Towards Automatic Variant Analysis of Ancient Devotional Texts

Amir Hazem[1] **Béatrice Daille**[1] **Dominique Stutzmann**[2] **Jacob Currie**[2] **Christine Jacquin**[1]
(1) Université de Nantes, LS2N, France
(2) IRHT-CNRS, France
{amir.hazem, beatrice.daille, christine.jacquin}@ls2n.fr
{dominique.stutzmann, jacob.currie}@irht.cnrs.fr

Abstract

We address in this paper the issue of text reuse in liturgical manuscripts of the middle ages. More specifically, we study variant readings of the *Obsecro Te* prayer, part of the devotional Books of Hours often used by Christians as guidance for their daily prayers. We aim at automatically extracting and categorising pairs of words and expressions that exhibit variant relations. For this purpose, we introduce a linguistic classification that allows to better characterize the variants than edit operations. Then, we study the evolution of *Obsecro Te* texts from a temporal and geographical axis. Finally, we contrast several unsupervised state-of-the-art approaches for the automatic extraction of *Obsecro Te* variants. Based on the manual observation of 772 *Obsecro Te* copies which show more than 21,000 variants, we show that the proposed methodology is helpful for an automatic study of variants and may serve as basis to analyse and to depict useful information from devotional texts.

1 Introduction

Among the most popular texts of the late middle ages were Books of Hours, used by Christians as a guidance book for their daily prayers. Appearing in the thirteenth century, in France, the Netherlands, and England and, later on, in Italy, Spain, and many other European countries, Books of Hours constitute one of the bestsellers of the late medieval period. Books of Hours evolved over the years and additional texts were included. Mostly written in Latin, they often include parts in Vernacular languages (esp. French). The whole was arranged in a particular repetitive structure that varied in its details depending on times of the day, seasons, liturgical use, patrons, origin (Wieck, 1988; Hindman and Marrow, 2013), etc.

Despite their success, the content of Books of Hours has been rarely studied on a large-scale in NLP, mainly due to the lack of available transcriptions. few of them are available. One textual element of Books of Hours which offers an opportunity for study is *Obsecro Te*. This devotional prayer to the Virgin Mary was manually transcribed and annotated based on 772 Books of Hours (Plummer and Clark, 2015). More than 21,000 textual variants were recorded. Plummer and Clark (2015) observed and reported three types of variants present in the *Obsecro Te* dataset, that is: (i) addition (marked "+", e.g. "peccatis + vel mortalibus" for *criminalibus peccatis / criminalibus peccatis vel mortalibus*), (ii) substitution (marked ":", e.g. "opera misericordia: misericordia opera" for *opera misericordia / misericordia opera*), and (iii) omission (marked "-", e.g. "-gloriosam" for *ostendem michi gloriosam / ostende michi*). This classification is roughly based on a surface assessment and does not allow a more fine-grained analysis of variants characteristic while no linguistic information is included. In order to study in a more precise way *Obsecro Te* variant readings, we adopt a linguistic classification based on both synformic and conceptual (similar words form) concepts (Laufer, 1988; Daille, 2017). Clark's variants consist in addition, suppression or omission operations at the word level. The same operation groups diverse linguistic operations. Substitution operation for instance, may refer to flexional variants (*crucem / cruce*), paradigmatic variants obtained by synonymic substitution (*gratie / indulgencie*), etc. Also, two consecutive substitution operations may characterise variant permutation (*opera misericordia / misericordia opera*). We conduct an automatic empirical study of the main unsupervised state-of-the-art approaches dealing with variant extraction and discuss our findings according to the proposed linguistic variant classification. Finally, we study variant-relation phenomena and the evolu-

Num	Obsecro Te 1	Obsecro Te 2
1	Obsecro Te domina sancta maria mater dei pietate plenissima summi	Obsecro Te domina sancta maria mater dei pietate plenissima summi
2	regis filia mater gloriosissima mater orphanorum consolatio	regis filia mater gloriosissima mater orphanorum consolatio
3	desolatorum via errantium **salus in te** sperantium virgo ante	desolatorum via errantium **salus et spes in te** sperantium virgo ante
4	partum virgo in partu **et** virgo post partum **Fons misericordie**	partum virgo in partu virgo post partum
5	fons salutis et gratie fons pietatis et leticie fons consolationis	fons salutis et gratie fons pietatis et leticie fons consolationis
6	et indulgencie Per illam sanctam ineffabilem leticiam	et indulgencie **Et** per illam sanctam inestimabilem leticiam
7	qua exultavit spiritus tuus in illa hora quando tibi per gabrielem	qua exultavit spiritus tuus in illa hora quando tibi per gabrielem
8	annunciatus filius dei fuit	**archangelum** annunciatus **et conceptus** filius dei fuit
9	Et per illud divinum mysterium quod tunc operatus est spiritus sanctus	Et per illud divinum mysterium quod tunc operatus est spiritus sanctus **in te**

Table 1: Comparison of the first lines of two *Obsecro Te* variants. Text in red indicates *Obsecro Te* variants.

tion of *Obsecro Te* readings from a temporal and geographical axis and discuss several aspects of Books of Hours.

This work constitutes a first step in the automatic study of Book of Hours content in order to discover the similarities and differences in practices of the middle age. The similarities can for instance serve to detect structural, geographical or terminological correlations between Books of Hours. Whether issued from different regions of the same country or from different countries of medieval Europe.

2 Books of Hours and *Obsecro Te*

Books of Hours contain a set of prayers to be used at eight hours of the day. The structure and content of Books of Hours vary from one book to another and this particularity is certainly due to the nature of textual transmission in a world before the printing press. Books of Hours did not appear as such until the thirteenth century. Before, other types of books were used. For their daily prayers, Christians adopted the Psalter previously used by the Jews for their devotions. Over the years, a number of additional texts came to enrich the Psalter, such as, antiphons, canticles, hymns, readings from the Bible, etc. The whole was arranged in a repetitive structure that varied in its details depending on times of the day and seasons. Also, a calendar was used to record local saints, days and feast's seasons. Finally, rubrics were employed as guidance on what to say and when to say it. This resulted in a complex book known as breviary. The breviary was used by clerks and was not intended to be used by lay people for whom it was too complex. However, the desire of lay people to imitate monastic practices resulted in the creation of a simpler book, that was easier to use: the Book of Hours. Amongst the prayers in Books of Hours is the *Obsecro Te*, a supplication to the Virgin Mary. As the content of a Book of Hours may vary due to writing choice, local liturgical practices, etc., we aim in this paper to study the amount and nature of variants of *Obsecro Te*.

Table 1 shows an example of the first lines of two copies of *Obsecro Te*. Red are the variants according to an arbitrary lines alignment of the two texts (Plummer and Clark, 2015). As highlighted by the passages in red, several variants can be observed. In line 3, for instance, the words "*et spes*" in *Obsecro Te* 1 are added between the words "*salus*" et "*in*" ((Plummer and Clark, 2015) notes that "*salus + et spes*", while in line 4, the words "*et*" and "*Fons misericordie*" are omitted in *Obsecro Te* 2. Also, at lines 7-8, the Annunciation is addressed with the expression *per Gabrielem annunciatus* (*Obsecro Te* 1), while *Obsecro Te* 2 expands upon the passage by specifying the announcer, the *archangelum*, and the effect *et conceptus*. If the reasons of such variants are a matter of interpretation, we aim at depicting the most common ones. For that purpose, we define in the next section our proposed classification of the observed variants before presenting an empirical study for variant extraction and categorisation.

3 Obsecro Te Variant Categorization

We introduce in this section a new variant classification inspired by similar lexical forms (Synforms) introduced in (Laufer, 1988) and the terminological variant typology proposed in (Daille, 2017) applying to nominals.

3.1 Similar Lexical Forms (Synforms)

The concept of synforms was first introduced to deal with lexical confusions of English learners (Laufer, 1988). Synforms are defined at the word level and can be classified on the basis of their

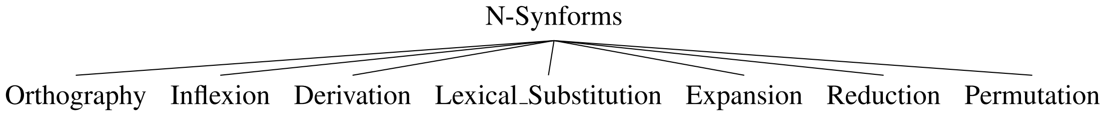

Figure 1: N-Synforms variant representation

similarity features. Words can be different in their affix and similar in their root, different in one phoneme, consonant or vowel. Usually, ten categories including letter addition, substitution and omission, are reported (Laufer, 1988; Kocic, 2008). These categories include: productive synforms with the same root and different suffixes (*considerable / considerate, successful / successive*); non-productive synforms with the same root and different suffixes (*credible / credulous, capable / capacious*); synforms which, although identical in consonants, have different vowels (*base / bias, manual / menial*); synforms with identical phonemes except for one consonant (*price / prize, extend / extent*), etc.

3.2 Lexical Similarity at the Word Ngram Level (N-Synforms)

We extend the similar lexical forms concept (Laufer, 1988; Kocic, 2008) to the word ngram level. Nonetheless, we do not exploit the ten categories presented in (Laufer, 1988), as it deals with confusions of English learners. Therefore, we keep the word level categorization of unigrams as defined in Daille (2017) and extend it using linguistic operations often applied to complex terms and to ngrams. Base on the copies of the *Obsecro Te* prayer and the variant annotations in (Plummer and Clark, 2015), we propose a linguistic representation of variant's typology that can be applied to word ngrams of any length. Our typology includes basic linguistic variants at the word level (orthography, inflexion, derivation), lexical substitution, as well as operations specific to sequence of words (reduction, expansion and permutation). Figure 1 illustrates our typology. We describe the proposed categorization as follows:

Orthography letter substitution (consonant or vowel) like *dilecto / delecto*;

Inflexion latin inflexions like *crucem / cruce*;

Derivation is defined as an operation which creates a new lexical unit from one existing word through modification processes such as affixation or convertion *dilecto (Adj)/ dilectissimo (Adj superlative)*;

Lexical substitution refers to any operation of substitution of a lexical unit by another. Lexical substitution allows variants in semantic relation, such as synonymy (*tribuas / concedas*), near-synonymy (*gratie / indulgencie*) and other variants with no clear semantic relation such as (*tribuas / obtineas*);

Expansion refers to several linguistic operations such as modification which specifies the nominal phrase, predication which inserts the nominal phrase into a nominal argument structure, coordination that emphasize an aspect (*criminalibus peccatis / criminalibus peccatis vel mortalibus*);

Reduction removes one of the lexical constituents of ngrams such as *ostendem michi gloriosam / ostendem michi*;

Permutation of the n-gram elements such as *criminalibus peccatis / peccatis criminalibus*.

Of course, like any typology, ours does not claim to be exhaustive. Nonetheless, it can be extended, if necessary, to other linguistic operations like composition. Also, variants that combine multiple operations like lexical substitution and expansion or substitution exist but they are marginal.

4 Variant Extraction Approaches

We introduce in this section four unsupervised state-of-the-art approaches to the task of variant extraction: Edit distance (Levenshtein, 1966), Jaccard Index (Jaccard, 1901), distributional bag of words (Harris, 1971) and its adaptation to variable length variants extraction and finally, distributed word embeddings (Mikolov et al., 2013; Arora et al., 2017).

4.1 Edit Distance (Levenshtein)

Edit distance, also known as the distance of Levenshtein (Levenshtein, 1966), aligns local similarities and differences between strings and calculates string-alignment. Distance is calculated from the number of necessary operations (insertions, deletions and substitutions) for transforming the string x into the string y. Among the edit distance applications, we find plagiarism detection and orthographic corrections. Edit distance formula is represented as follows:

$$D(i,j) = min \begin{cases} D[i-1,j] + SuppCost(i) \\ D[i,j-1] + InsCost(i) \\ D[i-1,j-1] + SubCost(i,j) \end{cases} \quad (1)$$

where D(i,j) represents the distance between two ngrams i et j and $Suppcost(i)$, $InsCost(i)$ represent respectively the deletion, insertion costs of i. Finally, $SubCost(i,j)$ represents the substitution cost of i by j. When the three cost functions are put to 1, Edit distance is equivalent to Levensthein distance. The use of Edit Distance is based on the observation that several *Obsecro Te* variants may be synformic (graphically similar). For instance, *salvatione* is very close to *salvationis* or *salvationem*. In this case, Edit distance score is 2 between *salvatione* and *salvationis* (the letter *e* is substituted by *i* and the addition of *s*) and a score of 1 between *salvatione* and *salvationem* (addition of the letter *m*).

4.2 Jaccard Index

Jaccard Index (JI) (Jaccard, 1901) measures the degree of similarity between two sets. This is represented by the number of elements in common normalized by the elements of the two sets. One advantage of Jaccard Index is that it is insensitive to element's position and for this reason is not affected by element's permutation. This particularity makes the JI well suited to semantic variants of permutation type, such as *crucifixum vulneratum* and *vulneratum crucifixum*. In this case, JI score is 0 which means that the pair of variants is similar according to permutation property. JI formula is as follows:

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

where the two sets A and B correspond to two word ngrams, with B a variant candidate. The intersection and union are both considered at the character level.

4.3 Distributional Bag of Words

In the distributional Bag of Words (BoW) approach each word w is represented by its context vector (Harris, 1971). The context vector of w gathers all the words with which it appears in the corpus within a size n context window. The context window represents a set of surrounding words often close to the sentence level size. To measure the similarity between words, the cosine (Salton and Lesk, 1968) is applied between the context vector of w and all the word context vectors of the corpus. The closest word to w is a potential variant. We adapt BoW approach and extend it to the ngram level. The procedure remains the same, the main change lying in the context representation of each variant. Let us consider the following example: *Levitae autem in tribu* ***familiarum suarum*** *non sunt numerati cum eis*. The context vector of ***familiarum suarum*** is represented by the following ngrams: Unigrams: *Levitae, autem, in, tribu, non, sunt, numerati, cum, eis*; Bigrams: *Levitae autem, autem in, in tribu, non sunt, sunt numerati, numerati cum, cum eis*; 3grams: *Levitae autem in, autem in tribu, non sunt numerati, sunt numerati cum, numerati cum eis*; 4grams: *Levitae autem in tribu, non sunt numerati cum, sunt numerati cum eis*; and 5grams: *non sunt numerati cum eis*. Once the context vectors have been computed, an association measure is used as a way to better characterize the contextual relation between the head of the vector (**familiarum suarum**) and its constituents. We consider three different association measures: mutual information (Fano, 1961), discounted odds ratio (Evert, 2005) and log-likelihood (Dunning, 1993). Finally, to extract the candidates, we compute cosine similarity (Salton and Lesk, 1968) between all ngrams of the corpus. Our adaptation takes into account broken ngrams. Hence, in addition to the above cited ngrams, based on *non sunt numerati cum eis*, we add the following bigrams: *non numerati, non cum, non eis, sunt cum, sunt eis, numerati eis*. Therefore, we assume that the unigrams *sunt, numerati*, and *cum* may not appear or were omitted.

4.4 Word Embeddings

In the word embedding approach, each variant is represented by an embedding vector which is a linear combination of the word embeddings composing the variant (Arora et al., 2017). For instance,

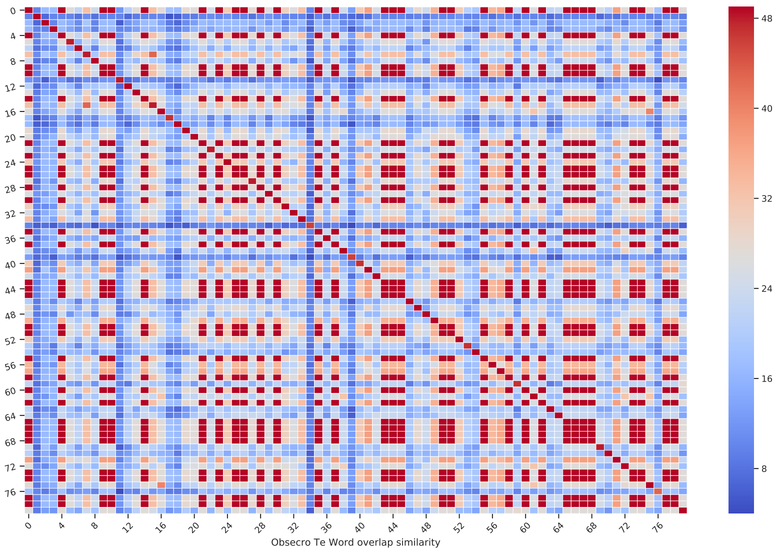

Figure 2: Word overlap similarity of 80 randomly selected *Obsecro Te* texts.

the embedding vector of *familiarum suarum*, is the sum of the word embedding vector of *familiarum* plus the word embedding vector of *suarum*. Finally, after computing the embedding vectors of all the ngrams which compose the corpus, cosine similarity is used to extract the variant candidates. The computation of the embedding vector of a given variant is represented as follows:

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

where A is a variant and n the number of words composing A. $Embedding(w_j)$ corresponds to the chosen embedding model of w_j. We use two pre-trained models: Word2Vec[1] and FastText[2].

5 Experimental Data

To evaluate the automatic extraction of *Obsecro Te* variants, we exploit the *Beyond Use* [3] database which contains variants extracted manually from 772 manuscripts (Plummer and Clark, 2015). The given prayer contains 49 segments (passages) defined arbitrarily. This segmentation allowed Clark to compare each line of the *Obsecro Te*, manuscript by manuscript, and to extract 21,329 variants, of which 3,298 distinct variants. In order to study the impact of variant length, we build four distinct evaluation lists. Each one corresponds to an ngram size. Hence, we obtain a list of unigrams that contains only unigrams as variants; a list of bigrams that contains only bigrams as variants and so on. We do not go beyond fourgrams because very few ngrams are characterized by a length longer than four in the corpus. We finally build a fifth list that contains all the ngrams of the four previous lists as well as ngram variants of any length (20% of the variants have a variant of a different size).

6 Results

Our experimental procedure targets three points: (i) an empirical evaluation of *Obsecro Te* reading similarity; (ii) an empirical evaluation of automatic variant extraction; (iii) a qualitative variant analysis with regard to linguistics, geographic and diachronic changes.

6.1 Similarity of Obsecro Te Texts

There is substantial variation in the text of the prayer *Obsecro Te*. As has been shown in (Wieck, 1988; Plummer and Clark, 2015), the manual analysis of 772 *Obsecro Te* prayers revealed several dissimilarities as well as the existence of more than 21,000 variants. Figure 2 illustrates the simi-

[1] www.cs.cmu.edu/~dbamman/latin.html

[2] github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[3] http://www6.sewanee.edu/BeyondUse/

	Ngram size (size of the evaluation list)																			
	1 (208)				2 (82)				3 (53)				4 (28)				ALL (482)			
Method	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP	P	R	F	MAP
EditDist	**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 (OR)	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 (LL)	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**
W2V	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 2: Evaluation of EditDist, Jaccard, BoW and Embedding approaches (W2V and FastText). The results are presented in terms of precision (P), Recall (R) and Fmeasure (F) at top 10 as well as the mean average precision (MAP). Between parentheses we display, for each ngram size, the size of the evaluation list. For instance: 1(208) corresponds to 208 ngrams (variants) of length 1.

larities between 80 randomly[4] selected *Obsecro Te* texts. The similarity is measured in terms of word overlap. Strong similarities are shown by the dark red colour, while weak similarities by dark blue. Figure 2 shows that none of the 80 sampled *Obsecro Te* texts are identical. This empirical finding confirms the observations of Clark and supports the idea that different copies of the prayer *Obsecro Te* differ substantially from one another.

6.2 Automatic Variant Extraction

In this section, we aim at evaluating unsupervised approaches to variant extraction. Hence, no clue, such as verse or segment alignment, is considered in variant modelling. This leads to the assumption that any ngram extracted from the corpus is a variant candidate. The side effect of this assumption is its error productivity while many ngrams are not variants.

Table 2 illustrates the results of the implemented approaches. Edit distance shows the best results for unigram variants. Nonetheless, its performance significantly drops when variants are of length greater than 1. This can be explained by the large number of permutations that are not identified by Edit distance.

Jaccard Index obtains better results than Edit distance for ngrams greater than 1, which means that conversely to Edit distance, it better handles the permutation phenomenon. Our adaptation of the bag of words approach (BOW (LL)) using log-likelihood shows the best results on the entire evaluation list (ALL). This indicates that BOW (LL) better handles variants of variable length. The lower results of BOW (MI) and BOW (OR) lead to the assumption that these two association measures fail to capture strong ngram association relations. The lack of training data can also explain this behaviour. The word embedding approach (w2v) shows the best Map scores for ngrams greater than 1. This suggests that w2v is the most appropriate when variants are not unigrams. The lower results for unigrams can be explained by the nature of the embedding models. Indeed, w2v and fastText are pre-trained models and many *Obsecro Te* words are not present in these models. Finally, a linear combination of the approaches has been carried out without significant improvements.

Rare Variants	**Category**
salvatione**m** / salvation**e**	inflectional
victori**a** / victoria**m**	inflectional
vi**s**erum / vi**sc**erum	orthographic
dolose / dolore	lexical substitution
gaud**i** / gaudi**o**	inflectional
ancilla tu**a** / famulo tu**o**	lexical substitution + inflectional (f./ m.)
michi annuncies / annuncies michi	permutation
sensum erigat mores **i**mponat / mores **co**mponat	reduction + lexical substitution
Frequent Variants	**Category**
gaudia / gaudio	inflectional (f./ m.)
misericordie / gratie	lexical substitution (Adjective)
domina / virgo	lexical substitution (Noun)
cordis dolorem / dolorem cordis	permutation
a dilecto filio / de filio	lexical substitution + reduction
regat / custodiat	lexical substitution (Verb)
super / per	lexical substitution (Preposition)

Table 3: Examples of extracted *Obsecro Te* variants.

If some phenomena can be detected such as synforms at the word level (with Edit distance for unigrams), permutations using Jaccard index, or lexical substitution using bag of words and embedding vector approaches, other phenomena are more dif-

[4] The number of texts was limited to 80 to enable a clear visualisation of the results. The same behaviour was observed over the entire *Obsecro Te* dataset.

ficult to handle, such as expansion and reduction variants where the two segments are of variable length. We also report that some words and expressions are often substituted by connectors (*et, a, que, de, in...*), such as *sanctam / et, de filio tuo / a, vulneratum / et, in omnibus / et in*. Very frequent connectors represent one of the most difficult variants to extract as they show a big discrepancy of distribution between the two elements.

6.3 Qualitative Variant Analysis

Variant categories can be analysed based on Edit distance, Jaccard Index, BoW and Embedding scores as follow: (i) if Edit distance score is lower than few characters (generally 3), we can effectively pinpoint, thanks to a regular expression, one of the three synformic categories (orthographic, inflectional or derivational); (ii) if Jaccard index score is equal to 0, we face a permutation; (iii) if we combine two criteria, i.e., high Edit distance score and low Jaccard index score, we extract variants that exhibit both expansion and permutation; (iv) lexical substitution variants can be extracted using BoW or Word embedding approaches. A high cosine similarity score has also been used to give more confidence about lexical substitution variants.

Based on the observation that a large number of variants (406) appears in less than five copies of *Obsecro Te*, we divide our analysis into two parts: rare variants and frequent variants. Table 3 reports some examples of variants identified by our automatic extraction. For rare variant pairs (*salvationem / salvatione*, for instance), each reported left side variant appears only in one copy, while its right side counterpart variant appears in hundreds of copies. Rare variants may indicate either a rare usage or a misspelling error. On the other hand, frequent variants may offer a high confidence in their usage.

We observe inflectional variants as rare or frequent variants: *salvationem* (singular accusative) and *victoria* (singular) appear only once in the corpus, while *salvatione* (singular ablative) and *victoriam* (singular accusative) appear respectively 961 and 966 times. In one case the accusative mode is used, while in the other, the ablative is used. Misspellings as rare variants: *viserum* and *vicerum* are both misspellings of *viscerum*. Lexical substitution applies mostly to frequent variants and leads to semantic variants. Synonym lexical substitution, such as *domina /virgo*, is encountered in every grammatical category (noun, verb, adjective, preposition). One exception is the rare variant *ancilla tua / famulo tuo*, the result of two linguistic operations, lexical substitution and inflection, that may refer to a customisation of Books of Hours according to its owner, either a woman or a man. We expect that looking to the whole text of Books of Hours and increasing the number of Books of Hours, this variant will be more frequent. Indeed, Books of Hours are personal objects, and are not intended to be shared. Finally, the last example of rare variants shows the application of two lexical operations, reduction and lexical substitution.

Rare (freq=1)	Frequent (freq >500)
in me instruat (Savoy)	instituat
ancilla tua n (Netherlands)	famulo tuo
sensum sursum dirigat (Paris)	cursum dirigat
famule tue leonarde (Provence)	famulo tuo
aliis rebus quas (Val d'Oise)	illis rebus in quibus
in cruce denudatum (Netherlands)	ante crucem nudatum
siscientem ac hely (Paris)	sicientem fel apponi
mea et desideria (Paris)	et desideria mea
venias et festine (Netherlands)	veni et festina
bene per me (Amiens)	me bene per
omni auxilio consilio (Netherlands)	omni consilio
cursum meum regat (Besançon)	cursum dirigat
scicientem fel aponi (Bourges)	sicientem fel apponi
venias et sustines (Valenciennes)	veni et festina
pace omni salvatione (Besançon)	omni salvatione pace
petitionibus et requestis (Western Fr)	orationibus et requestis
et etiam abundantiam (Val d'Oise)	etiam habundantiam
in omnibus etiam (Central France)	et in omnibus
deus filius tuus (Netherlands)	filius dei
mentem sensum et (Netherlands)	mentem erigat
gratie et salutis (Paris)	salutis et gratie
in ea elevatum (Netherlands)	in ipsa levatum
regat et mentem (Paris)	regat mentem
veni et festinam (Rouen)	veni et festina
probet et vota (Mons)	probet vota
cursum sensum erigat (Paris)	cursum dirigat
honnestam et honnourabilem (Mons)	honestam et honorabilem
venies et festinas (Netherlands)	veni et festina
meum in consilium (Rouen)	et consilium
horam et diem (Netherlands)	diem et horam

Table 4: Examples of 3 gram variants. First column shows variants that appear only once. Column 2 shows the corresponding frequent variants.

Figure 3 illustrates the number of *Obsecro Te* (and therefore, Books of Hours) produced between 1375 and 1530 and used in this experiment[5]. This

[5] Given that the Books of Hours are not dated by their scribes, a date range is generally devised by scholars. The

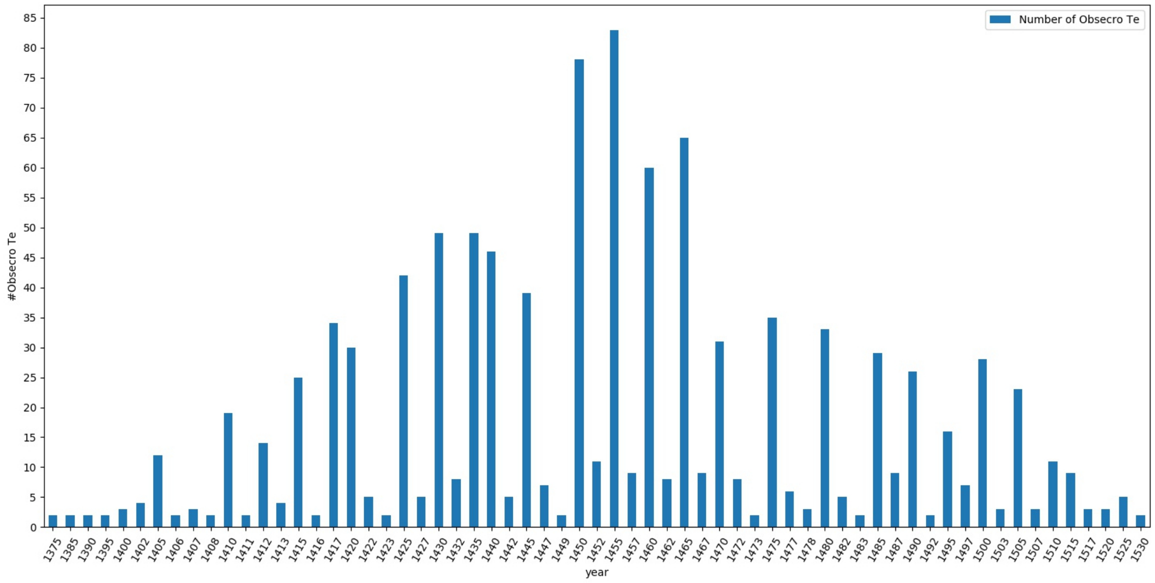

Figure 3: Number of Obsecro Te prayers per year

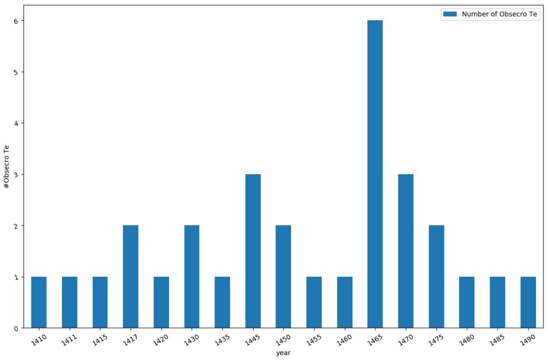

Figure 4: Obsecro Te per Year: 3gram variants over temporal axis

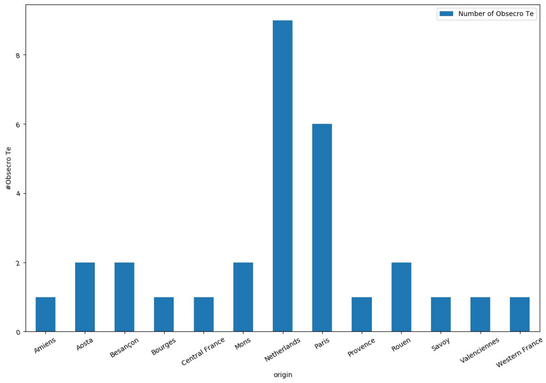

Figure 5: Obsecro Te Origin: 3gram variants over geographical axis

corpus illustrates how the production of Books of Hours increased until the second half of the fifteenth century and decreased afterwards. A larger corpus suggests that production reached it highest level during the last third of the fifteenth century (Stutzmann, 2019). An empirical overview through approximately a century and half of Book of Hours production, does not allow to draw a direct relation in the diachronic change of *Obsecro Te* prayer copies. Nonetheless, our method allows to target a variant category and to observe its years of use and origins. Table 4 provides some examples of rare 3gram variants of which it is not always obvious to assign a linguistic category. We can notice that rare 3gram variants are often expansion of frequent bigrams *omni auxilio consilio / omni consilio*. Figures 4 and 5 illustrate the use of rare orthographic 3gram variants that have less than ten character substitutions (respectively per year and per origin). This is performed by fixing the edit distance score to 10 (which means a maximum of ten substitutions) and choosing only the variants that appear only once in the corpus. We obtain for instance, the variant pair obtained by lexical substitution *petitionibus et requestis* / orationibus et requestis, where the former appears only once and the latter appears 983 times. From a geographic perspective, Netherlands is the country that produces this synformic category followed by Paris. Both places are the ones producing the

corpus in (Plummer and Clark, 2015) has a strong focus on the mid-fifteenth century, with a maximum of 80 copies ascribed to the year 1455. Figure 3 illustrates the number of witnesses by year using the arithmetic mean between the extreme dates, which explains the peaks on round numbers, particularly those ending in 0 and 5, and the important variation from one year to another largest number of copies, so that it comes as no surprise that scribes generates more variants, including the rare ones that we have isolated here. From a temporal perspective, however, we see that 3gram variants mostly appeared in 1465 and 1470. This is unexpected, since the maximum number of manuscripts in the corpus is for the years 1450 and 1455. This increase is perhaps correlated with the higher production levels of Books of Hours in the last third of the century (not strictly represented in Clark's corpus) whose variety would be reflected in Clark's corpus, but this would not explain why the 1460s and 1470s are more variant than the end of the century. We may now formulate an original hypothesis, that we observe here a loosening of the copying discipline for the *Obsecro Te* as a very common text, perhaps due to the multiplication of workshops or to other causes such as text memorisation, resulting in the emergence of many new, isolated variants. Even though our analysis cannot draw factual conclusions for now, it can nonetheless guide experts to analyse such phenomena.

7 Conclusion

We conducted for the first time a large-scale study of medieval devotional texts for the purpose of variant analysis. We used linguistic operations rather than edition operations to characterise variants in order to facilitate the interpretation of variants. We also design a suitable methodology for their detection that we hope will help medievalists in their research. If the automatic variant extraction is encouraging, further investigations are certainly needed to distinguish between orthographic in one hand, and inflectional and derivative variants in the other hand. Some computational methods well designed to deal with a particular variant detection fail when they face problematic cases: word embedding approach does not succeed to detect lexical substitutions showing a difference of distributions between the two elements, typically those substitutions that imply connectors. None of the methods is adapted to discover expansion and reduction at the n-gram level. This work constitutes a first step in the automatic study the content of Book of Hours in order to discover temporal and geographical correlations between Books of Hours, whether issued from different regions of the same country or from different countries of medieval Europe.

Acknowledgments

This work is part of the HORAE project (Hours - Recognition, Analysis, Editions) and is supported by the French National Research Agency under grant ANR-17-CE38-0008. We would like to thank professor Gregory Clark for making available *Obsecro Te* dataset and the annotations of variants.

References

Sanjeev Arora, Liang Yingyu, and Ma Tengyu. 2017. A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

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

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

Stefan Evert. 2005. *The statistics of word cooccurrences : word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

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

Sandra Hindman and James H. Marrow. 2013. *Books of hours reconsidered.* Harvey Miller Publishers, London.

Paul Jaccard. 1901. tude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Socit Vaudoise des Sciences Naturelles*, 37:547–579.

Ana Kocic. 2008. The problem of synforms. *Facta Universitatis*, 6(1):51–59.

Batia Laufer. 1988. The concept of synforms (similar lexical forms) in vocabulary acquisition. *Language and Education*, 2(2):113–132.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling,

Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

John Plummer and Gregory T. Clark. 2015. Obsecro te. *Beyond Use: A Digital Database of Variant Readings In Late Medieval Books of Hours*. http://www6.sewanee.edu/BeyondUse/texts_list.phptexts=ObsecroTe.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

Dominique. Stutzmann. 2019. Résistance au changement ? les écritures des livres d'heures dans l'espace français (1200-1600). In *'Change' in Medieval and Renaissance Scripts and Manuscripts. Proceedings of the 19th Colloquium of the Comit international de palographie latine (Berlin, 16-18 September, 2015).*, pages 101–120, Turnhout. Brepols.

Roger. Wieck. 1988. *Time sanctified : the Book of Hours in medieval art and life.* G. Braziller, New York.