Towards Automatic Variant Analysis of Ancient Devotional Texts

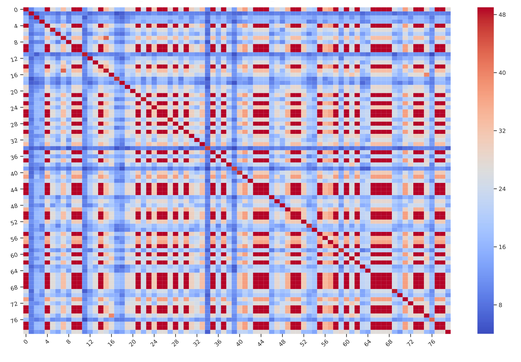
Amir Hazem Béatrice Daille Dominique Stutzmann
Jacob Currie Christine Jacquin
(1) LS2N, 2 Chemin de la Houssinière, 44322 Nantes
(2) IRHT, 40 Avenue d'Iéna, 75116 Paris
amir.hazem@ls2n.fr

Context

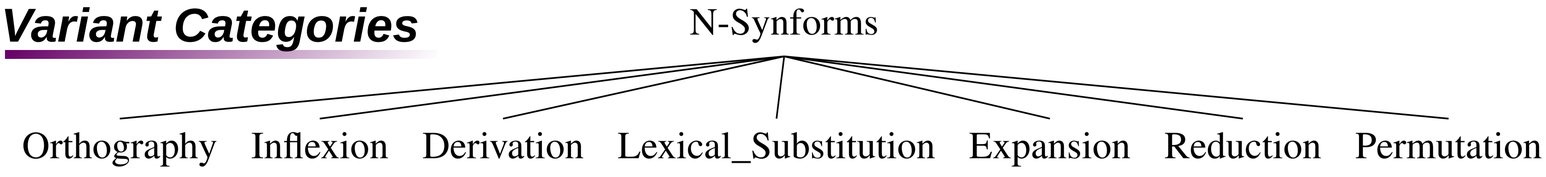
- text reuse in liturgical manuscripts
- variant readings of the *Obsecro Te* prayer categorizing pairs of expressions that exhibit variant relations
- linguistic classification
- study of *Obsecro Te* texts from a temporal and geographical axis

Folios of Books of Hours

Word overlap of 80 copies of *Obsecro Te*



Variant Categories



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

Inflexion → latin inflexions like ***crucem / cruce***

Derivation → creates a new lexical unit: affixation or convertion ***dilecto (Adj) / dilectissimo (Adj superlative)***

Lexical substitution → of a lexical unit by another. Variants in semantic relation: synonymy ***(tribuas / concedas)***, near-synonymy ***(gratie / indulgencie)*** and other semantic relation such as ***(tribuas / obtineas)***

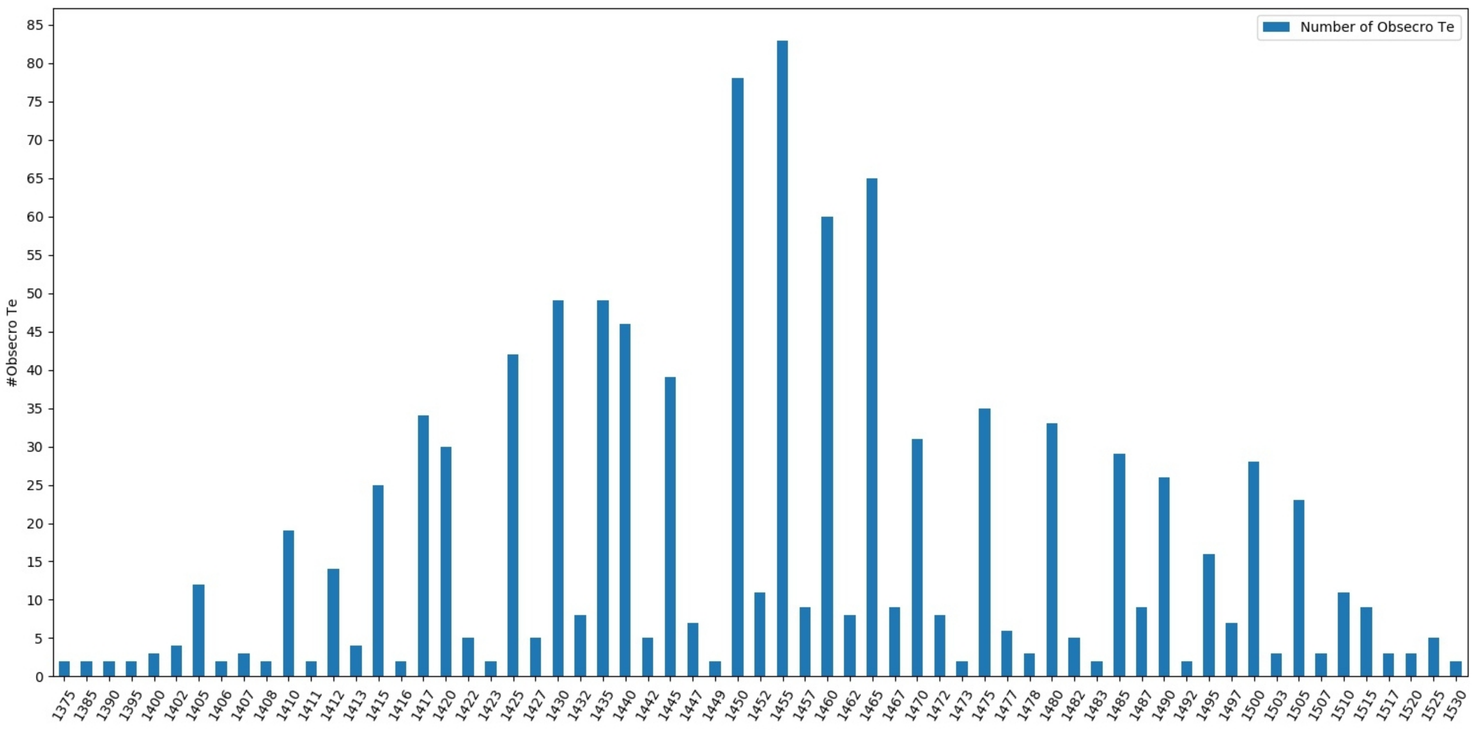
Expansion / Reduction → modification which specifies the nominal phrase, coordination that emphasize an aspect ***(criminalibus peccatis / criminalibus peccatis vel mortalibus)***

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

Alignment of two copies of *Obsecro Te* (9 first arbitrary lines)

N	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**

Number of *Obsecro Te* prayers per year



Resources

- **Beyond Use database**
- 772 manuscripts
- 21,329 variant inputs
- 3,298 distinct inputs
- 49 arbitrary lines

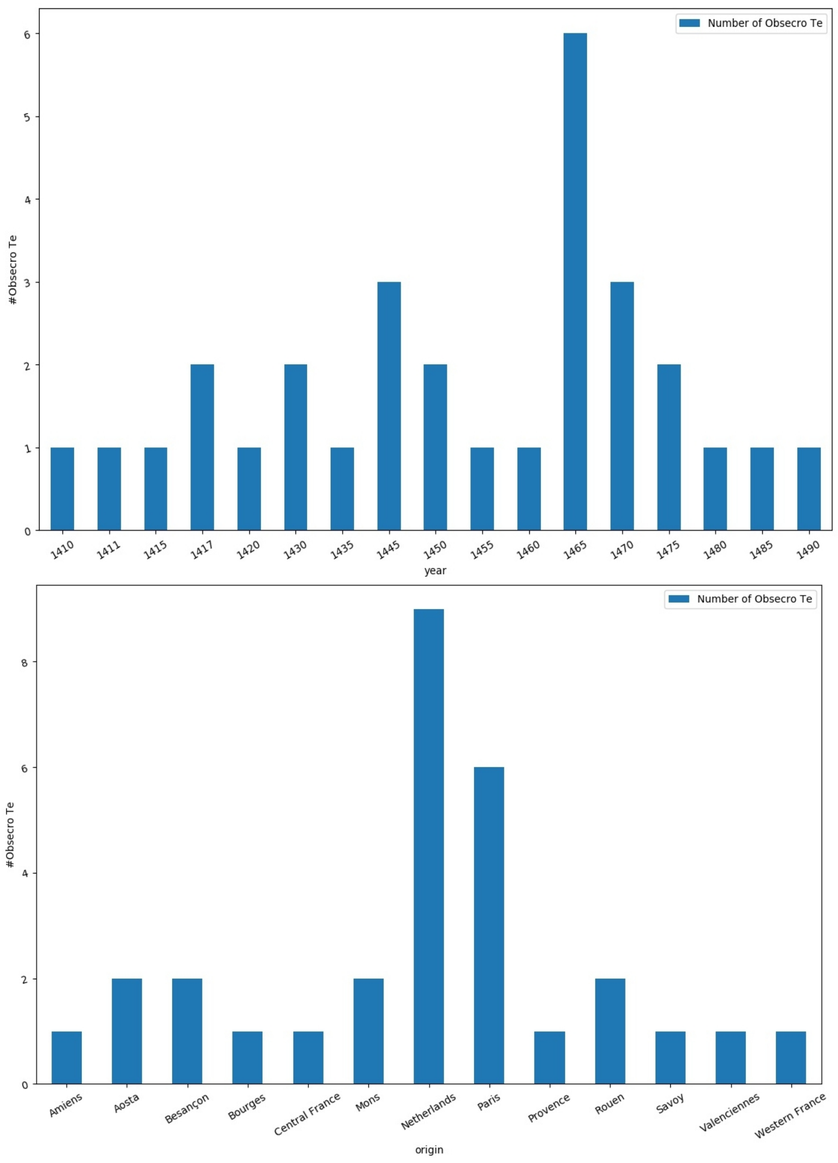
Methods

- Edit distance
- Jaccard Index
- adaptation of bag of words to ngrams
- word embeddings

Examples of 3 gram variants

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

3 gram variants over temporal and geographical axis



Results of automatic variant extraction

	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

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.

→Edit distance for unigrams
→Jaccard Index for permutation
→BoW (RV) for ALL
→Word Embeddings (Map) for ngrams > 1

Acknowlegments

This work is part of the **HORAE** project (**HO**urs - **R**ecognition, **A**nalysis, **E**ditions) 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.