# NON-LITERAL TEXT REUSE IN HISTORICAL TEXTS: AN APPROACH TO IDENTIFY REUSE TRANSFORMATIONS AND ITS APPLICATION TO BIBLICAL REUSE

## Göttingen–Hildesheim Workshop Oct. 2016

---

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, Marco Büchler

EMNLP 2016

October 13-14 2016, *University of Hildesheim*

# PHD STUDY

- PhD student at the Georg-August University School of Science (GAUSS), Program in Computer Science (PCS)
- Topic: Historical Text Reuse Style (the way text is reused) Investigation
- Supervisors:
  - Prof. Dr. Ramin Yahyapour
  - Prof. Dr. Dieter Hogrefe
  - Dr. Marco Büchler

# TABLE OF CONTENTS

# INTRODUCTION

# TEXT REUSE

Text Reuse:

- spoken and written repetition of text across time and space.

For example:

- citations, allusions, translations.

Detection methods are needed in different scholarly fields.

- They help to ensure clean libraries or identify fragmentary authors.

Text is often modified during the reuse process.

Paraphrasing and non-literal reuse challenges many approaches:

- Alzahrani et al. (2012)
  - study n-gram-, syntax-, and semantic-based detection approaches;
  - they find: as soon as reuse is slightly modified (words changed) most approaches fail.
- Barrón-Cedeño et al. (2013)
  - experiment with paraphrasing to improve plagiarism detection;
  - they found that complex paraphrasing with a high density challenges plagiarism detection, and
  - that lexical substitution is the most frequent plagiarism technique.

# TEXT REUSE DETECTION

Not only Alzahrani et al.'s but most NLP research focuses on English, although many inflecting languages exist. Ancient languages are particularly problematic as they come with:

- variants due to transmission.
- incomplete testimonials.
- diverse reuse.

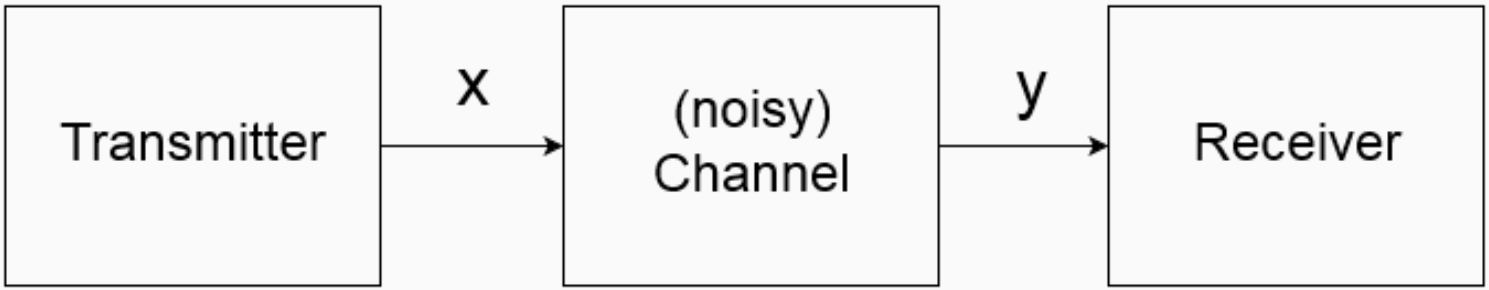
One solution: Reuse Style Investigation.

- i.e., how is reuse generally transferred and how literal is it?

# APPROACH

Inspired by Shannon's noisy-channel (Shannon, 1949) & the Kolmogorov Complexity (Li and Vitáni, 2008), we study Ancient Greek and Latin text reuse to understand how text is transferred.

- We identify operations that characterize word changes.
- We show how linguistic resources can help detect non-literal reuse.
- We complement the automated approach with a manual analysis.



# METHODOLOGY

# RESEARCH QUESTIONS

- RQ1. What is the extent of non-literal reuse in our datasets?
- RQ2. How is the non-literally reused text modified in our datasets?
  - RQ2.1. How can linguistic resources support the discovery of non-literal reuse?
  - RQ2.2. What are the limitations of an automated classification approach relying on linguistic resources?

# STUDY DESIGN

We:

1. define operations reflecting literal reuse and semantic replacements.
2. develop an algorithm that looks for identical, similar words, morphological changes and semantic changes.
3. apply it to our two data-sets (next slide).
4. manually analyze a smaller sample of our reuse using further operations.

# DATA-SETS - ANCIENT GREEK AND LATIN DATA-SET

"Salvation for the Rich"
Clement of Alexandria
Christian theologian, 2nd cent.

- Known for his retelling of biblical excerpts
- Reuse annotated by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 199 verse-reuse-pairs
- Pointing to 15 Bible books

Extracts from 12 works & 2 collections
Bernard of Clairvaux
French abbot, 12th cent.

- Known for his influence on the Cistercian order and his work in biblical studies
- Reuse extracted by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 162 verse-reuse-pairs
- Pointing to 31 Bible books

# BIBLICAL REUSE EXAMPLES

| more literal | Bible verse | Bernard reuse |
|---|---|---|
| **Proverbs 18 3** | **impius cum in profundum venerit** peccatorum **contemnit** sed sequitur eum ignominia et obprobrium (*When the wicked man is come into the depth of sins, also contempt comes but ignominy and reproach follow him)* | **Impius** , **cum venerit in profundum** malorum , **contemnit** (*When the wicked man is come into the depth of evil*) |

| less literal | Bible verse | Clement reuse |
|---|---|---|
| **1Cor 13 13** | νυνὶ **δὲ μένει πίστις** , **ἐλπίς** , **ἀγάπη** , **τὰ τρία ταῦτα μείζων δὲ τούτων ἡ ἀγάπη** (*And now remain faith, hope, love, these three; but the greatest of those is love.*) | **πίστει** καὶ **ἐλπίδι** καὶ **ἀγάπῃ** (*faith, and hope, and love* - in dative case) |
| | | **ἀγάπην** , **πίστιν** , **ἐλπίδα** (*love, faith, hope* - in accusative case) |
| | | **μένει δὲ τὰ τρία ταῦτα** , **πίστις** , **ἐλπίς** , **ἀγάπη** · **μείζων δὲ** ἐν **τούτοις ἡ ἀγάπη** *(and remain these three, faith, hope, love; but the greatest among them is love)* |

| non-literal | Bible verse | Clement reuse |
|---|---|---|
| **Mt 12 35** | ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἀγαθοῦ θησαυροῦ ἐκβάλλει ἀγαθά , καὶ ὁ **πονηρὸς** ἄνθρωπος ἐκ τοῦ πονηροῦ θησαυροῦ ἐκβάλλει **πονηρά** . (*A good man out of good storage brings out good things , and an evil man out of the evil storage brings evil things .*) | Ψυχῆς , τὰ δὲ ἐκτός , κἂν μὲν ἡ ψυχὴ χρῆται καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ **πονηρῶς** , **πονηρά** , ὁ κελεύων ἀπαλλοτριοῦν τὰ ὑπάρχοντα (*[are whitin the] soul, and some are out, and if the soul uses them good, those things are also thought of as good, but if [they are used as] bad, [they are thought of as] bad; he who commands the renouncement of possessions*) |

# LINGUISTIC SUPPORT – LEMMA RESOURCES

We aggregate:

- Biblindex' Lemma Lists
  - 65,537 Biblical Greek entries
  - 315,021 Latin entries
- Classical Language Tool Kit (CLTK) (Johnson et al., 2014)
  - 953,907 Ancient Greek words
  - 270,228 Latin words
- Greek New Testament of the Society of Biblical Literature[1] & Septuaginta (Rahlfs, 1935a; UPenn) 59,510 word-lemma-pairs

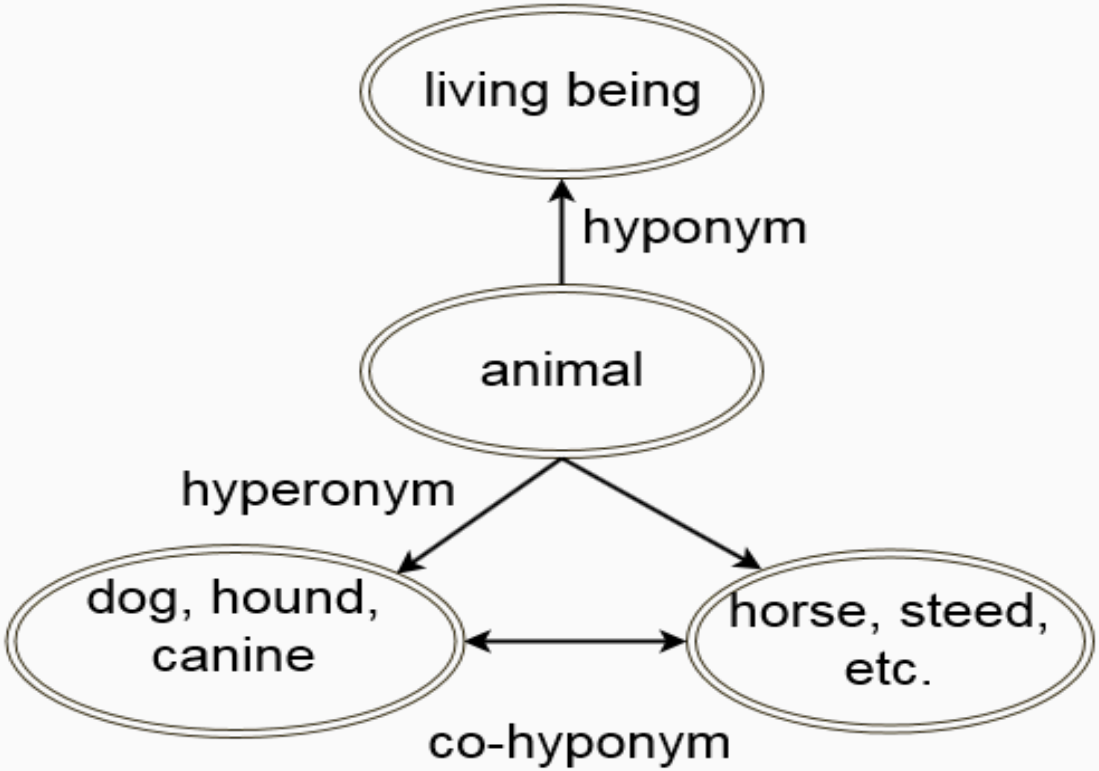
[1] Logos Bible Software `http://sblgnt.com/about/`

# LINGUISTIC SUPPORT – ANCIENT GREEK WORDNET (AGWN)

99K synsets

of which 33K contain Ancient Greek and 27K Latin words

(Bizzoni et al., 2014; Minozzi, 2009)



# TRANFORMATION OPERATIONS

**Table 1:** Operation list for the automated approach

| operation | description | example |
|---|---|---|
| *NOP(reuse_word, orig_word)* | Original and reuse word are equal. | *NOP(maledictus,maledictus)* |
| *upper(reuse_word, orig_word)* | Word is lowercase in reuse and uppercase in original. | *upper(kai,Kai)* - in Greek |
| *lower(reuse_word, orig_word)* | Word is uppercase in reuse and lowercase in original. | *lower(Gloriam,gloriam)* |
| *lem(reuse_word, orig_word)* | Lemmatization leads to equality of reuse and original. | *lem(penetrat,penetrabit)* |
| *repl_syn(reuse_word, orig_word)* | Reuse word replaced with a synonym to match original word. | *repl_syn(magnificavit,glorificavit)* |
| *repl_hyper(reuse_word, orig_word)* | Word in Bible verse is a hyperonym of the reused word. | *hyper(cupit,habens)* |
| *repl_hypo(reuse_word, orig_word)* | Word in Bible verse is a hyponym of the reused word. | *hypo(dederit,tollet)* |
| *repl_co-hypo(reuse_word, orig_word)* | Reused word and original have the same hyperonym. | *repl_co-hypo(magnificavit,fecit)* |
| *NOPmorph(reuse_tags, orig_tags)* | Case or PoS did not change between reused and original word. | *NOPmorph(na,na)* |
| *repl_pos(reuse_tag, orig_tag)* | Reuse and original contain the same cognate, but PoS changed. | *repl_pos(n,a)* |
| *repl_case(reuse_tag, orig_tag)* | Reuse and original have the same cognate, but the case changed. | *repl_case(g,d)* - cases genitive, dative |
| *lemma_missing(reuse_word, orig_word)* | Lemma unknown for reuse or original word. | *lemma_missing(tentari, inlectus)* |
| *no_rel_found(reuse_wword, orig_word)* | Relation for reuse or original word not found in AGWN. | *no_rel_found(gloria,arguitur)* |

# QUALITATIVE COMPLEMENT

We manually analyze:

- 60 Ancient Greek & 100 Latin instances
- 192 &. 224 replacements
- Using `ins(word)`, `del(word)` and replacements:
  - `NOP`, `lem`, `repl_syn`, `repl_hyper`, `repl_hypo`, `repl_co-hypo`
- We assign morphological categories from Perseus' tag-set (Bamman and Crane 2011)
  - E.g., `repl_case_a_g` `repl_num_s_p`

**Table 2:** Excerpt from Perseus' tag-set

| Category | Value | Tag |
|---|---|---|
| person | first person | 1 |
| | second person | 2 |
| | third person | 3 |
| number | singular | s |
| | plural | p |
| | dual | d |
| tense | present | p |
| | imperfect | i |
| | perfect | r |
| | pluperfect | l |
| | future perfect | t |
| | future | f |
| | aorist | a |

# RESULTS

# LITERAL SHARE OF THE REUSE (RQ1)

What is the extent of non-literal reuse in our datasets?

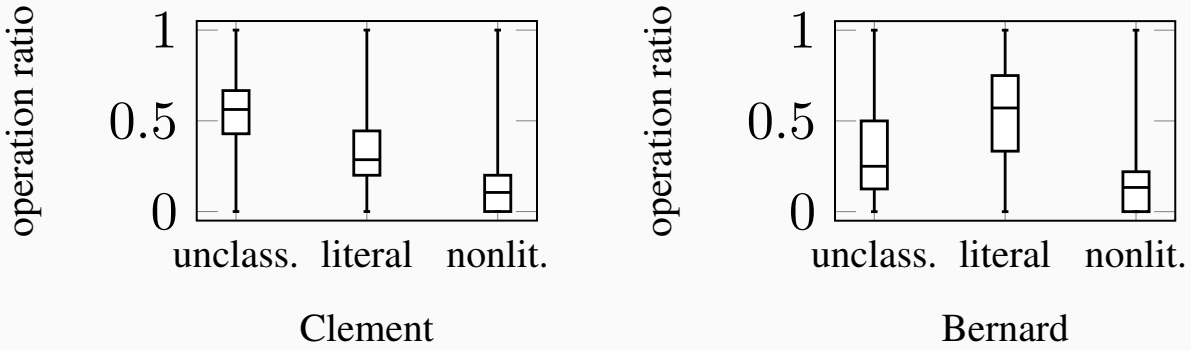



**Figure 1:** Ratios of operations in reuse instances. literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

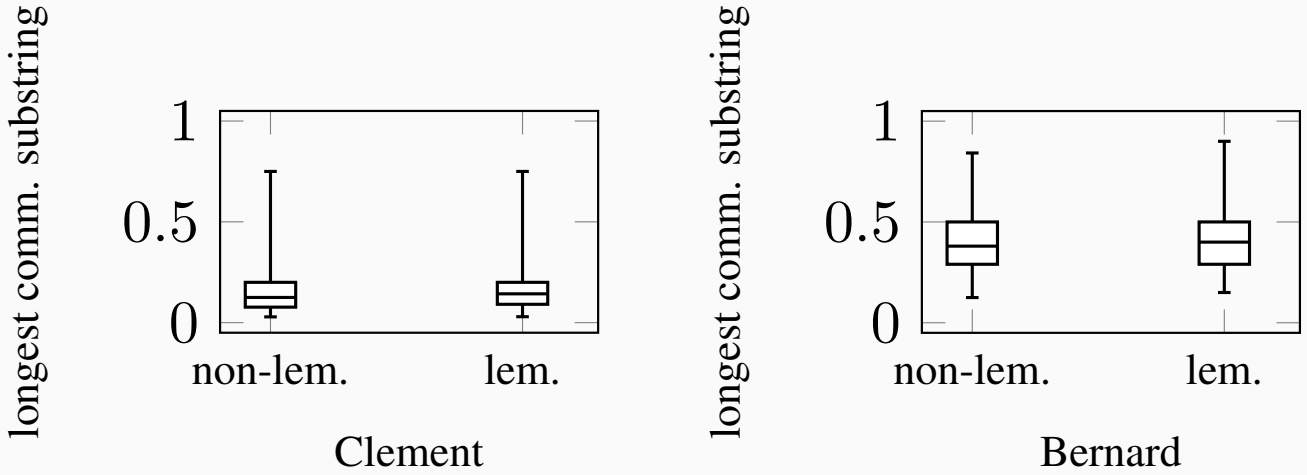



**Figure 2:** Ratios of literal overlap between reuse instances and originals.

# AUTOMATED APPROACH (RQ2.1)

How is the non-literally reused text modified in our datasets? (RQ2)
How can linguistic resources support the discovery of non-literal reuse? (RQ2.1)

**Table 3:** Absolute numbers of operations identified automatically.

| | literal | | | | non-literal | | | | unclassified | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | NOP | upper | lower | lem | syn | hyper | hypo | co-hypo | no_rel_found | lem_missing | total |
| Greek | 337 | 6 | 0 | 356 | 153 | 20 | 14 | 101 | 563 | 639 | 2189 |
| Latin | 587 | 0 | 44 | 102 | 60 | 14 | 28 | 68 | 347 | 85 | 1335 |

# AUTOMATED APPROACH (RQ2.1) - COVERAGE VALUES

Operations that successfully looked up a lemma: `lem_success`={`lem, syn, repl_hyper, repl_hypo, repl_co-hypo, no_rel_found`}, with `lem_missing` representing not found tokens in the lemmata.

$$\text{COV}_{\text{lem}} = \frac{\sum_{\text{Occ}(o)} o \in \text{lem\_success}}{\sum_{\text{Occ}(o)} o \in \text{lem\_success} \cup \{\text{lem\_missing}\}}$$

$$\text{COV}_{\text{AGWN}} = \frac{\sum_{\text{Occ}(o)} o \in \text{agwn\_success}}{\sum_{\text{Occ}(o)} o \in \text{agwn\_success} \cup \{\text{no\_rel\_found}\}}$$

We obtain a $\text{cov}_{\text{lem}}$ of **0.65** for our Greek and **0.88** for the Latin data-set. And a $\text{cov}_{\text{AGWN}}$ of **0.34** for our Greek and **0.33** for our Latin data-set.

Language resources help to get an idea of reuse components.

# QUALITATIVE APPROACH (RQ2.2)

How is the non-literally reused text modified in our data-sets? (RQ2)
What are the limitations of an automated classification approach relying on linguistic resources? (RQ2.2)

**Table 4:** Exceptions that prevent applying the operations.

| Exception | Quantity | |
|---|---|---|
| | Clement | Bernard |
| Word changed to antonym | 1[2] | 0 |
| Synonym and morphology changed | 1 | 16 |
| More than one morphological category changed | 1 | 7 |
| Synonym is multi-word expression | 3 | 5 |
| Many-to-many | 0 | 1[3] |

[2] "**the** God, the good (**one**)" (Clement) vs. "**none** is good but the God" (Bible).
[3] "judged calmly" (Bernard) vs. "fake friend" (Sal 12 18).

# DISCUSSION

- RQ1. The reuse is significantly non-literal
  - Techniques beyond stemming & even semantic closeness are required; suggestion: conceptualizing
- RQ2.1. Our results show
  - the possibility of supporting reuse detection with linguistic resources
- RQ2.2. Qualitative complement
  - especially the exceptions show that reuse detection needs looser relation (multi-to-multi-word) associations or implicit expert knowledge.

# CONCLUSION AND FUTURE WORK

# SUMMARY

We contributed:

- an automated approach to characterize how text is transformed between reuse and original,
- an application of the approach to two text data-sets where reuse was manually identified,
- empirical data based on the automated approach, complemented by a manual identification.

# FUTURE WORK

A more comprehensive study could strengthen the findings:

- using larger reuse data-sets,
- additional languages (inflecting vs. non-inflecting),

An automated approach for deriving an original text excerpt would be learning edit scripts (Kehrer, 2014, Chawathe et al., 1996).

# Thank you!

# REFERENCES


- Kurt Aland and Barbara Aland, editors. 1966. The Greek New Testament. Deutsche Bibelgesellschaft-United Bible Societies, 27 edition.
- Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. Trans. Sys. Man Cyber Part C, 42(2):133149.
- David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In Caroline Sporleder, Antal van den Bosch, & Kalliopi Zervanou (Eds) Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series, pages 7998, Berlin, Germany. Springer-Verlag.
- Barrón-Cedeño and Marta Vila and M.Antònia Martí and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistic, 39(4):917947.
- Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of ancient greek wordnet. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).
- Kyle P. Johnson, Patrick J. Burns, Luke Hollis, Martín Pozzi, Amit Shilo, Stephen Margheim, Gitter Badger, and Eamonn Bell. 20142016. Cltk: The classical language toolkit. `https://github.com/cltk/cltk.DOI10.5281/zenodo`. 44555 v0.1.32.
- Laurence Mellerin. 2014. New ways of searching with biblindex, the online index of biblical quotations in early christian literature. In Claire Clivaz, Andrew Gregory, and David Hamidovic, editors, Digital Humanities in Biblical, Early Jewish and Early Christian Studies, chapter 11, pages 175192. Brill, Leiden.
- Laurence Mellerin. 2016. Biblindex. `http://www.biblindex.mom.fr/`.
- Ming Li and Paul Vitáni. 2008. An Introduction to Kolmogorov Complexity and Its Applications 3. Auflage. Springer New York.
- Stefano Minozzi, 2009. Innsbrucker Beiträge zur Sprachwissenschaft, volume 137, chapter The Latin WordNet Project, pages 707716. Institut für Sprachen und Literaturen der Universität Innsbruck, Innsbruck.
- Alfred Rahlfs, editor. 1935a. Septuaginta. Württembergische Bibelanstalt, 9 edition. 1971.
- Alfred Rahlfs, editor. 1935b. Septuaginta, id est Vetus Testamentum Graece juxta LXX interpretes. Rahlfs. 2 vol., 1950.
- C. E. Shannon, A Mathematical Theory of Communication Urbana, IL: University of Illinois Press, 1949 (reprinted 1998).
- Gribomont J. Weber R., Fischer B., editor. 1969, 1994, 2007. Biblia sacra juxta vulgatam versionem. Deutsche Bibelgesellschaft.

# APPENDIX

# BIBLINDEX BIBLE EDITIONS AND CLEMENT EDITION

Old Testament:

- Alfred Rahlfs, editor. 1935. Septuaginta, id est Vetus Testamentum Graece juxta LXX interpretes. Rahlfs. 2 vol., 1950.

New Testament:

- Kurt Aland and Barbara Aland, editors. 1966. The Greek New Testament. Deutsche Bibelgesellschaft-United Bible Societies, 27 edition.

Latin Bible:

- Gribomont J. Weber R., Fischer B., editor. 1969, 1994, 2007. Biblia sacra juxta vulgatam versionem. Deutsche Bibelgesellschaft.

Clement Edition:

- Clément d'Alexandrie, Quel riche sera sauvé ?, Quis dives salvetur, P. A. O à Sources Chrétiennes, col. 537, p. 100 ff., 2011.

# BERNARD VOLUMES

- Bernard de Clairvaux, Sermons sur le Cantique 1-15, Sermones super Cantica Canticorum 1-15, P. A. O à Sources Chrétiennes, col. 414, 1996.
- Bernard de Clairvaux, Sermons sur le Cantique 16-32, Sermones super Cantica Canticorum 16-32 , P. A. O à Sources Chrétiennes, col. 431, 1998.
- Bernard de Clairvaux, Sermons sur le Cantique 33-50, Sermones super Cantica Canticorum 33-50, , P. A. O à Sources Chrétiennes, col. 452, 2000.
- Bernard de Clairvaux, Sermons sur le Cantique 51-68, Sermones super Cantica Canticorum 151-68 , P. A. O à Sources Chrétiennes, col. 472, 2003.
- Bernard de Clairvaux, Sermons sur le Cantique 69-86, Sermones super Cantica Canticorum 69-86 , P. A. O à Sources Chrétiennes, col. 511, 2007.
- Bernard de Clairvaux, Amour de Dieu, Liber de diligendo Deo, P. A. O à Sources Chrétiennes, col. 393, 1993.
- Bernard de Clairvaux, Lettres 1-41, Epistolae 1-41, P. A. O à Sources Chrétiennes, col. 425, 1997.
- Bernard de Clairvaux, Lettres 42-91, Epistolae 42-91, P. A. O à Sources Chrétiennes, col. 425, 1997.
- Bernard de Clairvaux, Lettres 92-163, Epistolae 92-163, P. A. O à Sources Chrétiennes, col. 425, 1997.
- Bernard de Clairvaux, Lettres 363-495, Epistolae 363-495, SBO VIII, 311-447 1957-1977.
- Bernard de Clairvaux, Éloge de la nouvelle chevalerie, Liber ad milites Templi De laude novae militiae, P. A. O à Sources Chrétiennes,col. 367, 1990.
- Bernard de Clairvaux, Précepte et la Dispense, Liber de praecepto et dispensatione, P. A. O à Sources Chrétiennes, col. 457, 2000.
- Bernard de Clairvaux, Grâce et le Libre Arbitre, Liber de gratia et de libero arbitrio, P. A. O à Sources Chrétiennes, col. 393, 1993.
- Bernard de Clairvaux, Degrés de l'humilité et de l'orgueil, Liber de gradibus humilitatis et superbia, SBO III, 13-59, 1957-1977.
- Bernard de Clairvaux, Vie de saint Malachie, Vita sancti Malachiae episcopi, P. A. O à Sources Chrétiennes, col. 367, 1990.
- Bernard de Clairvaux, Sermon lors de la mort de Malachie, Sermo in transitu sancti Malachiae, P. A. O à Sources Chrétiennes, col. 367, 1990
- Bernard de Clairvaux,Sentences, Sententiae, SBO VI-2, 1-256, 1957-1977.
- Bernard de Clairvaux, Conversion (Aux clercs sur la conversion), Sermo de conuersione ad clericos, P. A. O à Sources Chrétiennes, col. 457, 2000.
- Bernard de Clairvaux, Louange de la Vierge Mère, Homiliae super Missus est (In laudibus Virginis Matris), P. A. O à Sources Chrétiennes, col. 390, 1993.
- Bernard de Clairvaux, Sermon pour la Nativité de la Bienheureuse Vierge Marie, Sermo in natiuitate Beatae Mariae Virginis, SBO V, 275-288, 1957-1977.
- Bernard de Clairvaux, Sermons pour la naissance de saint Victor, Sermones varii : In natali sancti Victoris, col. 526, 2010.

# LICENCE