

Arabic Corpus – Corpus Description

July 2026 – Bastien KINDT

To date, the Arabic Corpus amounts to 12.940 word-forms, enriched by lemmatization and POS-tagging. This document lists alphabetically the authors and works currently included in the corpus. For information and terms of collaboration: info-gregori@uclouvain.be.

ANTIOCHIAN MENOLOGIUM

Date: 11th cent. AD

This work is supported by the CEDRAC (USJ - Beirut), the GREgORI project (UCLouvain), and the Emmy Noether Research Group «Religious Conflict and Mobility. Byzantium and the Greater Mediterranean, 700–900» (Eberhard Karls Universität Tübingen) to develop a specialized AI model dedicated to the linguistic analysis of Christian Arabic. Ground truth manually annotated (lemmatization and part of speech tagging) will be gradually made available to researchers.

Work: [Introduction](#)

Work-tag: [Ant. Men. – Introduction](#)

Word-forms count: 4.441

Topic: menologion

Transmission: manuscripts

Text translated from another language: no

Editions: YÜHANNĀ ‘ABD AL-MASĪḤ († 11e s.), *Ma‘īn al-ḥayāt, al-Markab al-sā‘ir fī mīnā’ al-naḡāt, autrement connu comme Al-Dūlāb*, ed. H. IBRAHIM (*Patrimoine Arabe Chrétien*, 31. *Textes et Études de Littérature Arabe Chrétienne Ancienne*), Beyrou – Jounieh, 2020, p. 61-75.

Lemmatization and POS-tagging: CEDRAC (USJ - Beirut) and GREgORI project (UCLouvain)

Work: [Passio Sancti Demetrii \(BHG 498\)](#)

Work-tag: [Ant. Men. – 30. Pass. S. Dem.](#)

Word-forms count: 2.941

Topic: menologion

Transmission: manuscripts

Text translated from another language: yes

Editions: YÜHANNĀ ‘ABD AL-MASĪḤ († 11e s.), *Ma‘īn al-ḥayāt, al-Markab al-sā‘ir fī mīnā’ al-naḡāt, autrement connu comme Al-Dūlāb*, II, ed. H. IBRAHIM (*Patrimoine Arabe Chrétien*, 31. *Textes et Études de Littérature Arabe Chrétienne Ancienne*), Beyrou – Jounieh, 2021, p. 295-303.

Lemmatization and POS-tagging: CEDRAC (USJ - Beirut) and GREgORI project (UCLouvain)

GREGORY OF NAZIANZUS

Date: 4th cent. AD

Work: [Homily 1](#)

Work-tag: [Greg. of Naz., Or. 1](#)

Word-forms count: 824

Topic: homiletic

Transmission: manuscripts

Text translated from another language: no

Editions: Gregorius Nazianzenus. *Opera. Versio Arabica antiqua*, II. *Orationes* I, XLV, XLIV (Arab. 9, 10, 11), ed. L. TUERLINCKX (*Corpus Christianorum. Series Graeca*, 43. *Corpus Nazianzenum*, 10).

Lemmatization and POS-tagging: Jacques Grand'Henry and Laurence Tuerlinckx

Work: [Homily 11](#)

Work-tag: Greg. of Naz., Or. 11

Word-forms count: 1.594

Topic: homiletic

Transmission: manuscripts

Text translated from another language: no

Editions: Gregorius Nazianzenus. *Opera. Versio Arabica antiqua*, IV: *Orationes* XI, XLI (arab. 8, 12), ed. J. GRAND'HENRY (Corpus Christianorum. Series Graeca, 85. Corpus Nazianzenum, 97).

Lemmatization and POS-tagging: Jacques Grand'Henry and Laurence Tuerlinckx

Work: [Homily 27](#)

Work-tag: Greg. of Naz., Or. 27

Word-forms count: 2.071

Topic: homiletic

Transmission: manuscripts

Text translated from another language: no

Editions: L. TUERLINCKX, *Histoire de la version arabe de Grégoire de Nazianze : édition critique des Orationes 27 et 38, étude de la tradition manuscrite et identification des divers états de texte*, Thèse de doctorat, Université catholique de Louvain, Louvain-la-Neuve, 2007.

Lemmatization and POS-tagging: Jacques Grand'Henry and Laurence Tuerlinckx

SULAYMĀN IBN ḤASSAN AL-ĠAZZĪ

Date: 10-11th cent. AD

This work is supported by the CEDRAC (USJ - Beirut), the GREgORI project (UCLouvain), and the Emmy Noether Research Group «Religious Conflict and Mobility. Byzantium and the Greater Mediterranean, 700–900» (Eberhard Karls Universität Tübingen) to develop a specialized AI model dedicated to the linguistic analysis of Christian Arabic. Ground truth manually annotated (lemmatization and part of speech tagging) will be gradually made available to researchers.

Work: [Theological Treatises](#)

Work-tag: Sul. Al-Gazzi, Treatises

Word-forms count: 8.451

Topic: theology

Transmission: manuscripts

Text translated from another language: no

Editions: *I trattati teologici di Sulaymān Ibn Ḥassan Al-Ġazzī*, ed. P. LA SPISA (Corpus Scriptorum Christianorum Orientalium, 648. Scriptores Syri, 52), Leuven, 2013, p 1-84.

Lemmatization and POS-tagging: CEDRAC (Chadi Kahwaji - USJ - Beirut) and GREgORI (Bastien Kindt)

GREgORI From raw text to tagged corpus
NLP solutions for Ancient Greek and
other languages of the Christian East

