# Comparing Topic Model Stability Between Finnish, Swedish, English and French

Simon Hengchen[0000−0002−8453−7221], Antti Kanner[0000−0002−0782−1923], Eetu Mäkelä[0000−0002−8366−8414], and Jani Marjanen[0000−0002−3085−4862]

COMHIS
University of Helsinki, Helsinki, Finland,
{simon.hengchen;anti.kanner;eetu.makela;jani.marjanen}@helsinki.fi

## 1 Abstract

In the recent years, topic modelling has gained increasing attention in the humanities. Unfortunately, little has been done to determine whether the output produced by this range of probabilistic algorithms is revealing signal or is merely producing noise, nor how well it performs on other languages than English. In this paper, we set out to compare topic model stability of parallel corpora in Finnish, Swedish, English, and French, and the effect of lemmatisation on those languages.

## 2 Context

Topic modelling (TM) is a well-known (following the work of (6; 7)) yet badly understood range of algorithms within the humanities. While a variety of studies within the humanities make use of topic models to answer historical questions (see (3) for a thorough survey), there is no tried and true method that ascertains that the probabilistic algorithm[1] reveals signal and is not merely responding to noise. The rule of thumb is generally that if the results are interesting and reveal a prior intuition by a domain expert, they are considered correct – in the sense that they are a valid entry point into a humongous dataset, and that the proper work of historical research is to be then manually carried out on a subset selected by the algorithm. As pointed out in previous work (10; 4), this, combined with the fact that many humanistic corpora are on the small side, "the threshold for the utility of topic modelling across DH projects is as yet highly unclear." Similarly, topic instability "may lead to research being based on incorrect foundational assumptions regarding the presence or clustering of conceptual fields on a body of work or source material" (4).

Whilst topic modelling techniques are considered language-independent, i.e. "use[] no manually constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like" (8), they encode key

[1] In this work, we choose to use "topic modelling" as a synonym of Latent Dirichlet Allocation (LDA) (2).

assumptions about the statistical properties of language. These assumptions are often developed with English in mind and generalised to other languages without much consideration. We maintain that these algorithms are not language-independent, but language-agnostic at best, and that accounting for discrepancies in how different languages are processed by the same algorithms is necessary basic research for more applied, context-oriented research – especially for the historical development of public discourses in multilingual societies or phenomena where structures of discourse flow over language borders. Indeed, some languages heavily rely on compounding – the creation of a word through the combination of two or more stems – in word formation, while others use determiners to combine simple words. If one considers a white space as the delimitation between words and disregards punctuation (as is usually done with languages making use of the Latin alphabet), the first tendency results in a richer vocabulary than the second, hence influencing TM algorithms that follow the bag-of-words approach. Similarly, differences in grammar – for example, French adjectives must agree in gender and number with the noun they modify, something that does not exist in English – reinforce those discrepancies. Nonetheless, most of this happens in the fuzzy and non-standard preprocessing stage of topic modelling, and the argument could be made that the language neutrality of TM algorithms rests more on it being underspecified with regard to how to pre-process the language. Previous work has tackled this problem: indeed, (5) studies the effect of stemming and concludes that it either helps or hinders the task, depending of the corpus used. More recently, (9) closely look at the effect of lemmatisation on the interpretability of LDA on a morphologically-rich language, Russian.

In this poster, we set out to test topic model stability across languages with regards to corpus size and the effect of lemmatisation. We do so using a custom-made parallel corpus in Finnish, Swedish, English, and French. By selecting those languages, we have a glimpse of how a selection of different languages are processed by TM algorithms. While concentrating on languages spoken in Europe and languages of interest of our collaborative network of linguists, historians and computer scientists, we are still able to examine two crucial variables: one of genetic and one of cultural relatedness. French and Swedish belong to Indo-European (Romance and Germanic branches, respectively) and Finnish is a Finno-Ugrian language. Finnish and Swedish on the other hand share a long history of close language contact and cultural convergence. Because of this, Finnish contains a large number of Swedish loan words, and, perceivably, similar conceptual systems. English (also a language of the Germanic branch, yet highly influenced by French) will serve as a comparison point between all languages, as it is the language that is the most widely used with TM. By doing so, we go further than related work: we study more than one language, and we use lemmatisation rather than stemming – a more "linguistically-aware" choice.

## 3 Methodology

Building on (4), we use DBpedia (1)'s built-in multilingual graph structure to select entities that exist in all four languages, and extract the content of the `short abstract` entry: generally, a two-to-three-sentence text. Selecting the short abstracts rather than the full content of the corresponding Wikipedia page has the advantage that it "smoothes out" cultural differences: through the reduction of their size to a few sentences, all DBpedia entries have a relatively similar weight in their own respective language corpus as well as across languages.

To explore our hypothesis, we use a parallel corpus of born-digital textual data in Finnish, Swedish, English, and French. Once the corpus, made of 115,547 documents, is constituted, it becomes possible to apply LDA (2) – a parametric topic modelling algorithm that is the most widely used in the humanities.

The resulting models for each language are stored, the corpora reduced in size, LDA is re-applied, the models are stored, corpora re-reduced, etc. Topic models are compared manually between languages at each stage, and programmatically between stages, for all languages. The same workflow is then applied to the lemmatised version of the above-mentioned corpora, and results compared across languages, sizes, and linguistic preprocessing.

# Bibliography


[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. Springer (2007)

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

[3] Brauer, R., Fridlund, M.: Historicizing topic models, a distant reading of topic modeling texts within historical studies. In: International Conference on Cultural Research in the context of "Digital Humanities", St. Petersburg: Russian State Herzen University (2013)

[4] Hengchen, S., O'Connor, A., Munnelly, G., Edmond, J.: Comparing topic model stability across language and size. In: Proceedings of the Japanese Association for Digital Humanities Conference 2016 (2016)

[5] Joachims, T.: Learning to classify text using support vector machines: Methods, theory and algorithms, vol. 186. Kluwer Academic Publishers Norwell (2002)

[6] Jockers, M.L.: Macroanalysis: Digital methods and literary history. University of Illinois Press (2013)

[7] Jockers, M.L., Mimno, D.: Significant themes in 19th-century literature. Poetics 41(6), 750–769 (2013)

[8] Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259–284 (1998)

[9] May, C., Cotterell, R., Van Durme, B.: Analysis of morphology in topic modeling. arXiv preprint arXiv:1608.03995 (2016)

[10] Munnelly, G., O'Connor, A., Edmond, J., Lawless, S.: Finding meaning in the chaos (2015)

[11] Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic biology 45(3), 380–385 (1996)