Réutilisation de textes dans les manuscrits anciens

Amir Hazem[1] Béatrice Daille [1] Dominique Stutzmann [2] Jacob Currie [2]
Christine Jacquin [1]
(1) LS2N, 2 Chemin de la Houssinière, 44322 Nantes
(2) IRHT, 40 Avenue d'Iéna, 75116 Paris
amir.hazem@ls2n.fr, beatrice.daille@ls2n.fr,
dominique.stutzmann@irht.cnrs.fr, jacob.currie@irht.cnrs.fr,
christine.jacquin@ls2n.fr

RÉSUMÉ

Nous nous intéressons dans cet article à la problématique de réutilisation de textes dans les livres liturgiques du Moyen Âge. Plus particulièrement, nous étudions les variations textuelles de la prière *Obsecro Te* souvent présente dans les livres d'heures. L'observation manuelle de 772 copies de l'*Obsecro Te* a montré l'existence de plus de 21 000 variantes textuelles. Dans le but de pouvoir les extraire automatiquement et les catégoriser, nous proposons dans un premier temps une classification lexico-sémantique au niveau n-grammes de mots pour ensuite rendre compte des performances de plusieurs approches état-de-l'art d'appariement automatique de variantes textuelles de l'*Obsecro Te*.

ABSTRACT

Text Reuse in Ancient Manuscripts

We address in this paper the issue of text reuse in liturgical books of the middle age. More specifically, we study variant readings of the *Obsecro Te* prayer, part of the devotional Books of Hours. The manual observation of 772 copies of *Obsecro Te* has shown more than 21,000 textual variants. In order to automatically extract and categorize them, we first introduce a semantico-synformic classification at the ngram level, then, we contrast several unsupervised state-of-the-art approaches for the automatic aquisition of *Obsecro Te* variants.

MOTS-CLÉS : Obsecro Te, Livres d'heures, Réutilisation de textes, Variantes textuelles.

KEYWORDS: Obsecro Te, Books of hours, Text reuse, Textual variants.

1 Introduction

La religion chrétienne utilise plusieurs types de livres liturgiques. Empruntant leurs principaux éléments à l'un d'eux, le bréviaire, les livres d'heures sont un recueil de prières à l'usage des fidèles (Leroquais, 1927). Souvent richement enluminés, et répandus dès le 13e siècle en France, au sud des Pays-Bas, en Angleterre et plus tard en Italie et en Espagne, ils constituent une part importante de l'ensemble des manuscrits médiévaux préservés et sont une source d'information sur la vie et la chrétienté au Moyen Âge. Ils reproduisent le contenu de livres réservés aux prêtres et au clergé et permettent aux laïques de prier, comme ceux-ci, selon les heures canoniales. Les livres d'heures ont un noyau en latin et des additions en langues vernaculaires (souvent en français) et font partie des textes les plus lus au Moyen Âge. Malgré leur succès à l'époque, il s'avère aujourd'hui que leur contenu

textuel reste très peu étudié. De plus, il existe très peu de livres d'heures transcrits et annotés. L'une des rares ressources sur le texte des livres d'heures est la base *Beyond Use*, qui contient, en particulier, une section sur l'*Obsecro Te* (Plummer & Clark, 2015). Cette prière à la Vierge a été transcrite et annotée manuellement à partir de plus de 772 livres d'heures [1]. Ainsi, plus de 21 000 variantes textuelles ont été enregistrées. Les variantes sont le résultat d'une opération d'addition, suppression ou substitution au niveau du mot. Une même opération regroupe des opérations linguistiques diverses. Ainsi une opération de substitution peut faire référence, entre autres, à des variantes flexionnelles (*crucem / cruce*), des variantes paradigmatiques obtenues par substitution synonymique (*gratie / indulgencie*). Deux opérations de substitution consécutives peuvent caractériser des variantes de permutation (*opera misericordia / misericordia opera*).
Nous abordons dans cet article la tâche d'extraction automatique de variantes textuelles dans les livres d'heures et plus particulièrement en utilisant *l'Obsecro Te* comme ressource d'évaluation. Ce travail qui constitue une première amorce, a pour but à terme, d'étudier le contenu textuel des livres d'heures afin de découvrir leurs différents usages et de déceler des similarités sur différentes granularités. Ces similarités pourraient servir par exemple à détecter des corrélations structurelles, géographiques et terminologiques entre livres d'heures provenant de différentes régions, d'un même pays ou de pays différents dans cette Europe médiévale. Nous examinons ces différences en proposant, dans un premier temps, une classification lexico-sémantique des variantes au niveau n-grammes de mots, pour ensuite rendre compte des performances de plusieurs approches état-de-l'art d'appariement automatique de variantes textuelles de l'*Obsecro Te*.

2 État de l'art

Un intérêt grandissant pour le traitement automatique du contenu textuel de manuscrits anciens commence à émerger avec un but majeur, qui est celui de pouvoir associer à la fois des analyses historiques et littéraires sur les réseaux textuels (Léonelli, 1985; Stutzmann, 2015; Dondi, 2016). La détection de variantes textuelles constitue un premier pas dans cette direction et plusieurs approches état-de-l'art dédiées à l'alignement et au plagiat par exemple, peuvent être envisagées. La plupart des approches traitant le plagiat ont été proposées et évaluées lors des campagnes PAN [2] de 2009 à 2015 (Belyy *et al.*, 2018). Parmi les approches les plus efficaces, nous pouvons citer celles à base de modèles par plongements de mots (Brlek *et al.*, 2016), celles utilisant les algorithmes génétiques (Kanhirangat & Gupta, 2016; Sanchez-Perez *et al.*, 2018) ou encore, celles à base de modèles thématiques (Le *et al.*, 2016). D'autres approches à base de réseaux de neurones profonds avec des architectures complexes peuvent aussi être envisagées, par exemple les réseaux à convolutions (CNN) (He *et al.*, 2015). Dans ce présent travail, ce type de modèles est difficilement applicable, d'une part, parce qu'il exploite le phénomène de la paraphrase, ce qui n'est pas ou peu le cas concernant les variantes de textes religieux, et d'autre part, le manque de données d'entraînements à disposition ne permet pas de réaliser un apprentissage efficace. Ainsi, nous abordons principalement des méthodes d'alignement classiques à base de similarité de chaînes de caractères et de mots comme la distance d'édition (Levenshtein, 1966) et l'indice de Jaccard (Jaccard, 1901), les approches distributionnelles (Firth, 1957; Harris, 1971) et les approches par plongements de mots (Brlek *et al.*, 2016; Arora *et al.*, 2017).

1. http://www6.sewanee.edu/beyonduse/
2. https://pan.webis.de

3 Variantes textuelles dans *l'Obsecro Te*

Nous présentons une nouvelle catégorisation de variantes inspirée de la similarité lexicale (similar lexical forms ou synforms) introduite par Laufer (1988) et de la typologie de variantes terminologiques proposée par Daille (2017) .

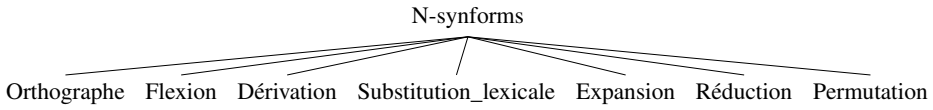
3.1 Similarité lexicale (Synforms)

Le concept de formes lexicales similaires (Similar lexical forms ou synforms en Anglais) a été introduit dans le but d'étudier les confusions lexicales des apprenants de l'anglais (Laufer, 1988). Les synforms sont définis au niveau du mot et sont classés selon différentes catégories de similarité comme des variantes : productives de même racine et de suffixes différents (*considerable / considerate, successful / successive*) ; non productives de même racine et de suffixes différents (*credible / credulous, capable / capacious*) ; ayant des consonnes identiques et des voyelles différentes (*base / bias, manual / menial*) ; ayant des phonèmes identiques à l'exception d'une consonne (*price / prize, extend / extent*), etc.

3.2 Similarité lexicale au niveau des séquences de mots (N-Synforms)

Nous étendons le concept de formes lexicales similaires (synforms) (Laufer, 1988; Kocic, 2008) au niveau des n-grammes de mots. Cependant, nous n'utilisons pas les 10 catégories présentées dans (Laufer, 1988) puisqu'elles ont été construites sur la base des confusions des apprenants de l'anglais. En revanche, nous conservons les catégories s'appliquant au mot seul (unigramme) communes à celle de Daille (2017) et étendons notre jeu de catégories avec certaines opérations linguistiques caractéristiques des termes complexes et s'appliquant aux n-grammes.

Notre observation à la fois des multiples versions de l'*Obsecro Te* et des annotations de celles-ci à l'aide d'opérations d'édition (Plummer & Clark, 2015) nous conduit à proposer une typologie de variantes motivée linguistiquement et pouvant s'appliquer à des séquences de mots de longueur variable. Notre typologie inclut les variations linguistiques classiques opérant sur le mot (orthographe, flexion et dérivation), la substitution lexicale, ainsi que les opérations spécifiques aux séquences de mots (réduction, expansion et permutation). La figure ci-dessous résume notre typologie :



Nous détaillons maintenant nos catégories de variantes :

Orthographe des substitutions de lettres au sein du mot (consonnes ou voyelles) comme *dilecto / delecto* ;

Flexion les flexions en cas du latin, comme *crucem / cruce* ;

Dérivation toute opération de dérivation morphologique pouvant engendrer ou non un changement de catégorie grammaticale, comme *dilecto (Adj)/ dilectissimo (Adj superlatif)* ;

Substitution lexicale toute opération de substitution d'une unité lexicale par une autre. La substitution lexicale permet de générer des variantes en relation de synonymie (*tribuas / concedas*), en relation de quasi-synonymie (*gratie / indulgencie*) mais aussi d'autres variantes sans relation sémantique claire (*tribuas / obtineas*) ;

Expansion les opérations linguistiques d'expansion sont la modification et la prédication comme *criminalibus peccatis / criminalibus peccatis vel mortalibus* ;

Réduction les opérations linguistiques de réduction sont la réduction lexicale et la réduction anaphorique comme *ostendem michi gloriosam / ostendem michi* ;

Permutation la permutation comme *criminalibus peccatis / peccatis criminalibus*.

Bien entendu, comme toute typologie, la nôtre ne prétend pas à l'exhaustivité. Elle pourra être étendue si nécessaire à d'autres opérations linguistiques comme la composition ou la coordination si celles-ci sont rencontrées. Des variantes combinant de multiples opérations, comme des substitutions lexicales associées à des expansions ou des permutations, existent mais elles sont rares.

4 Approches

4.1 Distance d'édition (Levenshtein)

La distance d'édition (Levenshtein, 1966) mesure la proximité entre deux mots x et y en attribuant un score prenant en compte le nombre d'insertions, de suppressions et de substitutions nécessaires pour transformer x en y. Plus le score est élevé, plus le nombre de changements est important et moins les mots sont similaires. Parmi les applications de la distance d'édition, nous retrouvons la détection de plagiat ou la correction orthographique. La formule de la distance d'édition est représentée ci-dessous :

$$D(i,j) = min \begin{cases} D[i-1,j] + SuppCout(i) \\ D[i,j-1] + InsCout(i) \\ D[i-1,j-1] + SubCout(i,j) \end{cases} \quad (1)$$

avec $D(i,j)$ la distance entre les n-grammes i et j, et $SuppCout(i)$ la fonction de coût de suppression de i, $InsCout(i)$ la fonction de coût d'insertion de i et $SubCout(i,j)$ la fonction de coût de substitution de i par j. Pour se ramener à la distance de Levenshtein les trois fonctions de coût sont mises à 1. Nous utilisons cette mesure dans la problématique d'extraction de variantes car certaines variantes latines observées dans *l'Obsecro Te* peuvent être très proches comme *salvatione* avec *salvationis* ou *salvationem*. Dans ce cas, la distance d'édition est très efficace pour détecter ces variantes. Nous obtenons par exemple un score d'édition de 2 entre *salvatione* et *salvationis* (la substitution de la lettre *e* par *i* et l'ajout du *s*) et un score de 1 entre *salvatione* et *salvationem* (1 ajout de la lettre *m*).

4.2 Indice de Jaccard

L'indice de Jaccard (Jaccard, 1901) mesure le degré de similarité entre deux ensembles. Ceci est représenté par le nombre d'éléments en commun entre les deux ensembles divisé par la totalité des éléments des deux ensembles. Plus il y a d'éléments en commun plus le score est proche de zéro

et plus les séquences sont similaires. L'un des avantages de l'indice de Jaccard est qu'il ne prend pas en compte la position des éléments dans les deux séquences. Cette mesure est donc efficace pour détecter les variantes de permutation en leur attribuant un score égal à 0. La formule ci-dessous exprime l'indice de Jaccard

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B} \tag{2}$$

où les deux ensembles A et B correspondent à deux n-grammes de mots, avec B une variante candidate. L'intersection comme l'union sont considérées au niveau du caractère.

4.3 Adaptation de l'approche distributionnelle

L'approche distributionnelle (par sac de mots) classique consiste à représenter chaque mot par son vecteur de contextes (Firth, 1957; Harris, 1971). Chaque contexte représente les mots qui entourent un mot donné selon une taille de fenêtre. Nous adaptons cette approche aux variantes de taille quelconque. Prenons l'exemple suivant : *Levitae autem in tribu* ***familiarum suarum*** *non sunt numerati cum eis*. Le vecteur de contextes de ***familiarum suarum*** comprendra tous les n-grammes qui l'entourent :

— Unigrammes : *Levitae, autem, in, tribu, non, sunt, numerati, cum, eis*
— Bigrammes : *Levitae autem, autem in, in tribu, non sunt, sunt numerati, numerati cum, cum eis*
— Trigrammes : *Levitae autem in, autem in tribu, non sunt numerati, sunt numerati cum, numerati cum eis*
— Quadrigrammes : *Levitae autem in tribu, non sunt numerati cum, sunt numerati cum eis*
— Pentagrammes : *non sunt numerati cum eis*

Une fois les vecteurs de contextes construits, nous effectuons le calcul de mesures d'association, afin de mesurer le degré de la relation contextuelle entre la tête du vecteur (*familiarum suarum* dans l'exemple) et chacun de ses contextes. Trois mesures d'association sont utilisées : l'information mutuelle (IM) (Fano, 1961), le rapport des cotes actualisé (RCA) (Evert, 2005) et le rapport de vraisemblance (RV) (Dunning, 1993). Enfin, pour extraire les candidats, nous mesurons à travers le Cosinus (Salton & Lesk, 1968) la similarité entre le n-gramme source et tous les n-grammes candidats du corpus. Notre adaptation de l'approche par sac de mots prend aussi en compte les n-grammes creux (broken n-grams). Ainsi, en plus des n-grammes déjà cités précédemment, et partant du pentagramme *non sunt numerati cum eis*, nous rajoutons les bigrammes suivants : *non numerati, non cum, non eis, sunt cum, sunt eis, numerati eis*. Ceci en supposant que les unigrammes *sunt, numerati*, et *cum*, aient été absents ou omis du corpus à un moment donné. Cette procédure est répétée pour chaque taille de n-gramme.

4.4 Approche par plongements de mots

L'approche par plongements de mots (word embeddings) consiste à représenter une variante par un vecteur de plongements. Ce vecteur est calculé à partir d'une combinaison linéaire des vecteurs de plongements des mots qui composent la variante (Arora *et al.*, 2017). Si nous reprenons l'exemple *familiarum suarum*, son vecteur de plongements sera l'addition des vecteurs de plongements des mots *familiarum* et *suarum*. Après le calcul des vecteurs de plongements de tous les n-grammes du corpus, nous classons les candidats à l'aide de la mesure du cosinus. La formule ci-dessous présente le calcul par plongements de mots :

$$Embedding(A) = \sum_{j=1}^{n} Embedding(w_j) \tag{3}$$

avec A qui représente un n-gramme de mots et n le nombre de mots le constituant. $Embedding(w_j)$ correspond au modèle de plongements de mots utilisé. Le résultat de la formule correspond à un modèle de plongements de mots représentant le n-grammes A par : $Embedding(A)$. Une variante de ce modèle serait d'utiliser une somme pondérée pour chaque mot du n-gramme (Wieting *et al.*, 2016). Dans nos expériences nous utilisons deux modèles pré-entraînés pour le latin qui sont Word2Vec [3] et FastText [4].

5 Expériences et résultats

Nous avons utilisé la base de données *Beyond Use* [5] qui permet d'étudier les livres d'heures à partir de leurs textes. Cette base fournit une annotation manuelle de variantes textuelles de *l'Obsecro Te* présentes dans 772 manuscrits. Une prière *Obsecro Te* comporte 49 lignes arbitraires définies dans (Plummer & Clark, 2015). Chaque ligne a été comparée et annotée manuellement à la même ligne de la même prière dans les 771 autres copies. À chaque fois qu'une variante est rencontrée, elle est enregistrée comme nouvelle variante dans la base. De ce processus a résulté un corpus d'environ 21 329 entrées d'apparat et 3 298 entrées distinctes. Partant du principe que les informations concernant le type de prière et la segmentation en lignes ne sont pas connues a priori [6], nous n'utilisons pas cette information d'alignement pour extraire les variantes. L'évaluation est faite sur un jeu de tests que l'on divise en 4 listes distinctes. Chaque liste correspond à une taille de n-grammes. Ainsi, nous obtenons une première liste d'unigrammes qui ont comme variantes uniquement des unigrammes, une liste de bigrammes qui ont comme variantes que des bigrammes et ainsi de suite jusqu'aux quadrigrammes [7]. Pour finir, nous rajoutons une cinquième liste notée $Tout$ et qui englobe les quatre précédentes mais qui contient aussi des couples de variantes de tailles variables (environ 23 %).

	Taille des n-grammes (Taille de la liste d'évaluation)																			
	1 (208)				2 (82)				3 (53)				4 (28)				Tout (482)			
Méthodes	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP
DistEdit	**14.0**	**59.1**	**22.6**	48.3	1.82	10.4	3.11	4.65	2.83	8.49	4.24	6.04	2.85	8.06	4.21	5.43	7.01	28.1	11.2	23.1
Jaccard	11.4	50.8	18.7	37.9	7.80	**66.0**	13.9	48.7	**11.3**	**66.0**	**19.3**	38.2	7.85	43.0	13.2	22.8	7.12	35.7	11.8	25.3
BoW (IM)	10.2	46.2	16.8	17.3	5.24	45.3	9.40	12.5	9.24	51.9	15.6	14.8	3.21	15.6	5.33	10.5	2.54	10.8	4.11	8.36
BoW (RCA)	10.1	46.2	16.7	17.1	4.87	41.6	8.73	12.3	9.05	50.1	15.3	14.5	3.21	15.6	5.33	10.5	2.54	10.9	4.12	8.39
BoW (RV)	12.6	52.6	20.3	**48.5**	**8.04**	60.9	**14.2**	28.6	10.7	60.0	18.2	25.7	2.85	17.7	4.78	12.1	**9.70**	**41.7**	**15.7**	**31.9**
Word2Vec	7.74	33.7	12.5	23.3	6.95	63.3	12.4	**62.3**	9.43	65.0	16.4	**49.1**	**12.5**	**64.0**	**20.9**	**40.9**	3.89	21.6	6.60	17.2
FastText	6.39	30.2	10.5	28.7	6.95	60.9	12.4	59.7	9.43	63.9	16.4	41.1	12.1	57.3	20.0	29.0	3.25	19.5	5.57	11.6

TABLE 1 – Évaluation des approches état-de-l'art et notre adaptation de l'approche distributionnelle (BoW). Les résultats sont représentés en termes de précision (P), rappel (R) et F-mesure (F) au top 10 ainsi que la précision moyenne (MAP). Nous affichons entre parenthèses pour chaque longueur de n-grammes, la taille de la liste d'évaluation. Par exemple : 1(208) correspond à 208 n-grammes pour lesquels nous cherchons à obtenir des variantes unigrammes.

Le tableau 1 illustre les résultats des différentes approches implémentées. L'approche par distance d'édition présente les meilleurs résultats lorsque les variantes sont des unigrammes. En revanche, ses performances chutent de manière prononcée dès lors qu'il s'agit de n-grammes supérieurs à

3. http://www.cs.cmu.edu/~dbamman/latin.html
4. https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md
5. http://www6.sewanee.edu/BeyondUse/
6. Dans un cadre applicatif réel nous aurons à disposition une transcription via OCR de livres liturgiques.
7. Nous n'allons pas plus loin car il y a très peu de variantes de taille supérieure à 4.

1. De nombreuses variantes sont des permutations, une opération non appréhendée par la distance d'édition qui est sensible à l'ordre dans lequel apparaissent les éléments d'une variante. L'indice de Jaccard, et même s'il est légèrement moins bon que la distance d'édition, obtient des résultats nettement supérieurs dès lors que l'on passe aux n-grammes supérieurs à 1. Ceci est sans doute dû au fait qu'il ne soit pas sensible au phénomène de permutation. Notre adaptation de l'approche distributionnelle (BoW (RV)) utilisant le rapport de vraisemblance comme mesure d'association montre les meilleurs résultats sur la liste globale ($Tout$). Cette approche est celle qui gère le mieux les couples de variantes de taille variable. Les moins bons résultats de BoW (IM) et de BoW (RCA) montrent que l'information mutuelle (IM) et le rapport des cotes actualisé (RCA) sont moins aptes à capturer l'association des n-grammes dans les vecteurs de contextes. Le manque de données est aussi un facteur qui peut expliquer ces résultats. L'approche par plongements de mots (Word2Vec) montre les meilleurs scores en terme de Map pour les n-grammes supérieurs à 1, ce qui suggère que ce modèle est le plus adapté pour cette configuration (n-grammes > 1). Les moins bons résultats concernant les unigrammes peuvent cependant s'expliquer par le fait que Word2Vec et FastText sont des modèles pré-entraînés et des unigrammes de *l'Obsecro Te* n'y figurent pas. Une combinaison d'approches a été menée mais n'a pas montré d'amélioration significative. Si certains phénomènes peuvent être détectés comme, par exemple, les variantes orthographiques, flexionnelles ou dérivationnelles en utilisant la distance d'édition, les permutations en utilisant l'indice de Jaccard ou encore les substitutions lexicales synonymiques grâce aux approches distributionnelles ou par plongements de mots, d'autres variantes sont plus difficiles à détecter comme les variantes d'expansion ou de réduction, et bien entendu les variantes combinant plusieurs opérations linguistiques. Par ailleurs, nous rencontrons des substitutions lexicales problématiques où certains mots ou séquences de mots sont remplacés par des connecteurs (*et, a, que, de, in...*) comme : *sanctam / et, de filio tuo / a, in omnibus / et in*. L'apparition très fréquente des connecteurs les rend difficiles à modéliser pour un type particulier de variantes.

6 Conclusion

Cet article a présenté une première étude de l'extraction de variantes textuelles latines provenant de livres liturgiques datant de la fin du Moyen Âge. Si des résultats intéressants ont été observés, les méthodes mises en œuvre ne permettent pas de distinguer les variantes orthographiques des variantes flexionnelles ou dérivationnelles. De plus, même les méthodes adaptées à la détection de certaines variantes échouent sur des cas problématiques : la substitution synonymique est peu performante pour détecter les substitutions lexicales où les éléments substitués ont des distributions très différentes comme celles mettant en jeu des connecteurs. Aucune méthode ne s'est révélée efficace pour la détection des expansions ou des réductions. Néanmoins, ce travail constitue une première amorce qui appelle à continuer dans cette voie et qui peut aussi servir à d'autres tâches comme la segmentation des livres d'heures et la découverte de nouvelles connaissances issues de ce type de ressources.

Remerciements

Ce travail s'inscrit dans le cadre du projet HORAE (Hours - Recognition, Analysis, Editions) et a bénéficié d'une aide de l'Agence nationale de la recherche portant la référence ANR-17-CE38-0008. Nous tenons tout particulièrement à remercier le professeur Gregory Clark pour avoir mis à disposition les données de l'*Obsecro Te* et l'annotation manuelle des variantes textuelles.

Références

ARORA S., YINGYU L. & TENGYU M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, p. 1–11.

BELYY A., DUBOVA M. & NEKRASOV D. (2018). Improved evaluation framework for complex plagiarism detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 157–162 : Association for Computational Linguistics.

BRLEK A., FRANJIC P. & UZELAC N. (2016). Plagiarism detection using word2vec model. In *Text analysis and retrieval 2016 course project*, p. 4–7.

DAILLE B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19 of *Terminology and Lexicography Research and Practice*. John Benjamins.

DONDI C. (2016). *Printed Books of Hours from Fifteenth-Century Italy : The Texts, the Books, and the Survival of a Long-Lasting Genre.* Firenze : Leo S. Olschki.

DUNNING T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

EVERT S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. PhD thesis, University of Stuttgart.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communications*. Cambridge, MA, USA : MIT Press.

FIRTH J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, p. 1–32. Oxford : Blackwell.

HARRIS Z. S. (1971). *Structures mathématiques du langage*. Dunod. Traduit de l'Américain par C. Fuchs.

HE H., GIMPEL K. & LIN J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1576–1586 : Association for Computational Linguistics.

JACCARD P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 547–579.

KANHIRANGAT V. & GUPTA D. (2016). Detection of idea plagiarism using syntax - semantic concept extractions with genetic algorithm. *Expert Systems with Applications*, **73**.

KOCIC A. (2008). The problem of synforms. *Facta Universitatis*, **6**(1), 51–59.

LAUFER B. (1988). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, **2**(2), 113–132.

LE H., N. PHAM L., D. NGUYEN D., V. NGUYEN S. & N. NGUYEN A. (2016). Semantic text alignment based on topic modeling. p. 67–72.

LEROQUAIS V. (1927). *Les livres d'heures manuscrits de la Bibliothèque nationale*. Paris.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, **10(8)**, 707–710.

LÉONELLI M.-C. (1985). La dévotion aux saints d'après les livres d'heures confectionnés à avignon. In *Mémoires de l'académie de Vaucluse*, volume 6, p. 327–335.

PLUMMER J. & CLARK G. T. (2015). Obsecro te. *Beyond Use : A Digital Database of Variant Readings In Late Medieval Books of Hours*. http://www6.sewanee.edu/BeyondUse/texts_list.php\?texts=ObsecroTe.

SALTON G. & LESK M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, **15**(1), 8–36.

SANCHEZ-PEREZ M. A., GELBUKH A. F., SIDOROV G. & GÓMEZ-ADORNO H. (2018). Plagiarism detection with genetic-based parameter tuning. *IJPRAI*, **32**(1), 1–23.

STUTZMANN D. (2015). Les écritures des livres d'heures dans l'espace français (1290-1550). In *Proceedings of the 19th Colloquium of the Comité international de paléographie latine*.

WIETING J., BANSAL M., GIMPEL K. & LIVESCU K. (2016). Towards universal paraphrastic sentence embeddings. *International Conference on Learning Representations, CoRR*, **abs/1511.08198**.