Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 17–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Dating Ancient texts: an Approach for Noisy French Documents

Anaëlle Baledent[1, 2], Nicolas Hiebel[1], Gaël Lejeune[1]
[1] Sorbonne University, STIH - EA 4509, Paris, France ;
[2] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
anaelle.baledent@unicaen.fr, nicolas.hiebel@etu.sorbonne-universite.fr, gael.lejeune@sorbonne-universite.fr

Abstract

Automatic dating of ancient documents is a very important area of research for digital humanities applications. Many documents available via digital libraries do not have any dating or dating that is uncertain. Document dating is not only useful by itself but it also helps to choose the appropriate NLP tools (lemmatizer, POS tagger ...) for subsequent analysis. This paper provides a dataset with thousands of ancient documents in French and present methods and evaluation metrics for this task. We compare character-level methods with token-level methods on two different datasets of two different time periods and two different text genres. Our results show that character-level models are more robust to noise than classical token-level models. The experiments presented in this article focused on documents written in French but we believe that the ability of character-level models to handle noise properly would help to achieve comparable results on other languages and more ancient languages in particular.

Keywords: Old documents, Text Mining, Document Dating, Corpus, Digital Humanities, Textual Document Dating

1. Introduction

Nowadays, a large number of historical documents is accessible through digital libraries among which we can cite EUROPEANA [1] or GALLICA [2] among other Digital Humanities (DH) digitization projects. This allows libraries to spread cultural heritage to a large and various audience (academics, historians, sociologists among others). It is also a great opportunity to have such an amount of data usable in various projects including NLP projects.

However, exploiting these documents automatically can be difficult because of the their various quality, their imperfect digitization, the lack of metadata or the fact that they exhibit a great variety of languages (among which under-resourced languages). Many documents will be difficult to access for researchers since it is difficult to unite them in a corpus, to rely on consistent metadata or to use NLP tools if the data is too noisy.

In particular, it is difficult for DH researchers to use most of available data since the quality of the Optical Character Recognition (OCR) on ancient documents can make them impossible to process properly with classical NLP tools. Therefore, pre-processing and data cleaning is often mandatory to make them suitable for classical NLP pipelines. This need increases the cost of treating new corpora for DH researchers since choosing the appropriate NLP tools can even be difficult. The problems encountered can vary with respect to the languages used in the document or the period were the document has been printed but it remains an open problem. Therefore, the knowledge of the date of the document is not only useful by itself but also because it helps to choose the appropriate OCR configuration (Cecotti and Belaïd, 2005), the post-processing techniques after the OCR phase (Afli et al., 2016) or the appropriate NLP processing tools to use for a particular corpus (Sagot, 2019). Hence, we propose in this paper to investigate the problem of document dating in noisy documents.

The contribution of this paper is three fold : (I) we propose a corpus of around 8,000 ancient documents in French (published from 1600 to 1710), (II) we propose some methods to enrich the metadata and (III) we propose new ideas to evaluate the quality of digitized data in order to put the DH researcher in the center of the loop. In the experiments part we will focus on the document dating task but we believe that the corpus we developed and the rationale of our methods can be useful for other tasks.

In Section 2. we present related work on corpus construction and document dating. In Section 3. we present the corpus made available with the article and in section 4. we show some results on document dating on this corpus and compare our method with other state-of-the-art datasets. Finally in Section 5. we give some words of conclusion and present future orientations of this work.

2. Textual Document Dating

In this work we try to tackle the problem of document dating in the context of historical textual documents. One way to tackle this task is to define it as a classification task, each year (or another time granularity) being a class. (Niculae et al., 2014) proposed a text ranking approach for solving document dating. Temporal language models for document dating use mainly a token-level representation. (Popescu and Strapparava, 2013) develop the hypothesis that period changes come with topics changes and written information reflect these changes by used vocabulary. So, one can delimit epochs by observing the variation in word frequencies or word contexts like in recent works about semantic change (Hamilton et al., 2016).

In the same fashion, (de Jong et al., 2005) and (Kanhabua and Nørvåg, 2008) used probabilistic models: the authors assign each word a probability to appear in a time period. Semantic change is therefore leveraged to give a time stamp to a given document. Some authors proposed graph models to extract relationship between events related in the document in order to find the document focus time (Jatowt et al., 2013) or compute an appropriate time stamp for the document(Mishra and Berberich, 2016). Another interesting approach comes from (Stajner and Zampieri, 2013) who used

[1] https://www.europeana.eu/
[2] https://gallica.bnf.fr/

four stylistic features to find appropriate document dating: average sentence length, average word length, lexical density and lexical richness.

Several works on the subject of document dating involved preprocessing of texts (e.g. tokenization, morphosynctatic tagging or named-entity recognition) or external resources, like Wikipedia or Google Ngram in order to detect explicit features that can characterize the date of a document : named entities, neologisms or to the contrary archaic words ((Garcia-Fernandez et al., 2011); (Salaberri et al., 2015)) However, this implies to have access a clean plain text, or a text without too much OCR errors in order to apply data cleaning techniques. Indeed the majority of works exploits newspapers' articles, due to facility for collect them on web and a high precision for dating, and few works use digitized documents. In Section 3. we show how corpus construction can be an issue for these token-level models and why the corpus we wanted to process can be too noisy for them.

3. Corpus and Methodology

3.1. Corpus Construction

Corpus construction is a crucial aspect in Computational Linguistics (CL) and Digital Humanities (DH) fields: the corpus construction is one of the first steps in research. To obtain relevant results, the used corpora must meet specific criteria: genre, medium, topic among other criteria (see (Sinclair, 1996) or (Biber, 1993) for other criteria examples). It must also be adapted with research objectives: a classification task doesn't require same data that a literary analysis. Another question regarding corpus construction is the following: what NLP tools can be used for processing the corpus ?

With Internet one can easily access to a huge amount of texts and corpora. Despite this, researchers must be careful with the data sources : quality, authenticity, noisiness. Barbaresi (Barbaresi, 2015) mentions inherent problems with a web scrapper method to collect corpus: repeated and/or generated text, wrong machine-translated text, spam, multilanguage documents or empty documents. Documents exhibiting this kind of problems can impair the efficiency of classifiers or other NLP modules and force researchers to rebuild a new corpus or to clean the data manually.

Digital libraries provide many and various textual archives, easy to collect and often used in Digital Humanities in view of topics. Indeed, these corpora are also diversified that domains in Humanities and Social Sciences (HSS): 19th century newspapers, middle-age manuscripts or early modern prints,(Abiven and Lejeune, 2019).

However, these documents are not "born-digital" and are often available only in image format. The quality of the text one can extract from these images is far from perfect. So, OCR performances are lower than one can expect on a modern document and this deterioration has an impact on the usability of the data. Several works like (Traub et al., 2015) or (Linhares Pontes et al., 2019) showed that OCR errors has an important impact on NLP tools efficiency and subsequent expert analysis.

Therefore, correcting automatically OCR has become an important prior task to take more advantage of digitalized corpora ((Barbaresi, 2016) (Rigaud et al., 2019)). Automation of this post-processing may reduce financial and temporal costs as compared to manual correction. It is a great challenge for Digital Humanities since these costs can in some cases constitute the biggest part of DH projects budget.

Decade	# Docs (Ratio)	Mean size (± stdev) Characters	Words
1600	389 (5%)	24117 (± 25449)	3702 (± 3698)
1610	649 (8%)	20861 (± 21421)	3248 (± 3223)
1620	926 (12%)	18979 (± 18437)	3033 (± 2727)
1630	917 (12%)	20691 (± 22471)	3304 (± 3339)
1640	815 (10%)	21692 (± 20791)	3558 (± 3271)
1650	583 (7%)	28877 (± 27754)	4725 (± 4306)
1660	552 (7%)	33739 (± 26172)	5698 (± 4266)
1670	489 (6%)	29887 (± 22052)	5150 (± 3655)
1680	630 (8%)	28355 (± 21519)	5023 (± 3677)
1690	802 (10%)	29554 (± 23751)	5276 (± 4106)
1700	791 (10%)	34302 (± 30191)	5928 (± 5030)
1710	427 (5%)	31620 (± 29799)	5461 (± 5151)
All	7970	26276 (± 24577)	4407 (± 3998)

Table 1: Statistics on the GALLICA dataset

3.2. A Dataset for Document Dating

The corpus we mainly use for our experimentations has been collected on the French digital library GALLICA. From GALLICA it is possible to access to a large amount of digitized historical and various documents and we wanted to see how we can apply NLP techniques to old documents were the OCR makes a lot of errors. Some textual documents have also plain text access, in fact a non corrected OCR output.

On the GALLICA website, advanced search's tab allows a search with different filters like date of publication, language, type of document or theme. For this experiment, we selected all Latin and French documents with plain text access and dated between 1600 and 1720. It represents about 8,000 documents. With the search API we exported a research report in CSV format and transformed it in a JSON file. Each document has an unique identifier and has metadata among which title, author(s), editor, date and other descriptions[3].

We took advantage of this research report to download all the documents in HTML. We developed a tool that scrapes the text and sorts the documents according to different kinds of metadata[4]. Four versions for each text are extracted by this tool in order to fulfill different needs : (i) plain text with dates inside the documents; (ii) plain text where dates have been removed (with regular expressions); (iii) text with HTML tags and dates; (iv) text with HTML tags and without date. For assuring that we have the appropriate date for each document, we took advantage of the date indicated in HTML metadata. Documents for which the metadata exhibited an uncertain date like *16*, *16??*, *16.*. or a time period (*1667-1669*) have been discarded.

Table 1 exhibits the statistics on the dataset we extracted

[3]Metadata present in the resource associated with this paper

[4]GITHUB repository to be made public

from GALLICA. In order to perform comparisons with other approaches we also used two other corpora of ancient French documents of another period (1800-1950) which had also OCR issues: Deft 2010 challenge on document dating (Grouin et al., 2010) where the objective was to give the good decade for a given text.

3.3. Training a Temporal model

We propose a method that takes advantage of noisy corpus to enrich metadata. The rationale of our method is to be as much independent of pre-processing steps because the lack of language dedicated resources (few NLP tools exist for ancient languages and their efficiency can be put into question). This can help DH researchers to process more easily new datasets since models robust to noise can avoid research projects to use too much resources in data preparation. For the GALLICA corpus we split the data into a training set (70%) and a test set (30%) and maintained the imbalance between the different classes. For the DEFT2010 corpora, the data was already separated between train and test so we kept it in order to ease comparisons with previous approaches.

We aim to find models suitable for noisy data so we got inspiration from recent works that showed that character-level models perform well for document dating (Abiven and Lejeune, 2019). We compare character-level representation to word-level representations in order to assess their respective advantages. We present our first results in Section 4..

4. Evaluation

In this Section, we first present results on the the Gallica dataset, then we use the exact same configuration to train a temporal model for the DEFT2010 challenge dataset.

4.1. Evaluation Metrics

For evaluation purposes, we use two different metrics. First, we use macro f-measure rather than micro f-measure to compare different models for document dating since the corpus we built from GALLICA is quite imbalanced. Then, since all the classification errors do not have the same impact, in other words when we have a document from 1650 it is better to predict 1640 than 1630, we wanted to have another measure. We choosed to use a Gaussian similarity (here after Similarity), as defined by Grouin *et al.* (Grouin et al., 2011) in order to measure how much there is a difference between the predicted decade and the real decade. It is computed as follows (with pd being the predicted decade and rd being the real decade):

$$Similarity(pd, rd) = e^{-\pi/10^2(pd-rd)^2}$$

This measure has the good property to highlight systems that produce smaller errors: an error of two decades is worst than two errors of one decade (see Table 2 for an excerpt of this similarity measure outcome).

4.2. Results on the GALLICA Dataset

Table 3 shows an extract of the results we obtained. It appeared that Decision Trees give good results and Random Forest (with 10 estimators) even better ones. Character 1-grams give good results and considering longer N-grams improves results until $N = 4$. With $N > 4$ there is no improvement and at some point the results get even worse, this observation is consistent with previous experiments with this kind of features (Brixtel, 2015). Longer N size seems to interfere with generalization. With a random forest classifier and token-level features (token n-grams with $1 <= N <= 3$) we obtained at the best 0.85 in similarity if we discard tokens that include non-alphanumeric characters and 0.93 if we do not discard them. This shows that punctuation, and in general short sequences of characters, are very useful for this kind of task even if they offer worse performances than character n-grams. Another interesting result is that this token-level model achieves only a 46.3% score in macro F-measure. These features exhibit more errors, resulting in a worse F-measure, but the errors are closer to the target.

Figure 1 exhibits the confusion matrix on the GALLICA dataset with our best classifier. One can see that most classification errors are low range errors, this is consistent with the high similarity score the classifier achieves. As presented before, this model outperforms the best token-level model (Figure 2) in F-measure but the difference in similarity is less significant. When comparing the first line of the two confusion matrices one can see that the number of true positives (first cell of the line) is logically higher with the character-level model. However, the false negatives (rest of the line) are in fact very close to the target class, the token-level model shows a bit less errors of 3 decades and more.

$\|pd - rd\|$	0	1	2	3	4	5	6	
SIMILARITY	1	0.97	0.88	0.75	0.60	0.46	0.31	...

Table 2: Similarity measure between pd the predicted decade and rd the real decade

N-gram size	Decision Tree	Random Forest
$1 \leq N \leq 1$	F = 31.62 S = 0.851	F = 35.32 S = 0.877
$1 \leq N \leq 2$	F = 51.23 S = 0.907	F = 58.86 S = 0.931
$1 \leq N \leq 3$	F = 59.49 S = 0.926	F = 66.436 S = 0.947
$1 \leq N \leq 4$	F = 64.6 **S = 0.933**	**F = 71.43** **S = 0.950**
$1 \leq N \leq 5$	**F = 65.1** **S = 0.933**	F = 69.8 S = 0.945
$2 \leq N \leq 2$	F = 51.17 S = 0.905	F = 58.30 S = 0.928
$2 \leq N \leq 3$	F = 59.94 S = 0.927	F = 67.16 S = 0.948
$2 \leq N \leq 4$	F = 64.06 **S = 0.934**	F = 70.53 S = 0.948
$2 \leq N \leq 5$	F = 65.00 **S = 0.934**	F = 70.87 S = 0.948

Table 3: Extract of the results obtained on the GALLICA dataset. Macro F-measure (F) and Similarity (S)

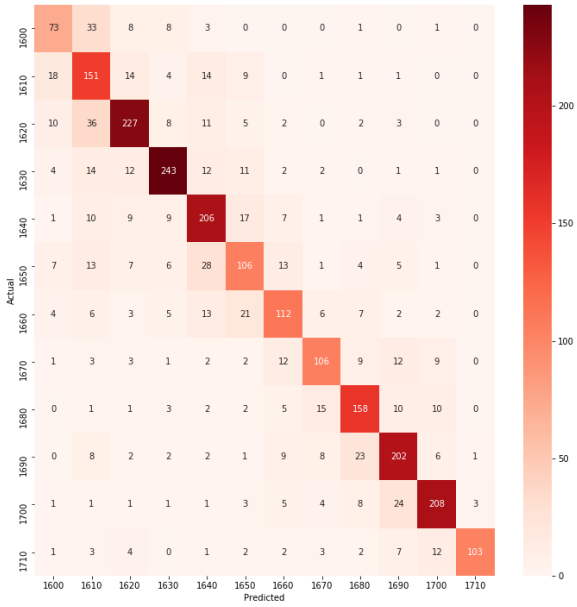

Figure 1: Character-level model (n-grams with $1 <= n <= 4$): confusion matrix for the best classifier (Random Forest with 10 trees) on the GALLICA corpus, F-measure= 71.43, Similarity =0.950

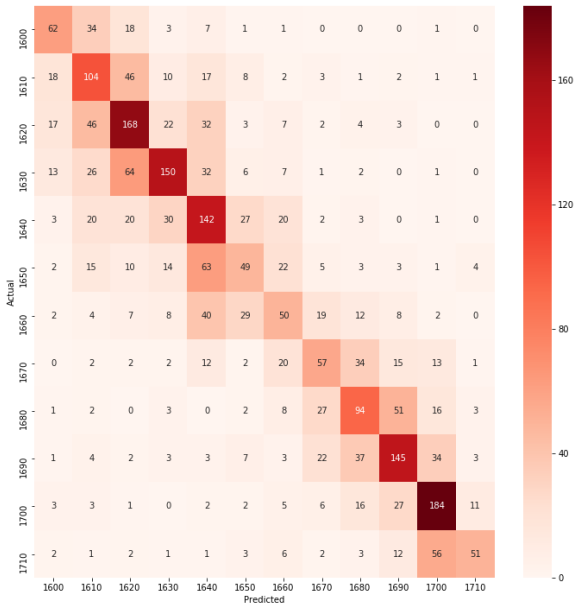

Figure 2: Token-level model (n-grams with $1 <= n <= 2$): confusion matrix for the best classifier (Random Forest with 10 trees) on the GALLICA corpus, F-measure= 46.27, Similarity =0.928

4.3. Results on the DEFT2010 dataset

In Figure 3 we present the results obtained with the same classifier trained and tested on the DEFT2010 dataset. With an F-measure of 32.8 its results are comparable to the best performer (F=33.8) for that challenge which is promising since we did not perform any kind of feature engineering dedicated to this dataset, we just used the same kind of features and the same classifier parameters. We can see

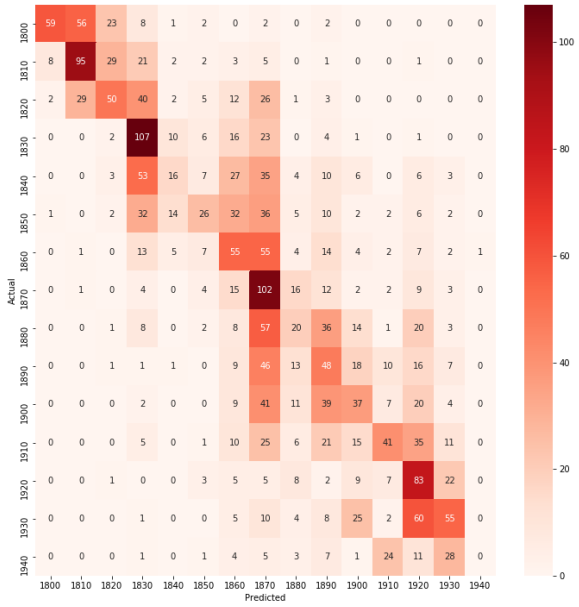

Figure 3: Character-level model (n-grams with $1 <= n <= 4$): confusion matrix for a Random Forest classifier with 10 trees trained and tested on the DEFT2010 dataset, F-measure= 32.81, Similarity =0.872

that most classification errors occur on the previous or next decade. Two interesting things occur however, the 1870 is the most prone to False Positives. It is interesting since this class represent the middle of the period. The 1940 decade does not contain any True Positive. This can be linked to a historical reason since most of the newspapers of this period were not authorized so that there is no clear tendency regarding the printing methods used during this period, illustrating a limit of the character-based models.

5. Conclusion and Perspectives

In this paper we proposed a dataset suited for ancient documents dating. This dataset contains more than 8k documents in French written between 1600 to 1710. The documents in this dataset exhibit a poor quality due to a bad and not post-corrected OCR. Our results show that this should not be a problem for document dating since noise in texts does not seen to impair document dating results. To the contrary, OCR errors seem to be good features to detect the printing time of the original document. We showed that a character-level model can take advantage of noise to improve classification results as compared to a classical token-level model. On a comparable dataset (DEFT2010) from a different time period (1800 to 1940) we show that the exact same features and classifier configuration achieved results close to the state-of-the-art. We believe this is an important result since post-correction of texts can be a very costly operation. This result shows that one can perform NLP task without requiring perfect datasets as input. In the future it would be interesting to see in a larger scope what is the impact of bad digitization on subsequent Natural Language Processing tasks.

6. Bibliographical References

Abiven, K. and Lejeune, G. (2019). Automatic analysis of old documents: taking advantage of an incomplete, heterogeneous and noisy corpus. *Recherche d'information, document et web sémantique*, 2(Numéro 1).

Afli, H., Qiu, Z., Way, A., and Sheridan, P. (2016). Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Barbaresi, A. (2015). *Ad hoc and general-purpose corpus construction from web sources*. Theses, ENS Lyon, June.

Barbaresi, A. (2016). Bootstrapped OCR error detection for a less-resourced language variant. In Stefanie Dipper, et al., editors, *13th Conference on Natural Language Processing (KONVENS 2016)*, pages 21–26, Bochum, Germany, September.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4):243–257, 01.

Brixtel, R. (2015). Maximal repeats enhance substring-based authorship attribution. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 63–71, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Cecotti, H. and Belaïd, A. (2005). Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. In *8th International Conference in Document Analysis and Recognition - ICDAR'05*, pages 1045–1049, Seoul, Korea, August.

de Jong, F., Rode, H., and Hiemstra, D. (2005). Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168. Koninklijke Nederlandse Academie van Wetenschappen, 9.

Garcia-Fernandez, A., Ligozat, A.-L., Dinarelli, M., and Bernhard, D. (2011). Méthodes pour l'archéologie linguistique : datation par combinaison d'indices temporels. In *DÉfi Fouille de Textes*, pages –, Montpellier, France, July.

Grouin, C., Forest, D., Da Sylva, L., Paroubek, P., and Zweigenbaum, P. (2010). Présentation et résultats du défi fouille de texte DEFT2010 où et quand un article de presse a-t-il été écrit ? In *Actes de DEFT*, Montréal, QC, 23 juillet. TALN.

Grouin, C., Forest, D., Paroubek, P., and Zweigenbaum, P. (2011). Présentation et résultats du défi fouille de texte DEFT2011. quand un article de presse a-t-il été écrit ? à quel article scientifique correspond ce résumé ? In *Actes de DEFT*, Montpellier, France, 1er juillet. TALN.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Jatowt, A., man Au Yeung, C., and Tanaka, K. (2013). Estimating document focus time. In *CIKM*.

Kanhabua, N. and Nørvåg, K. (2008). Improving temporal language models for determining time of non-timestamped documents. volume 5173, pages 358–370, 09.

Linhares Pontes, E., Hamdi, A., Sidere, N., and Doucet, A. (2019). Impact of ocr quality on named entity linking. In Adam Jatowt, et al., editors, *Digital Libraries at the Crossroads of Digital Information for the Future*, pages 102–115, Cham. Springer International Publishing.

Mishra, A. and Berberich, K. (2016). Estimating time models for news article excerpts. In *CIKM*.

Niculae, V., Zampieri, M., Dinu, L., and Ciobanu, A. M. (2014). Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 17–21, Gothenburg, Sweden, April. Association for Computational Linguistics.

Popescu, O. and Strapparava, C. (2013). Behind the times: Detecting epoch changes using large corpora. In *IJCNLP*.

Rigaud, C., Doucet, A., Coustaty, M., and Moreux, J.-P. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. In *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, September.

Sagot, B. (2019). Development of a morphological and syntactic lexicon of Old French. In *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France, July.

Salaberri, H., Salaberri, I., Arregi, O., and Zapirain, B. n. (2015). Ixagroupehudiac: A multiple approach system towards the diachronic evaluation of texts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 840–845, Denver, Colorado, June. Association for Computational Linguistics.

Sinclair, J. (1996). Preliminary recommendations on text typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), June.

Stajner, S. and Zampieri, M. (2013). Stylistic changes for temporal text classification. In *TSD*.

Traub, M. C., van Ossenbruggen, J., and Hardman, L. (2015). Impact analysis of ocr quality on research tasks in digital archives. *SpringerLink*, pages 252–263, Sep.