was predominant in 5 documents (large weights) and moderately relevant in about 45 abstracts. This topic is the only one that addresses **trans-disciplinarity** in a plausible manner. In particular, the words *acteur*, *territoire* and *public* designate the inhabitants and local stakeholders, through their composition in the expressions *acteur(s) du(des) territoire(s)*. This topic clearly corresponds to the third theme of the conference (Section 2) and highlights research not only involving natural or social disciplines but also people and stakeholders affected by or contributing to *in situ* experimentation.

Our automated analysis found that 31 abstracts (more than 1/3) were not concerned with **Topic 1 (socio-sys)**. This suggests that, while the RZA has a major interest in research in interaction and relation with society [1], there was a substantial proportion of the abstracts that did not really take into consideration societal aspects.

Next we analyse the contributions with respect to their **interdisciplinarity**, here numerically evaluated by looking at the document-topic correlations. Figure 7 lists the 57 documents within at least two topics. In contrast, an important part of the contributions (24, almost 30%) are considered less inter- or transdisciplinary since they are mainly about one topic (Figure 8). A major part of the abstracts, 48, are concerned with the **hydro-sys** topic, as assessed earlier. Compared to the other topics, **biodiv-pred** and **agro-eco** topics were addressed in combination, but with lower weights. The **land-dyn** topic, of more difficult interpretation, represents 42% of the contributions identified as monotopic.

This analysis reveals that a very substantial part ($52,5\%$) of the abstracts involved transdisciplinary activities. With respect to inter-disciplinarity, more than 70% addressed two or more different topics. In most cases, one of the disciplines involved the hydrology topic.

Limitations of the corpus and the methodology The social and biophysical models [1] of the conceptual SES were identified by the topic models, respectively via the trans-disciplinary

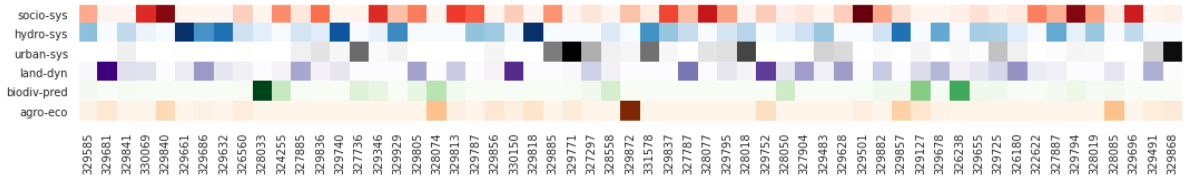

Figure 7: Overview of 57 abstracts with at least two relevant topics.

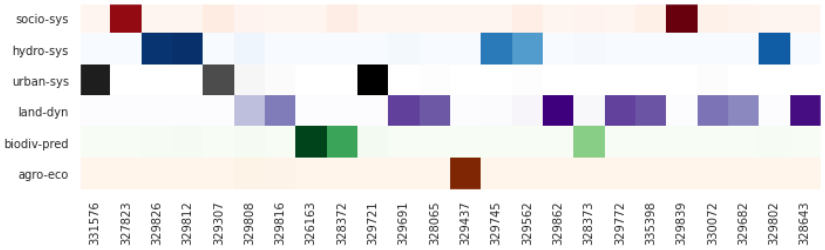

Figure 8: Overview of 24 abstracts with only one relevant topic.

Topic 1 socio-sys, and topics 2, 3, 5 and 6 for different ecosystems. However, it is worth noting that the so-called *adaptive management interface*, related to resilience, and *ecosystem services interface* were not identified.

Several explanations may be raised. Firstly, very few abstract were proposed on the resilience theme (Section 2). Secondly, the word *service* is very ambiguous in French (as it has many different meanings). Used in many abstracts, it is not discriminant enough to be associated to a unique topic. It often appears in bigrams (*e.g. services ecosystémiques*) where it designates more precise terms, but had no effect since the topic model only used unigrams. Thirdly, as with other computational methods, the grouping of terms into topics does not necessarily agree with the nested multi-purpose themes *a priori* defined by the organisers.

Therefore, in the future we plan to investigate how to sensibly integrate bigrams and trigrams (and more generally multi-word-expressions) into the topic model. Such an approach would also be meaningful in view of the development of a thesaurus, where we would expect concepts such as *adaptive management* and *ecosystem services*.

The topic model also points out that our corpus may not be appropriate for the construction of a thesaurus. Indeed, the top ranked words in the topics, considered most relevant for this corpus, are mostly general terms not suitable for a thesaurus – which is natural, since abstracts by definition remain summary. To target a thesaurus suitable for indexing data in a FAIR[2] manner, a corpus of metadata sheets would be more appropriate. Combining NLP tools and expert analysis to this end is projected for future work.

5. Conclusion

In this paper we used NLP methods, in particular topic modelling, to gain insights about the subjects of interest to the French long term socio-ecological research network RZA. For this we analysed the thematic structure of a corpus of abstracts from the network's 20th anniversary

[2]Findable, Accessible, Inter-operable and Reusable.

colloquium.

According to this thematic analysis, the RZA was found to go beyond the so-called disciplinary spheres [5], as it carries out trans- and inter-disciplinary studies in the fields of social and natural sciences, on different socio-ecosystems, and notably hydrosystems, urban environments and rural areas (agro-ecology). However, we also saw room for improvement since an important proportion (about 30%) of contributions was mainly focused on one topic.

Overall this work showed that the NLP techniques were very helpful, as they allowed a much more structured and in depth content analysis than the simple content survey provided by a frequency analysis (Fig. 1). However, as expected, the NLP expertise is not sufficient, domain expertise is essential and indispensable, in particular in future work for a thesaurus construction.

We make our code and datasets available to promote future research.

References

[1] V. Bretagnolle et al., Action-orientated research and framework: insights from the French long-term social-ecological research network, Ecology and Society 23 (2019) 10.

[2] R. Kates, T. Parris, A. Leiserowitz, What is sustainable development? goals, indicators, values, and practice, Environment: Science and Policy for Sustainable Development 47 (2005) 8–21.

[3] M. Schoon, S. Van der Leeuw, The shift toward social-ecological systems perspectives: insights into the human-nature relationship, Natures Sciences Sociétés (2015) 166–174.

[4] D. Vachez, Comparative study of thesauri in Environmental Sciences - Best practices in thesaurus design and FAIRification, 2021. URL: https://hal.archives-ouvertes.fr/hal-03264850.

[5] M. Mirtl, E. T. Borer, I. Djukic, M. Forsius, H. Haubold, W. Hugo, J. Jourdan, D. Lindenmayer, W. McDowell, H. Muraoka, D. Orenstein, J. Pauw, J. Peterseil, H. Shibata, C. Wohner, X. Yu, P. Haase, Genesis, goals and achievements of long-term ecological research at the global scale: A critical review of ilter and future directions, Science of The Total Environment 626 (2018) 1439–1462. URL: https://www.sciencedirect.com/science/article/pii/S0048969717334204.

[6] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[7] D. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL: http://arxiv.org/abs/1301.3781.

[9] J. Dick et al., What is socio-ecological research delivering? a literature survey across 25 international ltser platforms, Science of The Total Environment 622-623 (2018) 1225–1240.