The "tongueprint" as language identification tool: Elaborating the proof-of-concept

Raf van Rooy **Flavio Massimiliano Cecchini** **Isabelle Maes**
KU Leuven
Oude Markt 13, 3000 Leuven, Belgium
raf.vanrooy@kuleuven.be, flavio.cecchini@live.it, isabelle11maes@gmail.com

Relevant UniDive working groups: **WG1**, WG3, WG4

1 Introduction

This proposal aims to introduce the ERC-StG project ERASMOS, especially its language-technological aspects. This project focuses on classical bilingualism in the Renaissance, defined as the competence to use Latin and Ancient Greek actively in both speaking and writing, in a historical context where they interacted with vernacular languages like Italian and Dutch and other learned languages like Hebrew and Arabic. ERASMOS moreover analyzes the forms and functions of Renaissance classical bilingualism in comparison to its ancient pendant. The focus is on the long 15th century: from the start of Manuel Chrysoloras' crucial teaching in Florence (1397) to Erasmus of Rotterdam's death (1536), two highly influential scholars who shaped the Renaissance uses of Greek and Latin (see, respectively, e. g. Thorn-Wickert, 2006; Van Rooy, 2023).

Judging by the state of the art, Renaissance classical bilingualism has three main manifestations.

1. Only one of them, Latin-Greek translation, has attracted extensive attention (e. g. Pade, 2020; also see the *Catalogus Translationum et Commentariorum* [CTC][1]), but even there many lacunas remain.

2. The humanists' bilingual compositions, including letter collections and poetical cycles, remain largely unstudied. For some big names, they have been edited (e. g. Erasmus) but never approached through the lens of bilingualism.

3. The humanists' Latin-Greek code-switching (CS), occurring in diverse text types and oral discourse, has been the topic of only two recent international workshops, cf. (Barton and Van Rooy, In press).

[1] http://catalogustranslationum.org/

A major problem to be tackled for a large-scale investigation into Renaissance classical bilingualism is how to organize it systematically, since until now scholarship has operated largely on a case-study basis.

2 Tongueprint: a language identification pipeline

The ERASMOS project aims to map who wrote in the two classical languages, where, and on which themes. The classical bilingual corpus is so vast that it can only be efficiently navigated using automated techniques. These will help arrive at a global assessment of the linguistic make-up of classical bilingual texts available in print, in modern editions or, if these are lacking, in early modern ones. The main goal is to script a pipeline that consists of the following actions (see the proof-of-concept in Van Rooy and Mercelis, 2022, pp. 2–5, whence Figure 1 is taken):

1. detection of alphabets/fonts used;

2. Optical Character Recognition (OCR) of the text;

3. detection of the language(s) used;

4. calculation of number, proportion and distribution of languages in terms of words;

5. calculation of the density of code-switches per 1000 words and their average word lengths;

6. analysis of the level of code-switches (within or between sentences, paragraphs, . . .);

7. visualization of the data.

The project will build on advances in language recognition software and CS, and go beyond them by using better language models.[2] We also aim to insert a quotation estimate in the pipeline in order

[2] For CS, see (Molina et al., 2016; Liu and Smith, 2020; Volk et al., 2022), and in general the *Computational Approaches to Linguistic Code-Switching* workshop that took place in 2021.

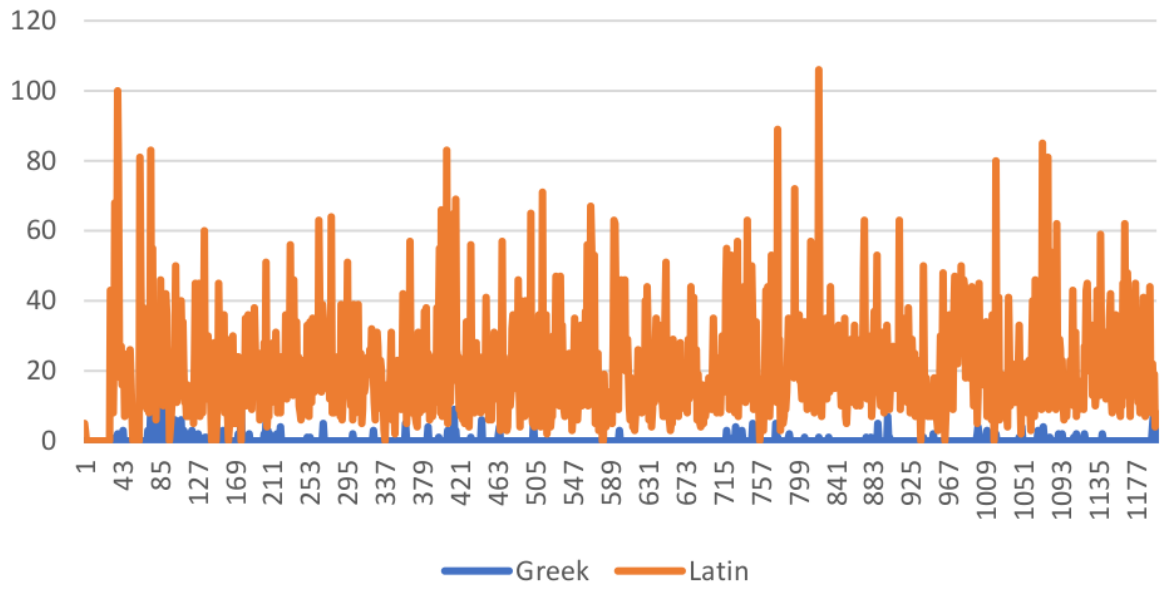

Figure 1: Distribution of Latin (orange) *vis-à-vis* Greek (blue) in the *Praise of Folly* by Erasmus (Van Rooy and Mercelis, 2022, pp. 2–5).

to answer the question: what is the share of quoted Greek versus newly composed Greek?

3 Aims of the tongueprint in ERASMOS

The tongueprint will be applied to the ancient texts in the *Code-Switching in Roman Literature* (CSRL) database[3] and especially to a representative corpus of Renaissance texts, in order to compare classical bilingualism in antiquity and the Renaissance formally on a macrolevel. Erasmus of Rotterdam, in particular, will receive close attention in comparison to classical examples like Cicero, Suetonius, and Martial. In the context of the ERASMOS project, we also aim to develop an open-access corpus of Erasmus' *œuvre*. We intend to generate tongueprint analyses for his works, which will allow for discovering tendencies across genres and time, as we expect him to evolve from no CS to extensive CS in the years 1490–1536, given that he only learned Greek around 1500, and to have Greek in literary works like *Praise of Folly* but less in manuals he wrote.

The initial aim of the tongueprint for ERASMOS is to conduct historical literary research, comparative across the ages but also among humanists from Chrysoloras to Erasmus' contemporaries. Additionally, by involving correspondence in the research, the tongueprint can help tracking historical sociolinguistic trends. However, the tongueprint does not aim to look at Latin and Greek alone but to be sensitive to early modern language varieties more broadly, including Hebrew and the vernacular languages. This will contribute to a discipline like book history, in which scholars are only on the verge of integrating multilingualism into their field, especially on the metadata level. The rudimentary book-historical practice of giving simplistic non-quantitative language tags (e. g. "Latin") to describe individual editions on platforms such as the *Universal Short Title Catalogue* (USTC)[4] may stand as a token of this lack of engagement with multilingualism. The tongueprint statistics will offer a finer-grained picture. Further, a detailed study of how such a CS is realized entails the analysis of its morphosyntactic structure: to this end, the tongueprint will include (manual or automated) linguistic annotation of such passages, and the choice naturally falls on a formalism geared towards multilingual, typologically oriented investigation, in particular the one of the Universal Dependencies project[5] (UD). We observe that UD already considers CS,[6] and also includes some code-switching treebanks.

[3] https://csrl.classics.cam.ac.uk/

[4] https://www.ustc.ac.uk/

[5] https://universaldependencies.org/

[6] https://universaldependencies.org/foreign.html

More broadly, ERASMOS hopes to prioritize multilingualism more, as this is a phenomenon that sometimes tends to be ignored in current NLP research. As we are at the very beginning of the ERASMOS project, we aim to focus in our presentation on the tongueprint as a proof-of-concept and how we can elaborate it to a full-fledged tool, taking into account which obstacles and challenges we face or may possibly face. We aim to highlight the great benefits of the tool and possible pitfalls in developing it, for which we hope to solicit feedback from peers.

References

William Barton and Raf Van Rooy. In press. Introduction. *Journal of Latin Cosmopolitanism and European Literatures (*JOLCEL*)*.

Shijia Liu and David Smith. 2020. Detecting *de minimis* Code-Switching in Historical German Books. In *Proceedings of the 28th International Conference on Computational Linguistics (*COLING*)*, pages 1808–1814, Barcelona, Spain (online). International Committee on Computational Linguistics.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the Second Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, TX, USA. Association for Computational Linguistics (ACL).

Marianne Pade. 2020. *"Conquering Greece": On the Correct Way to Translate in Fifteenth-Century Humanist Translation Theory*, number 17 in Acta Conventus Neo-Latini, chapter 3. Brill, Leiden, the Netherlands.

Lydia Thorn-Wickert. 2006. *Manuel Chrysoloras (ca. 1350–1415)*. Number 92 in Bonner romanistische Arbeiten. Peter Lang, Lausanne, Switzerland.

Raf Van Rooy. 2023. *Erasmus, an Unsuspected Superspreader of New Ancient Greek?* , number 4 in Euhormos: Greco-Roman Studies in Anchoring Innovation, chapter 10. Brill, Leiden, the Netherlands.

Raf Van Rooy and Wouter Mercelis. 2022. The art of code-switching: Toward a "tongueprint" of multilingual literary personas in Erasmus' *Praise of Folly* and Aleandro's journal? *Leuven Working Papers in Linguistics*, 9:1–16.

Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Silvan Schroffenegger, Raphael Schwitter, Phillip Ströbl, and Benjamin Suter. 2022. *Nunc profana tractemus*. Detecting Code-Switching in a Large Corpus of 16th Century Letters. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2901–2908, Marseille, France. European Language Resources Association (ELRA).