Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač (Eds.)

Lexicography and Semantics

Proceedings of the
XXI EURALEX International Congress
8–12 October 2024
Cavtat, Croatia

Nakladnik:	Institut za hrvatski jezik
Za nakladnika:	Željko Jozić
Knjiga:	Lexicography and Semantics, Proceedings of the XXI EURALEX International Congress
Urednice:	Kristina Štrkalj Despot Ana Ostroški Anić Ivana Brač
Tehničko uređenje i računalni slog:	Elena Vrbanić
Oblikovanje naslovnice:	Elena Vrbanić
Tisak:	

ISBN 978-953-7967-77-2

Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač (Eds.)

Lexicography and Semantics

Proceedings of the

XXI EURALEX International Congress

8–12 October 2024
Cavtat, Croatia

TABLE OF CONTENTS

Chapter II. Lexicography and Language Technologies

Chapter III. Dictionary Writing Systems and Lexicographic Tools....285

Chapter IV. Reports on Lexicographical and Lexicological Projects....361

EURALEX XXI

CROATIA 2024

FOREWORD

This book contains papers presented at the XXI EURALEX International Congress, Cavtat, Croatia, 8–12 October 2024.

The main topic of the XXI EURALEX International Congress was *Lexicography and Semantics*.

Lexicography and semantics, particularly lexical semantics, share a deep connection, as the accurate representation of meaning is a fundamental aspect of both disciplines. The intersection of lexicography and semantics is a fertile ground for exploring how words convey meaning and how meaning is structured in language.

Over the past century and a half, various significant theoretical approaches to understanding word meaning have left their mark on the field of lexicography. Notably, the influence of semantic theories on lexicography and the compilation of digital lexical resources has become particularly prominent with the rise of cognitive semantics, and more specifically, the advent of frame semantics, which has played a pivotal role in enhancing the way lexicographers analyse and describe word meanings, resulting in more comprehensive lexical resources.

These resources include traditional dictionaries, thesauri, and linguistic corpora, but also extend to more specialized databases, ontologies, word embeddings, and semantic networks. This remarkable range of lexical resources has had a profound impact on the field of semantics, stimulating new lines of research and enriching our understanding of how language works. These resources have found applications beyond these two linguistic fields: They are used in fields like information retrieval, sentiment analysis, machine translation, and Large Language Models development, demonstrating the practical relevance of the synergy between lexicography and semantics, which is very well represented at this conference.

EURALEX 2024 fostered valuable discussions, and promoted collaboration within the fields of lexicography and semantics, with a particular focus on cognitive approaches and their applications.

Papers, posters, presentations of new projects, and software demonstrations were invited on the topic *Lexicography and Semantics* (theoretical and applied perspectives) as well as on all lexicographically relevant topics including, but not limited to, the following fields:

- Monolingual, bilingual, and multilingual lexicography, including commercial dictionary publishing
- The dictionary-making process
- The use of dictionaries, and the needs of the user
- Lexicography and language technologies: the use of computational linguistics, natural language processing and computer science technologies and language resources in lexicography
- Terminology, terminography, and specialized forms of lexicography
- Lexicography of lesser-used and under-researched languages
- Phraseology and collocations

- Lexicography and etymology
- Lexicological issues of lexicographical relevance
- Reports on lexicographical and lexicological projects
- Academic lexicography, historical lexicography, scholarly lexicography
- Theoretical linguistic research, especially those branches most directly related to the interests and needs of compilers of dictionaries and/or computer lexicons, including word meaning, syntax, word combinations (phraseology), morphology, phonology, etc.

All submissions underwent a double-blind peer review process by at least two members of the Scientific Committee, whose support we deeply appreciate (see page 17). Decisions to accept or reject submissions for presentation at the congress and full papers for publication in the conference proceedings were based on the average review scores, and in many cases, further evaluation by members of the Programme Committee, to whom we are immensely grateful (see page 14). We extend our special gratitude to the EURALEX Board members who supported us as part of the Programme Committee: **Annette Klosa-Kückelhaus**, **Iztok Kosem**, **Robert Lew**, and **Philipp Stöckle**. Without the expertise and dedication of all colleagues serving on the Scientific and Programme Committees, we would not have been able to maintain the high academic standards of EURALEX congress presentations and proceedings. Thank you!

Our plenary speakers were invited to explore various aspects of the congress theme, *Lexicography and Semantics*. However, given the high relevance and timeliness of Large Language Models, it is no surprise that this topic resonates throughout all the keynotes and the entire conference program. We are deeply grateful to the keynote speakers for their timely, relevant, thought-provoking, and engaging talks.

In his plenary talk *Lexical Semantics, Lexicography and LLMs*, **Dirk Geeraerts** explores the evolving role of Large Language Models (LLMs) in lexical studies from both lexicographic and theoretical linguistic perspectives. From a theoretical perspective, LLMs currently constitute the epitome of distributional semantics, and distributional semantics is eminently suited as a methodological basis for usage-based cognitive semantics, allowing for a convergence of major theoretical trends in lexical semantics. From the lexicographical perspective, Geeraerts draws attention to the too often ignored process through which lexicography not only gave a major descriptive impetus to the development of corpus linguistics, but also specifically contributed to an essential step in the emergence of computational methods for corpus research: LLMs are a tool with lexicographic roots, at least to some extent. But, given that the tool has grown well beyond its original format, how does that affect its relationship to lexicography?

Tony Veale's talk *You Talk Funny! One Day Me Talk Funny Too* explores the intriguing challenge of teaching Large Language Models (LLMs) to understand and generate humour. While AI has made significant strides in various areas of human intelligence, humour remains a uniquely human trait that machines struggle to grasp. The talk emphasizes the importance of humour in AI, both for practical applications and in pushing the boundaries of human-machine interaction. It outlines current research efforts to equip LLMs with a "sense of humour" and highlights the complexities, such as navigating social norms and ensuring value alignment, given that humour often tests the limits of acceptability. The ultimate goal is to develop AI that can not only comprehend human jokes but also create its own, adding a new dimension to AI's capabilities.

In her talk *Case Studies in the Successes and Limits of Frame Semantics in Practical Lexicography*, **Kory Stamper** examines how frame semantics has influenced modern lexicography, particularly in the U.S., where traditional methods have long dominated. Frame semantics offers lexicographers a broader, more user-cantered approach to defining meanings, helping them overcome unconscious biases. However, recent technological and social shifts, such as the spread of misinformation, the rise of AI-generated text, and the changing use of online dictionaries, present challenges in further integrating frame semantics. The talk highlights case studies showcasing both the successes and limitations of this approach in the evolving field of lexicography.

Tiago Torrent talks about *Possible Futures for Semantic Lexical Resources in the Age of Artificial Intelligence* exploring the future of combining semantic lexical resources with Large Language Models (LLMs) in language technology. While LLMs have made significant strides in NLP, recent research shows their limitations, suggesting that hybrid approaches integrating LLMs with language resources may offer more potential. Tiago highlights possible research directions, including the application of the FrameNet model in developing tools to identify regions prone to gender-based violence and expanding into the field of multimodal NLP. These examples illustrate how blending AI with semantic resources could drive future advancements in language technology.

Lana Hudeček offers us valuable insight into the perspective of a lexicographer and the intricate process of dictionary making. The presentation on the *Croatian Web Dictionary – Mrežnik* highlights the development and evolution of this project, which began in 2017 and is now funded by the EU's NextGeneration program. Mrežnik is a free, online, corpus-based, monolingual, and normative e-dictionary of the Croatian language, consisting of modules for native speakers, schoolchildren, and non-native speakers. Utilizing tools like Sketch Engine and TLex, it integrates contemporary e-lexicography practices and computational linguistics methods. Mrežnik aims to be dynamic and continually updated, integrating features like gamification and recorded pronunciation. The project addresses the lack of an e-dictionary in Croatia and focuses on user-centered design, ensuring accessibility and relevance for various audiences.

I would like to extend special thanks to the keynote speakers of the *Figurative Language and Large Language Models* workshop: **Eve Sweetser** and **Alexander Ziem**.

Eve Sweetser, who has long been both an academic and personal role model, as well as a dear friend, discusses the current status and future directions of the Berkeley MetaNet project. The new MetaNet Wiki is under construction, and will (fairly soon) become accessible, including databases of French, English, and Spanish metaphors relating to (1) COVID and the pandemic; (2) cancer; and (3) climate change. This project has all of the scholarly community's shared concerns with the methods of metaphor identification and analysis: the talk therefore recaps the MetaNet procedure for searching, identifying, and verifying data. The particular focus of the talk is on the comparative aspects of the project. How can we use the wealth of documentation of English metaphors as a resource, without allowing our knowledge about English to "colonialize" our other databases?

Alexander Ziem's talk focuses on (a) examining the role of conceptual metaphors (CM) within a constructicon, (b) explaining how CMs—understood as frames with conventionalized mappings between a source and target frame—have been conceptually integrated into the constructicon, and (c) detailing the empirical methods used to identify and implement CMs within the system.

This book contains the full papers from keynotes, talks, posters, and software demonstrations presented at the XXI EURALEX International Congress. A total of sixty-seven full papers were accepted for publication. An alphabetical index at the end of this volume lists all authors and their details. The papers are thematically divided into the following chapters:

Chapter I. Lexicography and Semantics

Chapter II. Lexicography and Language Technologies

Chapter III. Dictionary Writing Systems and Lexicographic Tools

Chapter IV. Reports on Lexicographical and Lexicological Projects

Chapter V. Specialized Lexicography, Terminology, and Terminography

Chapter VI. Bi- and Multilingual Lexicography

Chapter VII. Dictionary (in) Use

Chapter VIII. Historical Lexicography.

The congress was organized by the Institute for the Croatian Language (ICL) in Zagreb, Croatia. Our sincere thanks go to all our colleagues at the ICL who supported the organization of the congress and the publication of both the Book of Abstracts and the Proceedings. We are especially grateful to **Željko Jozić** for his unwavering support, not only as the director of the Institute but also as a member of the Local Organizing Committee.

We would also like to extend our heartfelt thanks to all the **sponsors** (see page 15, 16), whose financial support made EURALEX 2024 possible. Without their generosity, this congress would not have taken place.

As Chair of the XXI EURALEX Organizing Committee, I would like to express my deep gratitude to all members of our Local Organizing Committee (see page 17). I would also like to extend my heartfelt thanks to **Annette Klosa-Kückelhaus**, who has been invaluable to us both as EURALEX President and as the exceptional organizer of EURALEX 2022. Her help and support have made this journey far smoother, and without her, it would have been significantly more challenging. Finally, a huge thank you goes to my co-editors, **Ana Ostroški Anić** and **Ivana Brač**, whose tireless efforts were invaluable not only in bringing this publication to life but also in ensuring the success of the entire conference.

Kristina Š. Despot
Chair of EURALEX 2024
September 2024

PART I: CONFERENCE OVERVIEW

Acknowledgements

We are deeply grateful to all those who generously supported the XXI EURALEX International Congress financially. Your contributions have been invaluable, and we sincerely appreciate your commitment to making this event a success.

Main Sponsors

Financira
Europska unija
NextGenerationEU

Hornby
A. S. Hornby Educational Trust

LEXiCAL
COMPUTING

JANAF

Sponsors

ZAGREB
KRAŠ
1911

Jamnica
since 1828

Republic
of Croatia
Ministry
of Culture
and Media
Republika
Hrvatska
Ministarstvo
kulture
i medija

REPUBLIKA HRVATSKA
Ministarstvo znanosti i
obrazovanja

HR CLARIN

We extend our heartfelt thanks to everyone who dedicated their time and expertise to reviewing the submissions and papers. Your invaluable contributions have greatly enriched the quality of our conference, and we are truly appreciative of your efforts.

We wish to express our deepest gratitude to the members of the Local Organizing Committee for their remarkable dedication and hard work in making this event a reality. We truly appreciate all the time and energy you've invested in creating a memorable and smooth experience for everyone.

Programme Committee

Annette Klosa-Kückelhaus
Iztok Kosem
Robert Lew
Philipp Stöckle
Ivana Brač
Milica Mihaljević
Ana Ostroški Anić
Kristina Š. Despot

Local Organising Committee

Kristina Š. Despot
Ana Ostroški Anić
Ivana Brač
Željko Jozić
Ivana Filipović Petrović
Ivana Lovrić Jović
Maja Matijević
Martina Pavić
Siniša Runjaić
Lobel Filipić

Scientific Committee

Andrea Abel
Arleta Adamska-Sałaciak
Špela Arhar Holdt
Petra Bago
Martina Bajčić
Vincent Balnat
Vuk-Tadija Barbarić
Verginica Barbu Mititelu
Slobodan Beliga
Vladimír Benko
Goranka Blagus Bartolec
Tomislava Bošnjak Botica
Petar Božović
Ivana Brač
Maja Bratanić
Úna Bhreathnach
Thomas Burch
Valeria Caruso
Paul Cook
Rute Costa
Gilles-Maurice de Schryver
Janet Decesaris
Idalete Maria Silva Dias
María José Domínguez Vázquez
Anna Dziemianko
Pamela Faber
Ilse Feinauer
Ivana Filipović Petrović
Edward Finegan
Carolina Flinz
Thierry Fontenelle
Judit Freixa
Polona Gantar
Radovan Garabík
Zoe Gavriilidou
Alexander Geyken
Voula Giouli
Oddrun Grønvik
Volker Harm
Kris Heylen
Lana Hudeček
John Humbley
Max Ionov
Dubravka Ivšić Majić
Miloš Jakubíček
Jelena Kallas
Virna Karlić

Boris Kern
Ilan Kernerman
Fahad Khan
Annette Klosa-Kückelhaus
Svetla Koeva
Veronika Kolářová
Kristina Koppel
Iztok Kosem
Simon Krek
Margit Langemets
Barbara Lewandowska-Tomaszczyk
Robert Lew
Marie-Claude L'Homme
Anja Lobenstein-Reichmann
Irina Lobzhanidze
Henrik Lorentzen
Magdalena Majdak
Tinatin Margalitadze
Saša Marijanović
Ivana Matas Ivanković
John P. McCrae
Peter Meyer
Milica Mihaljević
Fabio Mollica
Carolin Müller-Spitzer
Hilary Nesi
Henrik Nilsson
Pär Nilsson
Vincent Ooi
Christian-Emil Ore
Petya Osenova
Julia Ostanina-Olszewska
Ana Ostroški Anić
Gokhan Ozkan
Benedikt Perak
Sanja Perić Gavrančić
Ralf Plate
Boris Pritchard
Ida Raffaelli
Natascia Ralli
Geraint Paul Rees
Irene Renau
Ana Salgado
Thomas Schmidt
Hindrik Sijens
Ranka Stanković
Egon W. Stemle
Frieda Steurs
Philipp Stöckle
Kristina Štrkalj Despot
Giovanni Tallarico
Elsabe Taljard
Pius ten Hacken
Carole Tiberius
Lars Trap-Jensen
Anna Vacalopoulou
Carlos Valcárcel Riveiro
Tony Veale
Federica Vezzani
Craig Volker
Sabine Wahl
Ana Werkmann Horvat
Geoffrey Williams
Sascha Wolfer
Alexander Ziem
Tanara Zingano Kuhn
Mojca Žagar Karer
Slavko Žitnik

Overview of Keynotes and Workshops

Keynotes

Geeraerts, Dirk	Lexicography and Theories of Lexical Semantics (And the Other Way Round)
Hudeček, Lana	Croatian Web Dictionary – MREŽNIK
Stamper, Kory	Case Studies in the Successes and Limits of Frame Semantics in Practical Lexicography
Torrent, Tiago	Possible Futures for Semantic Lexical Resources in the Age of Artificial Intelligence
Veale, Tony	You Talk Funny! One Day Me Talk Funny Too. Investigating the Capacity of Large Language Models for Humour

Keynotes of the Workshop *Figurative Language and Large Language Models*

Sweetser, Eve	METANET – Its Current Status and Future Directions
Ziem, Alexander	Conceptual Metaphors in the German FrameNet Constructicon: From Conception to Implementation

Workshops

Despot, Kristina Š. and Ana Ostroški Anić	Figurative Language and Large Language Models
Krek, Simon	Large Language Models and Lexicography
Reese, Gerrraint Paul, Blanca Arias-Badia, Elisenda Bernal, and Sergi Torner	Lexicography and Accessibility

PART II: PROCEEDINGS

Leonardo Zilio and Besim Kabashi

USING NEURAL MACHINE TRANSLATION FOR NORMALISING HISTORICAL DOCUMENTS

Abstract The work with historical documents presents many challenges, not only because some sources are not well preserved, but also because grammar and spelling rules from older times were not always consistent. Still, these texts remain as a rich source of information from our history, and we could greatly benefit from the information that can be extracted from them. At the same time, the lack of spelling and grammatical consistency poses a problem for the application of computational tools, so most of the analysis work is done manually. To overcome this lack of consistency, researchers started normalising the spelling of historical documents, as this increases the performance of modern tools. Spelling normalisation is, however, also carried out manually most of the time. In this paper, we present some experiments that were done for automatically normalising historical documents in two languages: Portuguese and Albanian. Leveraging state-of-the-art large language models that were pre-trained for translation, we used corpora that were carefully curated and manually normalised to train new computational models. These models can automatically normalise documents in these languages, achieving new state-of-the-art BLEU scores above 90 for Portuguese, and up to 59 for Albanian, beating the task baselines.

Keywords historical linguistics; Albanian; Portuguese; spelling normalisation; computational linguistics; large language models; neural machine translation

1. Introduction

Historical documents have proven time and time again to be an enormous challenge for the automatic processing and extraction of information (cf. Quaresma & Finatto, 2020; Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022; Zilio et al., 2023; Vikomir & Herndon, 2024). While there is progress in the field of natural language processing (NLP) for historical texts, most of the tools are still developed with a focus on modern iterations of language, and only a few studies have been dedicated to the computational processing of historical documents in their original spellings.

To help mitigate issues caused by spelling differences, researchers started resorting to normalising the writing of historical documents (Piotrowski, 2012; Bollmann & Søgaard, 2016; Bawden et al., 2022; Vieira et al., 2024); that is, they started updating the spelling of historical texts using modern-day orthographic rules. This enables modern tools to extract data from historical documents with higher precision, as word forms in modern spelling have higher probability of having been present in their training data. However, normalisation work still is mainly done manually and is very time consuming.

Having this scenario as background, the objective of this study is to explore more automatised ways of producing a normalised text from a source historical document.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

As such, this paper extends the work of several researchers that already explored machine translation methods for text normalisation (cf. Bollmann & Søgaard, 2016; Domingo & Casacuberta Nolla, 2018; Domingo & Casacuberta Nolla, 2019), with the novelty that we investigate more recently introduced transformer-based large language models (LLMs). Another novelty is that we concentrate our efforts on languages that have been relatively less studied in terms of historical documents and automatic normalisation: Portuguese and Albanian. Our method is based on fine-tuning already existing neural machine translation (NMT) models, so we do not need large amounts of data, and a corpus composed of a few thousand words is already enough to achieve good results.

The main contributions of this paper are the following:

- An experiment using fine-tuned NMT models for the task of spelling normalisation of historical documents.
- The release of a new parallel corpus of Albanian texts, containing 12,677 tokens of original spelling and 12,836 tokens of normalised spelling, which can be used for fine-tuning NMT models.
- A comparison of performance for automatic normalisation in Portuguese and Albanian.

The remainder of the paper is organised in the following way: Section 2 discusses previous work done in the area of normalisation and automatic normalisation, while also commenting on previous efforts of applying natural language processing tools to historical documents; Section 3 presents the corpora that we used as basis for this study; Section 4 describes the methodology and the LLMs that we used for the automatic normalisation of historical documents; on Section 5 we present the results of the experiments for both languages, discussing the performance of individual LLMs; finally, on Section 6 we present our final remarks, discussing the main achievements of this study.

2. Related Work

Many authors have faced the challenge of using NLP methods for working with historical documents. Although we have not found any register for works dealing with historical data in Albanian in this setting, some studies dealt with the automatic spelling normalisation using machine translation models as basis. However, most of these studies do not take advantage of the more recent transformer models, and none that we could find uses the fine-tuning of large language models as a means to achieve spelling normalisation.

More recent studies involving NMT for spelling normalisation usually rely on an encoder-decoder, character-based architecture based on long-short term memory (LSTM) models (*cf.* Bollmann & Søgaard, 2016; Domingo & Casacuberta Nolla, 2018; Domingo & Casacuberta Nolla, 2019). While these studies make sense, by modelling the spelling normalisation problem as a character-based replacement, Tang et al. (2018) have already hinted that subword tokens can provide a better solution to character-based models. This points to the use of subword embeddings as basis for the automatic normalisation. However, the study by Tang et al. (2018) is the only one that we could find that tests subwords for this task, and neither Portuguese nor Albanian are among the tested languages.

In the subject of languages in focus, studies on automatic normalisation are concentrated around European languages. Bollmann (2019) developed a large comparison of automatic spelling normalisation methods for English, German, Hungarian, Icelandic, Portuguese, Slovene, Spanish, and Swedish. Other studies focused on fewer languages, such as the work of Domingo & Casacuberta (2019) for Slovene and Spanish, Bawden et al. (2022) for French, and Robertson (2017) for English, German, Icelandic, and Swedish. For Portuguese, we are only aware of the above-mentioned work by Bollmann (2019), who used a corpus of letters from the 15th to 19th century that was made available by the Post Scriptum project (CLUL, 2014). More recently, researchers at the University of Évora started working with text normalisation. Cameron et al. (2023) developed a categorisation of variants, which can support the normalisation of historical Portuguese texts, and Olival et al. (2023) described the normalisation of six documents belonging to the *Parish Memories* collection. In addition to text normalisation, there are studies that use historical data in different NLP tasks, such as named-entity recognition (cf. Ehrmann et al., 2023) and textual complexity (cf. Zilio et al., 2023).

As for the Albanian language we could not find any register of work being done on any automatised task related to historical documents, and this study is possibly the first of its kind, presenting a new normalised dataset and results for automatic spelling normalisation of historical documents.

3. Historical Corpora

3.1 The Portuguese Corpus

The Portuguese sample was collected from three medical handbooks published in the 18th century. It was originally released by Zilio et al. (2024) and is freely available for download on Github in several formats, including train, validation and test splits as TSV files[1]. Table 1 presents the number of tokens, sentences and overall statistics of the historical corpus as observed with AntConc (Anthony, 2005). As we can see, it is by no means a large corpus (especially considering today's age of big data), containing a little over 24 thousand tokens, but the results are nonetheless very impressive, as we will discuss later in Section 5.

This corpus was used mainly as a source of comparison for the work done on the Albanian language. This way, we add results from two new models that were not tested in the original paper (as can be seen in Section 5), and we discuss the results of a low-resourced language when compared to a higher-resourced language.

[1] Available at: https://github.com/uebelsetzer/automatic_normalisation_of_historical_documents.

Table 1: Description of the Portuguese historical corpus, considering both the original spelling and the normalised spelling. The type ratio represents the ratio between types in the original corpus and the types in the normalised corpus and is an indicator of spelling variation in the original documents. T/S = average tokens per sentence

	Portuguese historical corpus	
	Original	**Normalised**
Tokens	24,504	24,815
Types	5,584	5,341
Type ratio	N/A	1.05
Sentences	453	453
T/S	54.09	54.78

3.2 The Albanian Corpus

The Albanian corpus consisted of samples selected from three religious books from the 17th century. All tree samples were written by Pjetër Budi (1566–1622), whose Italian name variant was Pietro Budi. He was born in Gur i Bardhë, in Mat, a region in central Albania and he was an Albanian Catholic bishop.

The three selected books were published for the first time between 1618 and 1621. The first sample (ca. 20 pages) was selected from *Doctrina Cristiana*, or *Doktrina e Kërshtenë* (*The Christian Doctrine*). It was originally published in Rome, 1618 and has 286 pages (cf. Svane, 1985). The second sample (ca. 20 pages) was extracted from *Rituale Romanum*, or *Rituali Roman* (*Roman Ritual*). This book was published in Rome in 1621 and has 376 pages (cf. Svane, 1986a). The third sample (ca. 20 pages) was selected from *Speculum Confessionis*, or *Pasëqyra e t'rrëfyemit* (*The Mirror of Confession*). It was published in Rome in 1621 and has 409 pages (cf. Svane, 1986b). As usual at that time, the texts were printed following the spelling used by the author, which in some cases is not consistent.

Table 2: Description of the Albanian historical corpus, considering both the original spelling and the normalised spelling. The type ratio represents the ratio between types in the original corpus and the types in the normalised corpus. T/S = average tokens per sentence

	Albanian historical corpus	
	Original	**Normalised**
Tokens	12,677	12,836
Types	2,216	2,235
Type ratio	N/A	0.99
Sentences	307	307
T/S	41.29	41.81

Table 2 displays quantitative information about the whole corpus in its original and normalised versions. As it can be seen, the normalised corpus contains more types than the original, which indicates that, perhaps because of Albanian being a language with casus declination, the normalised version ends up having different spellings for words that were spelled in the same way in the historical documents. Another explanation for this could be that many words had to be split into separate words in the normalised version, resulting in more types.

4. Methodology

The methodology that we used in this study is very straightforward. We started by manually normalising the spelling of historical documents and then aligned the original and the normalised versions of the same text. These alignments are mostly done at the sentence level, however, for Albanian, we preferred to split sentences also at semi-colons, to have more segments available for training. These aligned segments were then used for fine-tuning existing transformer-based neural machine translation models, which were then applied to automatically normalise historical documents in the trained languages and domains.

The process of normalising a historical document consists of updating the spelling of words in a way that they match modern spelling standards of the language in focus. For Portuguese, the documents were normalised using modern Brazilian Portuguese spelling standards (cf. Zilio et al., 2024), while, for Albanian, the modern spelling norm (1972) was used. Here are examples of original and normalised versions of two short segments in Portuguese (Examples 1 and 2) and in Albanian (Examples 3 and 4):

1. He opiniaõ praticada pelo Cirurgiaõ Mór do nosso Hospital Real de Elvas Francisco Xavier , cuja sciencia he notoria . **(Original)**
2. É opinião praticada pelo Cirurgião-Mor do nosso Hospital Real de Elvas Francisco Xavier, cuja ciência é notória. **(Normalised)**
3. E kshtu mb atë ças vojta , zuna fiill përserii ; **(Original)**
4. E kështu më atë çast vajta , zura fill përsëri ; **(Normalised)**

As it can be seen from these examples, there are some different choices that were made by the normalisers of both languages. For instance, in Portuguese, the punctuation was also normalised to correspond to the standards of modern Brazilian Portuguese, while in Albanian the punctuation was kept as in the original text.

The two corpora presented in Section 3 were split into train, development, and test sets for each of the languages. This is standard procedure for preparing the corpus for fine-tuning NMT models. The two corpora were split a bit differently, based on the composition of each corpus.

In the case of Portuguese, full texts (i.e., individual chapters) were preserved in the train and development sets, and in the test set. This resulted in a division, based on number of tokens, of approximately 74% for training, 8% for development, and 18% for testing, as can be seen in Table 3.

Table 3: Train, development and test sets derived from the Portuguese historical corpus, considering both the original spelling and the normalised spelling. T/S = average tokens per sentence

	Train		Development		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	18,047	18,286	2,038	2,067	4,419	4,462
Types	3,386	3,213	826	803	1,372	1,325
Type Ratio	-	1.05	-	1.03	-	1.04
Sentences	342	342	38	38	73	73
T/S	52.77	53.47	53.63	54.39	60.53	61.12

Table 4 shows the data splits for Albanian. Because there is no clear subdivision in the texts that were used (except for coming from the three different sources), we simply split segments from each of the three samples into train, development and test set, making sure that segments from sentences that were split at semi-colon remained on the same set. This resulted in a more even split of 70% of the segments for training (~74% of tokens), 10% for development (~6% of tokens), and 20% for testing (~21% of tokens).

Table 4: Train, development and test sets derived from the Albanian historical corpus, considering both the original spelling and the normalised spelling. T/Seg = average tokens per segment

	Train		Development		Test	
	Original	Normalised	Original	Normalised	Original	Normalised
Tokens	9,345	9,484	709	695	2,623	2657
Types	1,830	1,838	299	291	777	770
Type Ratio	N/A	1	N/A	1.03	N/A	1.01
Segments	241	241	35	35	70	70
T/Seg	38.77	39.35	20.26	19.86	37.47	37.96

4.1 NMT models

The models we used in this study had to contain both Portuguese and Albanian in their set of trained languages, and they also needed to be trained in a many-to-many fashion, i.e., both languages had to have been presented both as source and as target during the training process. This severely limited our scope, as there are not many models that include Albanian in their training data, so we had to settle for three models, which were directly downloaded from the Huggingface repository (https://huggingface.co/):

- **m2m 100 (m2m100)**[2]: resulting from the work of Fan et al. (2021), the m2m models include a hundred languages trained in a many-to-many fashion, i.e., each language can be translated directly into the other, without using intermediate languages as point of reference. The base model contains 1.2

[2] For more information: https://huggingface.co/facebook/m2m100_418M.

billion parameters, but it was not supported by our graphics card, so we used a smaller version, with 418 million parameters.

- **NLLB-200 (NLLB)**[3]: the *No language left behind* paper (NLLB Team, 2022) had a huge impact on the machine translation community, as it offers the largest combination of languages to date, while focusing on low-resourced ones. While it contains many languages in its training data, it is also a less focused model, and while it works in advancing the machine translation state of the art for some low-resourced languages, it might not perform as brilliantly for highly resourced ones, such as Portuguese, when compared to models that include fewer languages. Again, we could not use the model's base version, because our graphics card did not support it, we instead used a distilled version that contains 600 million parameters.
- **SMaLL 100 (sm100)**[4]: as the name suggests, this is a much smaller version of the m2m100 model. Developed by Mohammadshahi et al. (2022), this model still contains all 100 languages trained in a many-to-many fashion, but its distillation process was focused on preserving information for low-resourced languages. Again, we have to observe how this focused distillation might have an impact on its performance for both Portuguese and Albanian.

All models were trained using an Nvidia RTX 4090 with 24GB of RAM using the standard *transformers* library as provided by Huggingface[5] (Wolf et al. 2020) for Python. All models were fine-tuned with the same parameters: learning rate of 2e-5, weight decay of 0.01, and 100 epochs (but only the best epoch was saved at the end based on BLEU score performance on the development set). Batch sizes had to be changed according to the model size, so that the GPU could load them into memory: 8 for m2m and sm100, and 6 for NLLB.

4.2 Evaluation metric

We evaluated all models using the BLEU score metric (Papineni et al., 2002), which is commonly used for assessing the performance of machine translation engines. There is much criticism against BLEU scores, because the metric indeed has several shortcomings when used for evaluating machine translation. BLEU is a metric that compares the number of n-grams in the target text with reference text(s) and produces a score from 0 to 100.

One of the downsides of BLEU is reproducibility. Because any size and combination of n-grams can be used, its reproducibility depends on a well-described methodology. To address this issue, Post (2018) introduced SacreBLEU, which is a standard way of obtaining BLEU scores. In this paper, we used the *evaluate* library's implementation of SacreBLEU for assessing BLEU scores for each model.

[3] For more information: https://huggingface.co/facebook/nllb-200-distilled-600M.

[4] For more information: https://huggingface.co/alirezamsh/small100.

[5] The *transformers* library is available on Github: https://github.com/huggingface/transformers.

Another downside of BLEU is that it assesses the target text using one or multiple reference texts. These references themselves may be questionable target texts, and they do not necessarily invalidate other equally correct translation options for a given source text. As such, a lower BLEU score might be just a reflex of different translation choices in the references. In our case, this issue is mitigated, because most of the time there is no alternative correct option for a given token in the normalisation pipeline. Most words in historical documents can only be normalised to one single form using modern spelling standards.

As a baseline for comparing the improvement that was achieved by the NMT models for the normalisation task, we used the original text and scored it against the normalised version. This way we have a baseline for what is the BLEU score in case no change is made in the original document. As for assessing the internal improvement of the NMT models with fine-tuning, we used their respective non-fine-tuned versions as baseline.

5. Results

Table 5 presents the results in terms of BLEU scores for off-the-shelf models, which serve as baselines for the fine-tuning, and fine-tuned models. The task baseline, i.e., the BLEU score of the original test set when compared to the normalised test set, is shown in the middle, as no off-the-shelf model was able to score higher than this baseline.

For Portuguese, the experiment with fine-tuning m2m100 provided a new state-of-the-art for automatic spelling normalisation in historical medical documents in the proposed dataset, as it beat the previous best score achieved by a fine-tuned mBART model. The fine-tuned m2m100 model not only had the best score, but its non-fine-tuned version also achieved the best off-the-shelf score. Models that were more focused on low-resourced languages, such as NLLB and sm100 achieved inferior results in this test set.

For Albanian, the results show that sm100 was the best model, technically tied with m2m100 because of rounding, but actually 0.0061 BLEU points ahead. As we can see from Table 5, normalising Albanian historical documents from the 17th-century is a much more complex task than 18th-century Portuguese, as shown by the task baseline of 16.78 BLEU points, compared to 62.07 in Portuguese. The off-the-shelf models had a really hard time achieving any sort of readable output, and sm100 had the best score, with 6 BLEU points, which goes to show how far behind low-resourced languages still are when compared to higher resourced ones. Upon fine-tuning, we can see a good improvement, moving the scores up to 59 BLEU points. This improvement for m2m100 and sm100 is actually higher for Albanian than for Portuguese, as it represents a jump of more than 50 BLEU points, while in Portuguese the jump was a bit over 30 BLEU points.

Table 5: BLEU-score results on Portuguese and Albanian test sets, separated into off-the-shelf models, task baseline, and fine-tuned models. * Scores for mBART were reproduced from Zilio et al. (2024) because this model had the former best performance for normalisation in the Portuguese dataset; unfortunately, this model does not cover Albanian.

	SacreBLEU	
Model	**18th-century Portuguese**	**17th-century Albanian**
Off-the-shelf models		
mBART*	30.73	N/A
m2m100	**57.98**	3.15
NLLB	40.64	2.97
sm100	57.44	**6.00**
Task baseline		
No changes in test set	62.07	16.78
Fine-tuned models		
mBART*	88.20	N/A
m2m100	**90.31**	**59.08**
NLLB	83.65	43.39
sm100	89.04	**59.08**

6. Discussion and Final Remarks

In this paper we leveraged large language models (LLMs) that were pre-trained for neural machine translation (NMT) to explore the feasibility of automatising the spelling normalisation of historical documents. As basis for our experiments, we used an existing corpus of medical texts in Portuguese and developed a completely new corpus of Albanian religious texts. This new Albanian dataset is being released with this paper and is available on Github[6].

We selected three LLMs that were pre-trained for NMT and included both Portuguese and Albanian as source and target languages. Each of these models includes a variety of languages, and we purposefully selected two models that were variants of each other, where one of the variants was a distilled version focused on low-resourced languages, so we could see the influence of removing embeddings for certain tokens when dealing with a higher resourced language, such as Portuguese, when compared to a low-resourced language, such as Albanian.

For Portuguese, a new state of the art in spelling normalisation was achieved by using the m2m100 model, which surpassed the performance of mBART, the previous state of the art for medical documents (cf. Zilio et al., 2024) by more than 2 BLEU-score points. This model also achieved the best off-the-shelf result (i.e., without fine-tuning), even if it still did not score above the task baseline. The distillation

[6] Available at: https://github.com/uebelsetzer/NMT4AlbanianNormalisation.

performed on m2m100 to create sm100 indeed reduced the size of the generated models, making them lighter to use: m2m100 models occupy 5.41GB of space, while sm100 models have a size of 3.72GB (approximately 69% of the size). However, this came at a small cost of performance for Portuguese, as the result, although marginally better than mBART, was still more than one BLEU-score point below its larger version.

For Albanian, the models achieved more than 40 BLEU points above the task baseline after fine-tuning. The distillation process focused on low-resourced languages did not seem to have much impact on the final scores, as sm100 did not achieve a much higher score than its non-distilled counterpart, and results from off-the-shelf m2m100 with 1.2 billion parameters (5.47 BLEU points) indicate that it could surpass sm100 upon fine-tuning.

As far as limitations go, there are a few elements that should be mentioned. One of the main restrictions of this paper is that we had a single RTX 4090 for running the experiments, which did not allow us to fine-tune larger models, such as the m2m 100 1.2GB[7] or the NLLB 200 3.3GB[8]. These models, when tested on the Portuguese dataset using their off-the-shelf versions, produced BLEU scores of 63.35 and 9.18, respectively. As such, this larger m2m100 was the first off-the-shelf model to score above the baseline, which points to potential state-of-the-art performance after fine-tuning, and, although the large NLLB model had an abysmal score before fine-tuning, it is a model that has great potential for improvement, as it has information collected from a couple hundred languages.

Another limitation was the size of the datasets. Because all the spelling normalisation work had to be done manually, collecting more data is a very time-consuming work. As such, the Albanian corpus, which started being constructed much later than the Portuguese corpus, is not as large, and thus there is probably room for improvement if we can manage to have a larger corpus with samples from more authors.

Even with these limitations, this study managed to contribute with a new dataset and new state-of-the-art models for automatic spelling normalisation for both Portuguese and Albanian. We hope that these resources will help advance the studies in Historical Linguistics, and we expect to use them to extend the existing datasets, as these models can potentially speed up the normalisation process, by turning it into post-editing instead of a fully manual task.

References

Anthony, L. (2005, July). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005. Proceedings. International Professional Communication Conference*, 2005. (pp. 729–737). IEEE.

[7] For more information: https://huggingface.co/facebook/m2m100_1.2B.

[8] For more information: https://huggingface.co/facebook/nllb-200-3.3B.

Bawden, R., Poinhos, J., Kogkitsidou, E., Gambette, P., Sagot, B., & Gabay, S. (2022, June). Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3354–3366).

Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis*, Minnesota, June 7, 2019. (pp. 3885–3898). Association for Computational Linguistics.

Bollmann, M., & Søgaard, A. (2016, December). Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers (pp. 131–139).

Cameron, H. F., Olival, F., & Vieira, R. (2023). Planear a normalização automática: tipologia de variação gráfica do corpus das Memórias Paroquiais (1758). *LaborHistórico*, Rio de Janeiro, ISSN, 2359-6910.

CLUL. 2014. P.S. Post Scriptum. *Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna.* https://www.clul.ulisboa.pt/projeto/ps-post-scriptum.

Domingo, M., & Casacuberta Nolla, F. (2018). Spelling normalization of historical documents by using a machine translation approach. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*: 28–30 May 2018, Universitat d'Alacant, Alacant, Spain (pp. 129–137). European Association for Machine Translation.

Domingo, M., & Casacuberta Nolla, F. (2019, September). Enriching Character-Based Neural Machine Translation with Modern Documents for Achieving an Orthography Consistency in Historical Documents. In *International Conference on Image Analysis and Processing* (pp. 59–69).

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, *56*(2), 1–47.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., et al. (2021). Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107), 1–48. https://doi.org/10.48550/arXiv.2010.11125

Mohammadshahi, A., Nikoulina, V., Bérard, A., Brun, C., Henderson, J., & Besacier, L. (2022, December). SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 8348–8359).

NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672.*

Olival, F., Cameron, H. F., Farrica, F., & Vieira, R. (2023). As Memórias Paroquiais (1758) do atual concelho de Vila Viçosa. *Callipole*: *revista de Cultura*, *29*, 85–128.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Piotrowski, M. (2012). *Natural language processing for historical texts.* Morgan & Claypool Publishers.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*: Research Papers (pp. 186–191), Brussels, Belgium.

Quaresma, P., & Finatto, M. J. B. (2020). Information Extraction from Historical Texts: a Case Study. In *DHandNLP @ PROPOR* (pp. 49–56).

Robertson, A. (2021). *Automatic Normalisation of Historical Text.* PhD thesis, School of Informatics, University of Edinburgh.

Svane, Gunnar (Ed.). (1985). *Pjetër Budi. Dottrina christiana (1618). With a transcription into modern orthography and a concordance (Sprog og Mennesker 9),* Institut for Lingvistik, Aarhus Universitet. Århus.

Svane, Gunnar (Ed.). (1986a). *Pjetër Budi. Rituale Romanum (1621). With a transcription into modern orthography and a concordance (Sprog og Mennesker 11),* Institut for Lingvistik, Aarhus Universitet. Århus.

Svane, Gunnar (Ed.). (1986b). *Pjetër Budi. Speculum Confessionis (1621). With a transcription into modern orthography and a concordance (Sprog og Mennesker 13)*, Institut for Lingvistik, Aarhus Universitet. Århus.

Tang, G., Cap, F., Pettersson, E., & Nivre, J. (2018, August). An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1320–1331).

Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401.* https://doi.org/10.48550/arXiv.2008.00401

Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., & Santos, I. (2021). Enriching the 1758 Portuguese Parish Memories (Alentejo) with named entities. *Journal of Open Humanities Data*, *7*(20).

Vieira, R., Cameron, H., Olival, F., Farrica, F., Finatto, M., Banza, A., Ribeiro, A., Trojahn, C. PLN e Humanidades Digitais In H. M. Caseli, & M. G. V. Nunes (Org.). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024.* https://brasileiraspln.com/livro-pln/2a-edicao/parte-dominios/cap-humanidades-digitais/cap-humanidades-digitais.html.

Vilkomir, K., & Herndon, N. (2024, April). Challenges of Automatic Document Processing with Historical Data. In *Proceedings of the 2024 ACM Southeast Conference on ZZZ* (pp. 50–59).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*: system demonstrations (pp. 38–45).

Zilio, L., & Finatto, M. J. B. (2022, March). Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022)*, Fortaleza, Brazil, 21st March, 2022 (pp. 1–10). CEUR.

Zilio, L., Finatto, M., Vieira, R., & Quaresma, P. (2023). A natural language processing approach to complexity assessment of 18th-century health literature. *Domínios de Lingu@gem*, *17*.

Zilio, L., Lazzari, R. R., & Finatto, M. J. B. (2024) Can rules still beat neural networks? The case of automatic normalisation for 18th-century Portuguese texts. In *Proceedings of the International Conference on the Computational treatment of Portuguese*, volume 2 (pp. 86–95). https://aclanthology.org/2024.propor-2.0.pdf

Acknowledgements

We would like to thank the Chair of Computational Corpus Linguistic at the Friedrich-Alexander-Universität Erlangen-Nürnberg for the support.

Contact information

Leonardo Zilio
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
leonardo.zilio@fau.de

Besim Kabashi
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
besim.kabashi@fau.de