Analysis of the French Parliamentary Debates of the Third Republic with Topic Modelling and Word Embedding. Methodological Challenges and First Results

Aurélien Pellet[1], **Fanny Lebreton**[1 2], **Nicolas Bourgeois**[1], , **Marie Puren**[1 3]
[1]MNSHS-Epitech, [2]Ecole nationale des chartes, [3]Centre Jean Mabillon
[1]Le Kremlin-Bicêtre (France), [2 3]Paris (France)
{aurelien.pellet, nicolas.bourgeois, marie.puren }@epitech.eu, fanny.lebreton@chartes.psl.eu

Abstract

In this paper, we present the work we are carrying out within the AGODA project. One of the objectives of this project is to extract the text of parliamentary debates of the Third Republic, digitised by the Bibliothèque nationale de France, and to analyse these texts with NLP techniques. In particular, we will present the measures we use to measure the quality of the OCR, and the first analyses we have conducted on this corpus.

1. Introduction

For the past sixty years, parliamentary debates have been frequently used by humanities and social sciences Chester1962, Franklin1993. The debates are indeed a valuable source for many disciplines, such as political science (Van Dijk, 2010), sociology (Cheng, 2015) or linguistics (de Galembert et al., 2013; Hirst et al., 2014; Rheault et al., 2016). History has also used these documents (Ouellet and Roussel-Beaulieu, 2003; Marnot, 2000; Ihalainen et al., 2016; Lemercier, 2021), but it is a source that is still too little used in France (Coniez, 2010).

Access to digitised and ocerised debates seems to have a positive effect on the number of historical works using these documents (Bonin, 2020; Mela et al., 2022). The same effect can be observed for other disciplines using textual data from contemporary debates (Fišer et al., 2018; Fišer et al., 2020). The objective of the AGODA project is thus to facilitate access to and use of French parliamentary debates of the Third Republic. AGODA [1] (2021-2022) is one of the five pilot projects supported by the DataLab of the National Library of France[2] (Puren and Vernus, 2021). It aims to create an online platform for consulting and exploring parliamentary debates in the Chamber of Deputies (1881-1940), transcribed in the *Journal officiel de la République française. Débats parlementaires. Chambre des députés : compte rendu in-extenso*, available online on *Gallica*[3], in the form of structured and semantically enriched textual data.

In this paper, we present the NLP analyses we conducted on the debates held between 1881 and 1899. In the framework of the AGODA project, we are mainly interested in the parliamentary cycle from 1889 to 1893[4]; but we apply topic modelling and word embedding on a larger corpus (1881-1899) because both methods (and especially word embedding) require a large amount of text.

2. A corpus of ocerised printed texts

The issues of the *Journal Officiel* available on *Gallica* have been digitised by the National Library of France and the archives of the National Assembly. Between 1881 and 1899, 2596 issues were published, or 50791 images[5]. The digital images of the documents available in JPG format can be downloaded via the *Gallica* API. The debates are also downloadable in TXT format. *ABBY FineReader* automatic transcription (OCR) software was used to extract the text of the debates on the fly, as they were being digitised. The generated text was made available online, but without extensive post-correction.

As shown in figure 2 (see section 3), the quality of the OCR is very variable - and can be particularly poor. Various factors contribute to the high variability in the quality of ocerised texts. The curvature of the page, due to the binding of the registers, has the effect of ”curving“ the text, sometimes even cutting off parts of it or casting shadows on the pages. In addition, the quality of the documents themselves (stains, overprinted text) is also at issue. We know that the quality of OCR output could have a negative impact on the linguistic analyses conducted on these texts (cf. section 4) (van Strien et al., 2020). We therefore considered fully post-correcting the texts; but the very poor quality of some of the OCR outputs and their high overall variability make this task particularly complex. For this reason, we chose to ocerise the parliamentary debates again, in order to obtain less faulty OCR outputs.

[1] Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale.

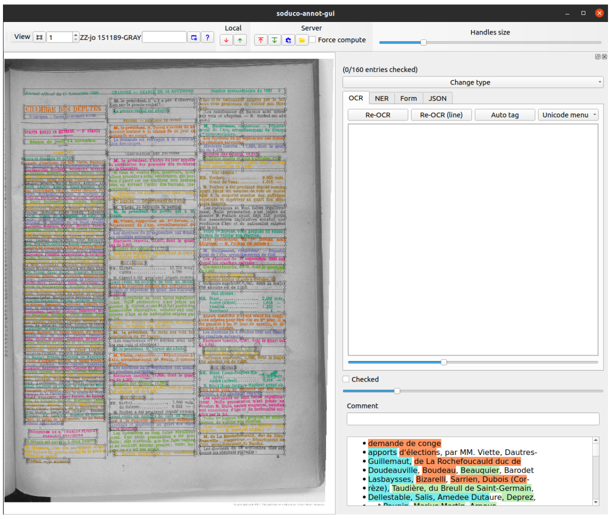
[2] https://www.bnf.fr/fr/les-projets-de-recherche

[3] Available on Gallica

[4] This parliamentary cycle or “5th *législature*” took place between 12 November 1889 and 14 October 1893.

[5] One image corresponding to one page.

To extract the text from the images, we use the OCR tool developed in the framework of the ANR SODUCO project[6]. Figure 1 shows a view of the tool. It should be noted that this tool not only ocerises texts but also recognises named entities (such as speakers' names). OCR is performed using the PERO OCR engine (Kišš et al., 2021; Kodym and Hradiš, 2021; Kohút and Hradiš, 2021), which performs particularly well on historical printed texts. Currently in private alpha version, this tool was used, for example, to prepare the data used in (Abadie et al., 2022). This dataset, which will be freely available on Zenodo [7], consists of texts ocerised from a corpus of printed trade directories of Paris from the XIXth century.



- demande de conge
- apports d'élections, par MM. Viette, Dautres-
- Guillemaut, de La Rochefoucauld duc de
- Doudeauville, Boudeau, Beauquier, Barodet
- Lasbaysses, Bizarelli, Sarrien, Dubois (Cor-
- rèze), Taudière, du Breull de Saint-Germain,
- Dellestable, Salis, Amedee Dutaure, Deprez,

Figure 1: Interface of the OCR tool developed by SODUCO

3. Evaluation of the OCR output quality

As we have pointed out, the quality of the OCR can have a significant impact on the results of NLP analyses: it is therefore particularly important to measure the quality of the OCR we obtain. In this section, we will focus on the comparison of methods for measuring OCR quality.

As a first step, we roughly evaluated the quality of the OCR results obtained via Gallica, by estimating the number of correct words present in the ocerised texts, using the French dictionary[8] provided with the Python library *pyspellchecker*[9]. . Figure 2 shows that the quality of the OCR varies greatly.

This evaluation method has important limitations. Firstly, we are working with a dictionary that is not well adapted to our documents; it is based on a French that

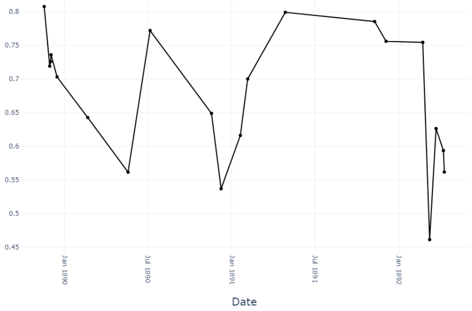


Figure 2: OCR quality evaluation (OCR retrieved from *Gallica*)

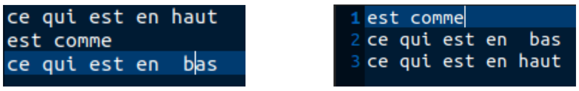
is too contemporary, including, for example, English words such as "*PC*" or "*deal*". Secondly, it is an unsupervised method that is less reliable than supervised methods using ground truth.

To improve our metrics, we are in the process of producing a ground truth, based on 100 randomly selected pages between November 1889 and November 1893. This ground truth will allow us to more accurately assess the quality of the OCR using Levenshtein distance. Levenshtein distance consists of calculating the minimum number of insertions, additions and substitutions necessary to obtain the ground truth from the OCR result. For example, the Levenshtein distance between "*As above se below*" and "*as abve so belo*" is 3.

Thanks to this ground truth, we will then be able to calculate the Character Error Rate (CER) and the Character Accuracy (CA) in OCR :

$$CER = \frac{Lev(text_{gt}, text_{ocr})}{len(text_{gt})}$$
$$CharacterAccuracy = max(0, 1 - CER)$$

The main problem with Character Accuracy is that it is very sensitive to layout analysis for OCR. Layout analysis is one of the main tasks of an OCR engine, but we would like to have a more flexible measure that does not take into account errors in the reading order of the document. This is why we chose to use the Flexible Character Accuracy implementation (Clausner et al., 2020) which is independent from the reading order.



```
ce qui est en haut
est comme
ce qui est en  bas
```

(a) Ground Truth

```
est comme
ce qui est en  bas
ce qui est en haut
```

(b) OCR

Figure 3: Exemple A

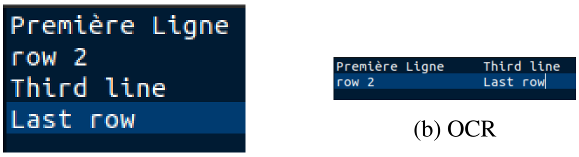
We can see in Table 1 how Flexible Character Accuracy gave an almost perfect result despite the incorrect reading order achieved by the OCR. This is because each block/character was almost always predicted correctly.

[6] https://soduco.github.io/, https://anr.fr/Projet-ANR-18-CE38-0013

[7] https://zenodo.org/record/6394464

[8] Which consists of a list of word frequencies

[9] https://pyspellchecker.readthedocs.io/en/latest/index.html



```
Première Ligne
row 2
Third line
Last row
```

(a) Ground Truth

```
Première Ligne      Third line
row 2               Last row
```

(b) OCR

Figure 4: Exemple B

Exemple	CharAccuracy	FlexCharAccuracy
A	0.45	1
B	0.39	0.95

Table 1: Character Accuracy vs Flexible Character Accuracy

4. Topic Modelling and Word Embedding Applied to Parliamentary Debates

We also wish to facilitate the exploration of these debates by offering new ways to "reading" them (Clavert, 2014). To gain an in-depth understanding of these documents, it is indeed necessary to adopt computational methods to analyse such a large corpus of sources (Pančur and Šorn, 2016; Bonin, 2020). In this section, we will present some examples of lexical analysis that have been performed on the parliamentary corpus (Bourgeois et al., 2022), using the original ocerised text provided by the National Library of France.

4.1. LDA

Latent Dirichlet Allocation is a Bayesian model based on a strong hypothesis (Blei et al., 2003), that fits extremely well our corpus. The underlying model is that there exist hidden variables, namely the topics, which consist of weighted lists of words (the more significant, the higher their probability). Then, every text from the corpus is generated by (1) picking at random a limited number of topics and (2) selecting words from these topics, according to their probability distribution. The role of LDA is to revert this generation process in order to retrieve the original topics, with the hope that their statistical coherence reflects some semantic homogeneity.

The difficulties of LDA include determining the number of topics, ensuring their coherence, naming them and aggregating those who are highly correlated (Newman et al., 2010). We can for example produce a large number of topics (Table 2 shows only 3 of the 40 topics identified), then use an agglomerative clustering to build coherent classes and proof-check them with a qualitative survey. Hence we obtain 15 classes with each a strong identity and limited correlation (Figure 5).

Topic 8	**Topic 11**	**Topic 15**
salaire	général	pari
question	commission	télégraphe
gouvernement	régiment	faire
jour	troupe	ingénieur
patron	monsieur	train
chambre	année	ligne
droit	jeune	chambre
syndicat	temps	personnel
délégué	faire	etat
monsieur	corps	administration
travail	soldat	employé
travaux	ministre	poste
ministre	homme	public
grève	loi	travaux
faire	an	service
mineur	guerre	agent
mine	service	ministre
loi	militaire	fer
compagnie	officier	chemin
ouvrier	armée	compagnie

Table 2: Three topics among 40: the working class (8), the army (11) and the state infrastructures (15).

With our new classes we can also look at their distribution over time - for example by looking at the weight of each topic for every month of our corpus and display it as a time series (Figure 6).

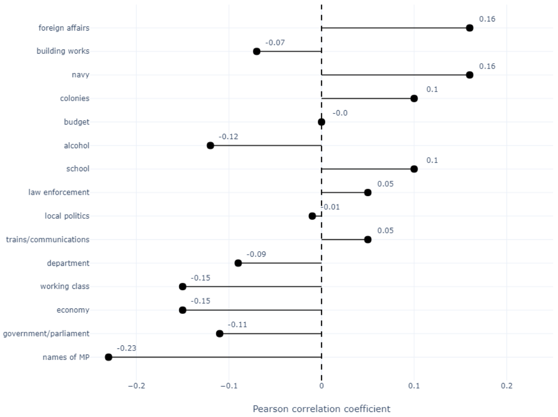


Figure 5: Correlation between the topic "army" and the other identified topics (by month).

The main drawback of this analysis is that a single parliamentary sitting is in fact a rather long text, in which a possibly large sequence of topics are addressed one after the other. It is therefore preferable to divide it into several smaller chunks of texts that better fit the hypothesis of the model. Theoretically, the structure of the document provides a perfect tool for this division, since the different parts of a parliamentary session are easily recognisable by their titles in capital letters.

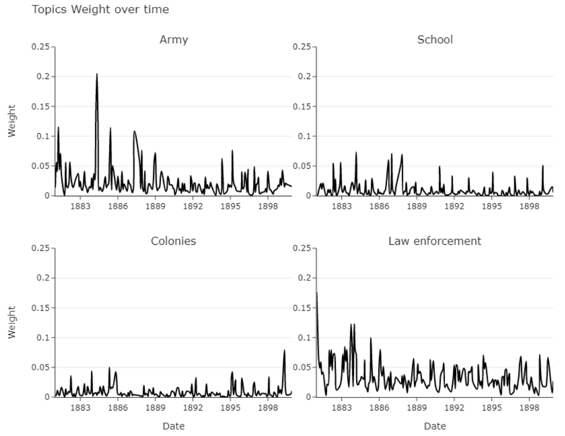


Figure 6: Time evolution of four classes (by month)

However, the recognition of these titles is very imperfect in the original OCR, so we resorted to fixed-length divisions. We hope that as the quality of the text improves, we will be able to use a semantic-based division instead of this arbitrary division.
Another limitation of LDA is that the topics consist of discrete probability distributions over words. This non-continuous method makes it difficult to cluster the topics. Hence to cluster the topics, we relied on the common words between them. However, two topics can be similar while having only a few words in common. For example, a topic about rail traffic and another about maritime traffic both talk about transport, but with different vocabulary. Furthermore, LDA is based only on the word frequency matrix, and therefore does not take into account the order of the words. For LDA, there is no difference between ”*the cat jumped over the fence*“ and ”*the fence jumped over the cat*“ - and yet the difference is is semantically huge. This is why, in addition to LDA, we used word embedding.

4.2. Word Embedding

By definition, Latent Dirichlet Allocation builds a limited number of large semantic units and allow little control over the process. Alternatively, we may use word embedding. Word Embedding consists in transforming words in such a way that they will be interpret-able by algorithm. We do that by assigning one vector to every words in our corpus' vocabulary. to reduce the dimension of the original space from tenth of thousands of forms to a hundred of axes, and then apply classical data science tools such a clustering or correlation analysis on the reduced space (Mikolov et al., 2013). Word embedding has also shown its value in the study of parliamentary debates (Rheault and Cochrane, 2020).

4.2.1. First approach

We first considered using a One-Hot Encode Vector : for a vocabulary of size V, each word is represented by a vector of size V with only 0 except a 1 at a specific location assigned to the word. Unfortunately its a pretty limited method since the word ”*war*“ will be as different from the word ”*wine*“ than it is from the word ”*army*“.

We also examined the use of a singular value decomposition based on co-occurrences matrix. A co-occurrence matrix is a matrix of size VxV, where V is the size of the vocabulary, and each entry in the matrix is 1 if two words co-occur and 0 otherwise. The purpose of this method is to project this matrix into a lower dimensional space to obtain our vector representation. Once again, it has limitations; in particular, it is a time-consuming and computationally intensive method. In addition, we obtain a very scattered matrix since most of the words do not co-occur. We then adopted an iterative method: the algorithm learns the word vectors a little more with each iteration of our model. To do this, we used the Word2Vec algorithm and more precisely CBOW.

4.2.2. Word2vec

CBOW consist in a neural network with one hidden layer (Rong, 2014). Every word is first seen as a one-hot vector and project into a lower dimensional space with a weight matrix average. It is therefore expected to maximise the probability of observing a target word in relation to the other words around it (context word). The words are then transformed into vectors that can best predict the target word which is itself a vector. For example, we want to predict the word ”*cat*“ in the sentence ”*the — jumped over the fence*“. Then a softmax function is applied to minimise an objective function, the cross-entropy. Gradient descent is used to update all parameters. What we are interested in is the weight matrix used to project the words into a lower dimensional space: these will be our embedded words.

Once we have our vectors, we cluster them before displaying them. We still need to apply dimension reduction to get a two dimensional representation. Principal component analysis can be used, but it is limited as it seeks to perform a linear projection. We therefore decided to use the t-SNE algorithm. Technically, this is not a dimension reduction method, but a two-dimensional projection that finds a density probability in the higher dimensional space, and tries to find the nearest density probability in a 2-dimensional space using KL divergence. We can see on Figure 7 how well our different clusters have been separated.

Skip-Gram is an alternative approach. It is the reverse of the previous approach: given a word, we try to predict the best possible context word by assigning vectors to them. Skip-Gram is generally considered to be more accurate, but it takes longer to compute. This is because of the objective function: the probability of observing all words based on a target word takes longer to optimise.

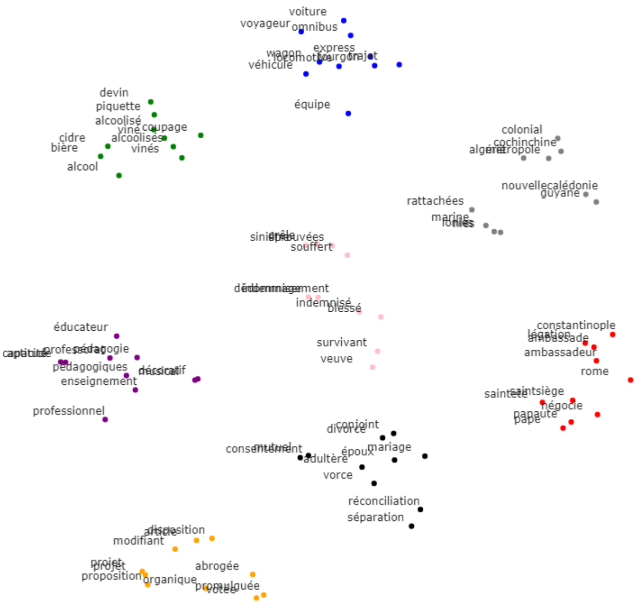


Figure 7: t-SNE projection of the centroïds of the clusters.

4.2.3. top2vec

In addition to word2vec, we used the top2vec algorithm (Angelov, 2020). It starts by embedding documents and vectors together using Doc2vec. Doc2vec extends Word2Vec by adding a document vector that will contribute to all training productions in addition to the word vectors (Le and Mikolov, 2014). Once all our document vectors have been trained, an HDBSCAN clustering algorithm - density-based spatial clustering of applications with noise - is used to group the documents into clusters (Campello et al., 2013). Topic vectors are found as the centroid of the document vectors for each dense area. We then search for the nearest word vectors to obtain the topics.

With top2vec, we obtain a large number (113) of highly coherent clusters (Table 4 shows three of the 113 clusters identified), which we can study in relation to each other, or in relation to other parameters such as time. We can recombine them, for example through agglomerative clustering (Figure 8): with some choices of linkage, we can find superclasses that are very similar to the topic models ; while with others, we get more detailed information about some aspects of the corpus.

Once we have the topic vectors, we can make a more natural clustering based on their vector representations. We can see an example of a cluster of topic obtained. On it seems to capture every topic related to network and transport.

However, word embedding is probably more sensitive than LDA to the quality of the OCR, since a clustering of documents requires that each text belongs to a single class, whereas several topics can be combined. Moreover a more appropriate segmentation of the parliamentary debates - according to the different sequences of a parliamentary sitting - should have a significant impact on the results.

Cluster 55	Cluster 68	Cluster 70
victimes	divorce	enveloppes
inondations	epoux	timbres
secourir	mariage	poste
eprouvees	conjugal	postale
orages	divorces	timbre
sinistres	adultere	recepisses
grele	conjugale	postes
secours	remarier	postaux
venir	separation	telegraphes
infortunes	indissolubilite	colis
ravages	conjoints	fixe
miseres	mutuel	recouvrements
catastrophe	separations	graphes
evenements	mari	postales
repartition	mariages	taxe
incendies	femme	decide
soulager	conjoint	soit

Table 3: Three clusters among 113: storms (55), divorce (68) and the post office (70).

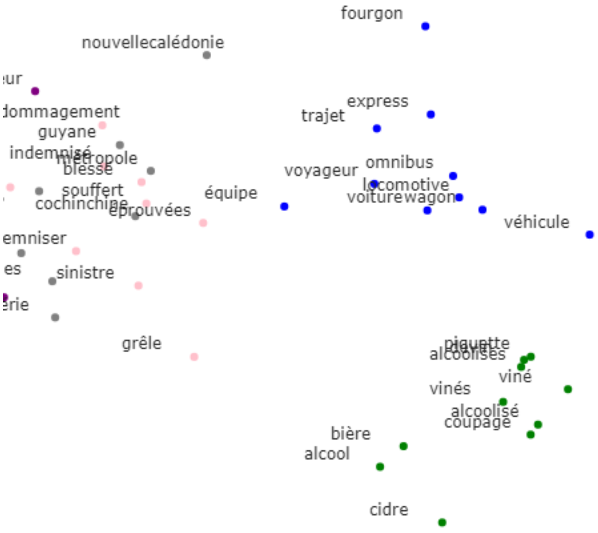


Figure 8: Excerpt from a t-SNE projection of the centroïds of the clusters.

5. Conclusions

The AGODA project aims to analyse texts from parliamentary debates of the Third Republic, based on digitised documents. As in any project working with digitised historical documents, one of the main obstacles is to extract (as clean as possible) text from images. We hope to achieve a very low error text at the end of the project, by combining both re-ocerisation of the documents and post-correction of the texts. We also hope

Cluster 55	Cluster 68	Cluster 70
paquebots	trains	strategiques
postal	freins	loulan
escale	wagons	ligne
messageries	mecaniciens	chemins
antilles	signaux	timbre
transatlantiques	train	tronçons
west	voyageurs	kilomètres

Table 4: A cluster made a three topics

to improve the results we obtain with topic modeling and word embedding. Although these "bag-of-words" techniques are not as sensitive to OCR quality as specific tasks such as name entity recognition, there is a strong incentive to use corrected text to perform natural language processing (van Strien et al., 2020; Mutuvi et al., 2018).

6. Acknowledgements

We would like to thank the National Library of France and Huma-Num for their support in the framework of the BnF DataLab.

7. Bibliographical References

Abadie, N., Carlinet, E., Chazalon, J., and Dumenieu, B. (2022). A Benchmark of Named Entity Recognition Approaches in Historical Documents. Application to 19th Century French Directories. DAS 2022 15th IAPR International Workshop on Document Analysis Systems `https://das2022.univ-lr.fr/`, La Rochelle.

Angelov, D. (2020). Top2vec: Distributed representations of topics.

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

Bonin, H. (2020). From antagonist to protagonist: 'Democracy' and 'people' in British parliamentary debates, 1775–1885. *Digital Scholarship in the Humanities*, 35(4):759–775.

Bourgeois, N., Pellet, A., and Puren, M. (2022). Using Topic Generation Model to explore the French Parliamentary Debates during the early Third Republic (1881-1899). In *DHNB 2022 – Digital Humanities in Action - Workshop "Digital Parliamentary Data in Action*.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Jian Pei, et al., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse Society*, 26(5):562–586.

Clausner, C., Pletschacher, S., and Antonacopoulos, A. (2020). Flexible character accuracy measure for reading-order-independent evaluation. *Pattern Recognition Letters*, 131:390–397.

Clavert, F. (2014). Vers de nouveaux modes de lecture des sources. In *Le temps des humanités digitales*. FYP EDITIONS.

Coniez, H. (2010). L'Invention du compte rendu intégral des débats en France (1789-1848). *Parlement[s], Revue d'histoire politique*, 2(14):146–159.

de Galembert, C., Rozenberg, O., and Vigour, C. (2013). *Faire parler le parlement: méthodes et enjeux de l'analyse des débats parlementaires pour les sciences sociales*. LGDJ-Lextenso éditions, Issy-les-Moulineaux.

Darja Fišer, et al., editors. (2018). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Darja Fišer, et al., editors. (2020). *Proceedings of the Second ParlaCLARIN Workshop*, Marseille, France, May. European Language Resources Association.

Hirst, G., Feng, V. W., C. C., , and Naderi, N. (2014). Argumentation, ideology, and issue framing in parliamentary discourse. In A.Z. Wyner E. Cabrio, S. Villata, editor, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Ihalainen, P., Ilie, C., and Palonen, K. (2016). *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books, Oxford, NY.

Kišš, M., Beneš, K., and Hradiš, M. (2021). At-st: Self-training adaptation strategy for ocr in domains with limited transcriptions. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science*, pages 130–146. Cham: Springer.

Kodym, O. and Hradiš, M. (2021). Page layout analysis system for unconstrained historic documents. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021: 16th International Conference*, page 492–506, Lausanne, Switzerland, september. Heidelberg: Springer-Verlag.

Kohút, J. and Hradiš, M. (2021). Ts-net: Ocr trained to switch between text transcription styless. In J. Lladós, et al., editors, *Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science*, pages 130–146. Cham: Springer.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. `https://arxiv.org/pdf/1405.4053v2.pdf`.

Lemercier, C. (2021). Un catholique libéral dans le débat parlementaire sur le travail des enfants dans l'industrie (1840). *Parlement[s], Revue d'histoire politique*, pages 197–208.

Marnot, B. (2000). *Les ingénieurs au Parlement sous la IIIe République*. CNRS histoire. CNRS Editions, Paris.

Mela, M. L., Norén, F., and Hyvönen, E. (2022). Digital Parliamentary Data In Action (DiPaDA 2022). workshop co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022). https://dhnb.eu/conferences/dhnb2022/workshops/dipada/.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Mutuvi, S., Doucet, A., Odeo, M., and Jatowt, A. (2018). Evaluating the Impact of OCR Errors on Topic Modeling. In *Maturity and Innovation in Digital Libraries. 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings*, pages 3 – 14.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Ouellet, J. and Roussel-Beaulieu, F. (2003). Les débats parlementaires au service de l'histoire politique. *Bulletin d'histoire politique*, 11(3):23–40.

Pančur, A. and Šorn, M. (2016). Smart big data : Use of slovenian parliamentary papers in digital history. *Contributions to Contemporary History*, 56(3):130–146.

Puren, M. and Vernus, P. (2021). AGODA : Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale. Inauguration du BnF DataLab.

Rheault, L. and Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Rheault, L., Beelen, K., Cochrane, C., and Hirst., G. (2016). Measuring Emotion in Parliamentary Debates with Automated Textual Analysis. *PLoS ONE*, 11(12).

Rong, X. (2014). word2vec parameter learning explained. https://arxiv.org/abs/1411.2738/.

Van Dijk, T. A. (2010). Political identities in parliamentary debates. european parliaments under scrutiny. In Cornelia Ilie, editor, *European Parliaments under Scrutiny: Discourse strategies and interaction practices*, pages 29–56. John Benjamins Publishing Company.

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 21:484–496.