UNIVERSITY OF WOLVERHAMPTON

UNIVERSIDAD DE MÁLAGA

NEW BULGARIAN UNIVERSITY

GHENT UNIVERSITY

EM TTI

With the support of the Erasmus+ Programme of the European Union

Definition Modelling for English and Portuguese: a comparison between models and settings

Anna Beatriz Dimas Furtado

Supervisors:

Dr María Rosario Bautista Zambrana (University of Malaga)

Prof Dr Ruslan Mitkov (University of Wolverhampton)

A report submitted in partial fulfilment of the requirements
for the Masters in Technology for Translation and Interpreting degree

June 1, 2022

Declaration

EUROPEAN MASTER'S IN TECHNOLOGY FOR TRANSLATION AND INTERPRETING (EM TTI)

University of Malaga, University of Wolverhampton

(i) This work or any part thereof has not previously been presented in any form to the Universities or to any other institutional body whether for assessment or for other purposes. Save for any express acknowledgements, references and/or bibliographies cited in the work, I confirm that the intellectual content of the work is the result of my own efforts and of no other person.

(ii) It is acknowledged that the author of any project work shall own the copyright. However, by submitting such copyright work for assessment, the author grants to the Universities a perpetual royalty-free licence to do all or any of those things referred to in section 16(i) of the Copyright Designs and Patents Act 1988 (viz: to copy work; to issue copies to the public; to perform or show or play the work in public; to broadcast the work or to make adaptation of the work).

(iii) I have read the University of Wolverhampton's Ethics Guidance[1]. Where ethical issues have been identified in relation to my proposed research project, I have sought ethical approval from the EM TTI Ethics Committee, explicitly indicating how I intend to address these in my research.

Name: Anna Beatriz Dimas Furtado

Signature: [Signature redacted]

Date: 01 June 2022

[1] https://www.wlv.ac.uk/research/research-policies-procedures--guidelines/ethics-guidance/

Abstract

Definitions are key for many areas of knowledge; they convey meaning, refer to product conceptualization and naming, facilitate communication, provide clarity, and pervade all areas of human activity. Hence, having access to definitions is essential for many professions, but it is crucial for translation and interpreting. Definition Modelling (DM) is a task concerned with automatically generating definitions from embeddings. While most approaches to DM covers only English, this project aims at generating definitions for both Portuguese and English. DM is tackled with two deep-learning models as a sequence-to-sequence task in this research. Experiments are performed in three different settings - monolingual, cross-lingual, and multilingual based on various corpora and different embeddings. Given the lack of resources, the first dataset for Portuguese DM is developed. Both intrinsic and extrinsic evaluation is conducted. Results show that adopting the pre-trained MT5 model yield better results than non-pre-trained models for monolingual settings. Besides that, Flair-embeddings fare better than both character-based and transformer-based embeddings in non-pre-trained embedding. Human evaluation suggests that automatically generated glosses are useful for translators, although post-editing may be required to achieve optimal quality.

Keywords: Definition Modelling. Deep Learning. Cross-lingual Learning. Multilingual Learning. Lexicography for Translation.

Resumo

As definições são fundamentais para muitas áreas do conhecimento; elas transmitem significado, referem-se à conceituação e nomeação de produtos, facilitam a comunicação, proporcionam clareza e permeiam todas as áreas da atividade humana. Portanto, ter acesso às definições é essencial para muitas profissões, mas é crucial para a tradução e interpretação. A modelagem de definições (DM) é uma tarefa que se preocupa em gerar definições automaticamente a partir de embeddings. Enquanto a maioria das abordagens de DM abrange apenas o inglês, este projeto visa gerar definições tanto para o português quanto para o inglês. A DM é abordada com dois modelos de aprendizagem profunda como uma tarefa de seqüência a seqüência nesta pesquisa. Os experimentos são realizadas em três configurações diferentes - monolíngue, inter-língua e multilíngue com base em vários corpora e diferentes embeddings. Dada a falta de recursos, é proposto o primeiro conjunto de dados para a DM portuguesa. São realizadas avaliações tanto intrínsecas quanto extrínsecas. Os resultados mostram que a adoção do modelo pré-treinado MT5 produz melhores resultados do que modelos não pré-treinados para ambientes monolíngues. Além disso, Flair embeddings produzem resultados melhores do que ambos embeddings baseados em caracteres e transformers. A avaliação humana sugere que as definições geradass automaticamente são úteis para tradutores, embora a pós-edição possa ser necessária para atingir a qualidade ideal.

Palavras-chave: Modelagem de definições. Aprendizagem profunda. Aprendizagem Transversal. Aprendizagem Multilingue. Lexicografia para Tradução.

Resumen

Las definiciones son clave para muchas áreas del conocimiento; transmiten significado, se refieren a la conceptualización y la denominación de productos, facilitan la comunicación, aportan claridad e impregnan todos los ámbitos de la actividad humana. Por eso, tener acceso a las definiciones es esencial para muchas profesiones, no obstante, es crucial para la traducción y la interpretación. El modelado de definiciones (DM) es una tarea que se encarga de generar automáticamente definiciones a partir de vectores. Mientras que la mayoría de los enfoques de DM sólo abarcan contenido en inglés, este proyecto pretende generar definiciones tanto para el portugués como para el inglés. En esta investigación, el DM es abordado con dos modelos de aprendizaje profundo como una tarea de secuencia a secuencia. Los experimentos se llevan a cabo en tres escenarios diferentes: monolingüe, multilingüe y multilingüe basado en varios corpus y diferentes vectores. Dada la falta de recursos, se propone el primer conjunto de datos para DM en portugués. Se realiza una evaluación intrínseca y extrínseca. Los resultados muestran que la adopción del modelo pre-entrenado MT5 produce mejores resultados que los modelos no pre-entrenados para entornos monolingües. Además, los vectores Flair obtienen mejores resultados que los vectores basados en caracteres y en transformadores en modelo no pre-entrenados. La evaluación humana sugiere que las definiciones generadas automáticamente son útiles para los traductores, aunque puede ser necesaria la posedición para lograr una calidad óptima.

Palabras clave: Modelización de definiciones. Aprendizaje profundo. Aprendizaje cruzado. Aprendizaje multilingüe. Lexicografía para la traducción.

Résumé

Les définitions sont essentielles dans de nombreux domaines de la connaissance ; elles transmettent le sens, renvoient à la conceptualisation et à la dénomination des produits, facilitent la communication, apportent de la clarté et sont omniprésentes dans tous les secteurs de l'activité humaine. Par conséquent, avoir accès aux définitions est essentiel pour de nombreuses professions, mais il est crucial pour la traduction et l'interprétation. La modélisation des définitions (DM) est une tâche qui consiste à générer automatiquement des définitions à partir des vecteurs. Alors que la plupart des approches de DM ne couvrent que l'anglais, ce projet vise à générer des définitions pour le portugais et l'anglais. Le DM est abordé avec deux modèles d'apprentissage profond comme une tâche de séquence à séquence dans cette recherche. Les expériences sont réalisées dans trois contextes différents - monolingue, interlingue et multilingue - sur la base de divers corpora et de différents contexte. Compte tenu du manque de ressources, le premier jeu de données pour le DM portugais est proposé. Une évaluation intrinsèque et extrinsèque est menée. Les résultats montrent que l'adoption du modèle MT5 pré-entraîné donne de meilleurs résultats que les modèles non pré-entraînés pour les paramètres monolingues. En outre, les incorporations Flair donnent de meilleurs résultats que les incorporations basées sur les caractères et les transformateurs dans le modèle non pré-entraînées. L'évaluation humaine suggère que les glosses générés automatiquement sont utiles aux traducteurs, bien qu'une post-édition puisse être nécessaire pour obtenir une qualité optimale.

Mots clés : Modélisation des définitions. Apprentissage profond. Apprentissage interlinguistique. Apprentissage multilingue. Lexicographie pour la traduction.

Contents

List of Acronyms

AI – Artificial Intelligence
API – Application Programming Interface
BART – Bidirectional Auto-Regressive Transformers
BERT – Bidirectional Encoder Representations from Transformers
BLEU – Bilingual Evaluation Understudy
CAT Tool – Computer Assisted Translation Tool
COBUILD – Collins Birmingham University International Language Database
CodWoe – Comparing Words and Dictionaries Embeddings CSS – Cascading Style Sheets
CTT – Communicative Theory of Terminology
DC – Definition Contexts
DfTs – Dictionaries for Translation
DL – Deep Learning
DM – Definition Modelling
ELMo – Embeddings from Language Models
EN – English
ES – Spanish
FFN – Feed-Forward Network
FR – French
FSL – Few-Shot Learning
FTL – Function Theory of Lexicography
GRUs – Gated Recurrent Units
GTT – General Theory of Terminology
HTML – HyperText Markup Language
ISO – International Organization for Standardization
IT – Italian
JSON – JavaScript Object Notation
kNN – k-Nearest Neighbour
LD – Lexicographic Definitions
LGP – Language for General Purposes
LRLs – Low-Resource Languages
LSP – Language for Specific Purposes
LSTM – Long-Short Term Memory Layer
METEOR – Metric for Evaluation of Translation with Explicit Ordering
ML – Machine Learning
MT5 – Multilingual Text-To-Text Transfer Transformer
NLP – Natural Language Processing

NN – Neural Networks
PT – Portuguese
QE – Quality Estimation
ROUGE – Recall-Oriented Understudy for Gisting Evaluation
RNN – Recurrent Neural Network
seq2seq – sequence-to-sequence
SM – Similarity Measures
SOTA – State-of-the-Art
TD – Terminological Definitions
URL – Uniform Resource Locator
UTF–8 – UCS (Unicode) Transformation Format
vec2sec – vector-to-sequence
word2vec – word-to-vector

List of Figures

List of Tables

Acknowledgements

> "Part of being an explorer is overcoming your fears, but when you do that, that's right when you get to see the magic."
>
> Welcome to Earth - NatGeo

Doing an MA is a marathon in itself. But doing an MA abroad amidst a pandemic was quite the journey. There were moments when I felt sad, anxious, desperate, and hopeless. But there were moments when I felt loved, empowered, and cherished because I had great people around me; to these people, I dedicate this work.

I started this MA thinking I knew something about python, only to find out that I solely had a glimpse of the tip of the iceberg. A whole new world was unfolded right before my eyes, and I felt marvelled but also afraid of never being able to make it till the end. But here I am at last.

Every single day was a struggle; I kept fighting every day. Fighting to understand this new area I wanted to dive in. Fighting to absorb these brand new concepts I didn't even know existed to start with. Fighting to cope with losing people to the COVID-19. Fighting anxiety, losing, standing up, and fighting again. Fighting to accept that my language abilities were not as good as I thought.

It was hard to understand and accept that I wasn't alone either. I owe a debt of gratitude to my partner in crime, Rafael. Thank you for translating the HEI++ Dataset into Portuguese, and for remotely granting access to your PC for scraping purposes. We have been together for years now, but you chose to do your MA in Dublin to stay close to me. You weren't physically present, but you showed up every single day. Thanks for being my safety net, where I could simply fall down and rest. We have the whole life together, but I will never be able to be half of what you mean to me.

My deepest heartfelt gratitude to my parents, who have never understood half of what I was doing but stood by my side anyway. While I was writing this thesis, my father fought for his life in the hospital. It was hard not to abandon everything and go to him. Dad, you inspired me to fight harder. Luckily, we both won, and I cannot wait to give you the warmest hug.

I thank Dr Bautista Zambrana for her immense patience in going through all my huge emails and endless doubts, and Prof Mitkov for his comments, supervision and support. It took me some time to process and understand where I was, what I was doing, but you were both patient throughout the whole process.

Surviving the lockdown in Malaga was the hardest thing I have ever done in my life, followed by this thesis. If I survived, it was only because I had the craziest MA-peers who also accepted this challenge. Many thanks to Nataliya, Annalisa, Stiven and Maral. We should have tried to write our literature reviews at the beach. Speaking of beaches, I'd like to thank Marine for being so lovely and agreeing to walk with me by the shoreline every week. I felt guilt, I though I was slow soo many times, but you were nothing more than lovely to me. You are my big Armenian sys. I simply admire you, and I hope we get to have fun again.

During my first year, I felt deeply lost. I didn't feel safe making mistakes, and a series of unfortunate events befell me. However, in my placement, out of a sudden things started to make sense. This is thankfully due to Dr Orasan, who supervised me and helped me see things differently. Before and during my data collection, every single person I asked about scraping told me that I should not proceed because of the challenge. Instead, you disagreed, and, more importantly, encouraged and believed in me. I am profoundly grateful for your patience and willingness to discuss the half-baked ideas I had. If I could build not one but two scrapers, then I could learn anything from then on.

A vast, enormous shoutout to the Coding Clinics, Booster Sessions, and all extra support I received from the RGCL team. Thank you to all teachers who allowed me to audit their classes. I don't know if you noticed, but I was trying to verify if I could learn through osmosis by showing up in all activities (!) ; at the end, I learnt a lot (not through osmosis, though), but I am sure I became the special chaser in the department.

A special thanks to Dr Ranasinghe, who displayed incredible patience when I used to send one email per day (I'm sure you realised my mind was more confused than word salad). Thank you very much for being the kindest teacher I've ever had. I made the most awful and careless mistakes, and you didn't judge or discourage me even once. Thank you for not stifling the million ideas I had. This gave me the strength to go from basics to beyond and lose the fear of playing with the code. Thank you for your precious time, for sharing the knowledge on cross-lingual, multilingual learning, MT5, and everything else with me. If I am here, it is greatly because of you.

I have to send the most sincere thanks to Dr Blain, who I kindly call the *police of the server*; thanks for all the inspiring tutorials (and they were many), the most carefully made classes (and slides!) and for simply listening. You must know that I admire you not only for your engineering skills but for being one of the most compromised lecturers I have ever met. You suggested me to incorporate QE in my thesis, and it worked! I would oftentimes show up with a minor bug (certainly was the end of the world to me), and you'd solve them and say that it was nothing. Well, for me, it meant everything. Thank you for seeing potential in me, for introducing me to Bash and the RGCL server, and for always being there. Lastly, thank you for reminding us that nobody is born with a degree and that everything is learnable if I have the guts to work hard.

Thanks to Dr Moze for being gentle and receptive. You have always been comforting to all of us, but I cannot forget all the times you reminded me to rest and take care of myself. Your lightness and fun mood have always made all meetings and classes we had such a great time that it was fun to learn again.

Many people were cheering me up from the other side of the globe. I want to express my undying gratitude to my Brazilian friends (Jeferson, Augusto, Gustavo, Patrícia, and many more) who never forgot about me even once. Many thanks to Dr Rebecchi for allowing me to audit one of her modules in the distance. Also, I am most grateful to Elisa (Dr Teixeira), who supervised me in my undergraduate studies and continued to follow my journey online. One month before finishing, I was falling apart but Cláudio (Dr Côrrea) pointed me back to the right direction. Back in the undergrad we would stay until late trying to put the glossary live. It was very challenging, but we had fun. Why couldn't I have fun again? Well, at some point, it came back to me. Thank you, both of you kept the fire of language technologies lit in my heart. I hope we can fly even higher in the future.

Thanks to the Union for Girls against Procrastination, formed by that's right, Lilit and I. I've never felt alone because we would always be together at the library, every day, sometimes Saturday, sometimes Sunday. We kept each other company, shared food, reviewed our codes, and felt the urge to destroy our laptops; we took care of each other. Lilit, I feel like you are my little sister. I hope I have not been too harsh on you because I've never doubted your abilities for even a second, and I won't allow you to do so. Thanks for listening, sharing, and living. Thank you for helping me with scraping, with writing, with LaTeX, and so much more... We've had a great time together, and I hope your journey in Malaga be as adventurous as mine was this whole time.

Most importantly, to the Creator of all things, thanks for sending the right people in the right moment; if the universe works in favor of some people, I am certainly one of the lucky ones.

A heartwarming thanks to both Malaga and Wolverhampton admin teams who followed and helped me navigate the never ending bureaucratic issues chasing me.

Lastly, I am grateful for have been granted an Erasmus Mundus scholarship. Many thanks to the EC and the Consortium for this opportunity. Gladly, I had the chance to see a little bit of the world, share different points of view, meet some of my academic idols, participate in courses I would never be able to pay for, and learn a lot.

I cannot believe I did this and it is now time to share it! If you read until here, I hope you know that is not the end, but just the beginning. Cheers to what is ahead.

Chapter 1

Introduction

Definitions are essential in the Information Society. They spread the knowledge throughout areas, cultures, and countries and permeate all areas of human activity. The translation and interpreting field is not different. As part of their everyday work, professionals working in the translation field rely on different resources to support their linguistic decisions. By and large, the relationship between translators and lexicographic resources has been thriving for years now.

Nevertheless, producing high-quality definitions is no trivial task and can be rather time-taking and labour-intensive, presenting a significant challenge for lexicographers and terminographers. From this perspective, recent advances in natural language processing and deep learning can foster the automatic generation of definitions and optimise the work of lexicographers, terminographers, and scholars engaged in this process with a view to facilitating resources for translators promptly.

Definition Modelling (DM) is a natural-language generation task first proposed in 2017 whose goal is to generate definitions from word representations (Noraset et al. 2016). The task's initial purpose was to evaluate the quality of embeddings and provide some understanding of the relationship between vectors and meaning (ibid).

At first, DM's goal was to evaluate the quality of word embeddings. Soon, it became clear its application in language learning (Yang et al. 2020) and lexicography (Bevilacqua et al. 2020). Moreover, recent investigations showed that it is possible to generate monosemic definitions (Noraset et al. 2016) and polysemic definitions (Ishiwatari et al. 2019) in English (Gadetsky et al. 2018), Chinese (Yang et al. 2020), French (Mickus et al. 2020), and other languages (Kabiri & Cook 2020).

In spite of that, investigations attempting to leverage comparable corpora to generate definitions in both Portuguese and English have never been carried out. Nevertheless, they can be key for developing bilingual and multilingual resources for translation and interpreting working with these languages.

Recent advances in deep-learning techniques have encouraged the inclusion and production of resources for languages other than English by employing different techniques. For example, in cross-lingual learning (Adams et al. 2017), similar languages are added in training, assuming that the model can learn from the similarities between the languages. In multilingual learning (Adams et al. 2017), two or more languages are combined to learn together, and outputs are generated for both languages simultaneously.

1.1 Research Questions

Thus, the ultimate goal of this thesis is to investigate whether it is possible to generate definitions for Portuguese and English by using two different models (a vec2seq baseline and the MT5 model) in three different settings (monolingual, cross-lingual, and multilingual). This goal is pursued by examining the following research questions (and sub-questions):

RQ 1: How effective are deep-learning techniques in generating glosses for Portuguese?

1. Is it best to cast definition modelling as a vector-to-sequence or as a sequence-to-sequence task?
2. Does the adoption of pre-trained models result in an increase in automatic scores or quality for Portuguese?
3. How effective are monolingual character-based embeddings at enabling the generation of definitions for Portuguese in a monolingual learning setting?
4. How effective are monolingual transformer-based embeddings at enabling the generation of definitions for Portuguese in a monolingual learning setting?

RQ2: Can we apply cross-lingual learning techniques to generate definitions for Portuguese?

1. How effective are monolingual character-based embeddings at enabling the generation of definitions for Portuguese in a cross-lingual learning setting?
2. How effective are cross-lingual transformer-based embeddings at enabling the generation of definitions for Portuguese in a cross-lingual learning setting?

RQ3: Can we apply multilingual-learning techniques to definition modelling?

1. Can we generate definitions for English and Portuguese simultaneously?
2. How effective are monolingual character-based embeddings at enabling the generation of definitions for Portuguese in a multilingual learning setting?
3. How effective are cross-lingual transformer-based embeddings at enabling the generation of definitions for Portuguese in a multilingual learning setting?
4. How effective are cross-lingual transformer-based embeddings at enabling the generation of definitions for English in a multilingual learning setting?

RQ4: Is it possible to align Portuguese-English glosses automatically?

1. is it possible to apply sentence embeddings and the cosine measure to map Portuguese definitions to English ones?

RQ5: Are the automatically generated glosses useful for translators?

1. Is post-editing still required?

As mentioned above, to answer these questions, two encoder-decoder transformer-based models are utilized to conduct the experiments. The baseline model comprises a non-pre-trained model provided by the Comparing Words and Dictionaries Embeddings (CodWoe) Sharedtask (Mickus et al. 2022). The second model harness the power of the massive Multilingual T5 (MT5) (Xue et al. 2021), fine-tuned as it is in its small version, to perform Definition Modelling.

Results are evaluated in terms of BLEU (Papineni et al. 2002), BERTScore (Zhang et al. 2019), and MoverScore (Zhao et al. 2019). Moreover, with a view to obtaining more information regarding the quality of the glosses, Quality Estimation is employed through TransQuest (Ranasinghe et al. 2020*a*). Should the system produce meaningful definitions, qualitative experiments are devised to assess whether automatically generated glosses are useful for translators and whether they prove to be of good quality. For these, questionnaires were be applied, and the inter-annotator agreement was calculated.

1.2 Research Hypothesis

For carrying out this research, the following hypotheses are raised:

1. Contexts are widely found in dictionaries.

2. Adopting a pre-trained model provides better results than training a model from scratch, given that pre-trained models can provide encyclopaedic and general knowledge, which lexicographers rely heavily on to produce definitions.

1.3 Research Objectives

In order to answer the research questions, the following goals were established:

1. Provide the first dataset to enable research on Portuguese DM.

2. Provide baseline results on which future studies about DM can be based.

3. Investigate whether pre-trained models can be applied to Portuguese DM.

4. Investigate which type of embedding (character-based or transformer-based) produces better results with the baseline architecture.

5. Investigate whether cross-lingual and multilingual settings can produce reasonable results for DM.

6. Investigate whether the MT5 model can perform DM.

7. Employ Quality Estimation for DM and provide quality-related information for translators.

8. Investigate whether DM can produce useful results for translators.

1.4 Original Contributions

Ultimately, this project harnesses the synergy between NLP, lexicography, and translation to bridge the existing gap in research on Portuguese Definition Modelling. Consequently, this thesis:

1. makes available the first dataset for Portuguese DM to the scientific community.
2. makes available to the scientific community the first dataset containing only multiword expressions (PT_HEI).
3. inaugurates DM for Portuguese by providing a baseline to which future works can be compared.
4. provides insights into generating glosses for two languages simultaneously by the first time.
5. applies cross-lingual and multilingual learning on Portuguese DM for the first time.
6. applies Quality Estimation on DM for the first time.
7. conducts for the first-time human evaluation on DM with translators.

1.5 Organisation of this thesis

This thesis is divided into 6 chapters. After the introduction, Chapter 2 offers the background knowledge and key concepts on lexicography for translation, deep learning, and definition modelling.

In Chapter 3, the methodology adopted to conduct this research is provided. First, the process of data collection, cleaning and pruning, and storage of data for all languages is described. Then, the two deep-learning models are presented, along with a in detail explanation of the experiments. Finally, both extrinsic and intrinsic evaluation are covered.

Then, Chapter 4 provides details on both intrinsic and extrinsic evaluation of the models, followed by Chapter 5, where the discussion of the results is presented along with the limitations of this research.

Lastly, Chapter 6 provides the conclusion for this thesis.

Chapter 2

Background Knowledge

Overview – This chapter introduces the theoretical background and the related work underpinning this research. To start with, Section 2.1 provides an overview of the key issues and trends in lexicography. Then, in Section 2.2, a comprehensive review of definitions and their features is offered. After that, deep learning core concepts and techniques are covered. Finally, a detailed survey of definition modelling and related work is presented.

2.1 Lexicography: Theories and Users

The end of the XX century marks the rise of the Information Society, defined byCastells (2003) as a network-functioning society founded upon the power supplied by information. The recent technological advancements made possible remarkable progress in daily life. Currently, cyberspace allows people to access information at such a rapid pace, as never seen in human history.

Technical and specialised knowledge are under constant creation and change. Now, more than ever, new knowledge needs to be spread in a timely manner; the coronavirus pandemic is a case in point. However, the systematisation of this knowledge in databases, glossaries, and dictionaries occurs at a much slower rate, posing a significant challenge to terminologists, lexicographers, and ultimately, translators.

Many different stakeholders are interested in specialised knowledge, and their needs vary accordingly. In any case, understanding a specialised concept is imperative, making definitions then a valuable resource for any potential audience. The main source where to find definitions is the dictionary par excellence.

Lexicographic products – lists of words from old texts, collections of words, lists of names and things – emerged in ancient civilisations (Béjoint 2016); the dictionary, however, is a modern invention, arising in the Middle Ages (ibid.). All these works (henceforth lexicographic resources) were intrinsically built to “show and explain, to teach and facilitate the acquisition of knowledge and the mastery of language, to educate and socialise, and ultimately to participate in the cohesion of the community" (ibid.).

Lexicography, therefore, is traditionally the discipline concerned with dictionaries and reference sources, whose main purpose is to provide a particular audience with information (Tarp 2018). While this definition of lexicography is widely accepted in the lexicographic community, the disciplinary status of lexicography is still somewhat controversial.

Some scholars consider lexicography as simply an "**art and craft**" (Landau n.d.), where it is an "instance of applied linguistics" (Meier 2003). Others, nevertheless, describe lexicography as **a science** subordinated to linguistics (Kudashev 2007, Tarp 2018), with its own theory, tasks, and methods. A third group believes lexicography is an independent **discipline** (Wiegand 2013); this view differentiates between lexicography (a scientific practice) and metalexicography (a scientific research area). Finally, a fourth group deems lexicography to be a fundamentally independent science with a solid interdisciplinary tendency (Tarp 2008).

Since lexicography has origins in ancient times, it has never depended upon linguistics or applied linguistics to operate; there are many lexicographic resources where linguistic knowledge has not been required (Tarp 2018). Moreover, lexicography is intrinsically interdisciplinary, mainly because it combines a range of disciplines to observe, describe, and produce its subject matter: dictionaries and other lexicographical reference works (ibid.).

In this sense, lexicography meets all the requirements to be considered a science (Tarp 2008): i) an object of study: dictionaries and lexicographical resources (and its planning, production, design, and usage); ii) it is underpinned by categories, concepts, hypothesis, and theories; iii) it covers its historiography (from both dictionaries and pre-theoretical ideas); iv) it contributes independently to methodologies; v) it is constituted of both practice and theory.

In the same fashion, much has been argued about whether there is a lexicographical theory (Atkins & Rundell 2008). Tarp (2018), nevertheless, classifies lexicographic theories as contemplative lexicographical theories, whose focus is only to describe the existing reality, mainly existing "old dictionaries", and transformative lexicographical theories, where the focus is on both describing the reality and guiding and improving lexicographic products (ibid.).

Despite the controversiality, many theories have been proposed in the field: the General Theory of Lexicography (Ščerba 1995, Wiegand 1998), the Theory of Bilingual Lexicography (Duda 1986), the Theory of the Lexicographical Example (Hausmann 2017), and the Theory of the Dictionary Form Mann (2011).

A significant shift in lexicography, however, occurred with the Function Theory of Lexicography (FTL) (Bergenholtz & Tarp 2002). Previously, only a few dictionaries placed the user and their needs at the centre of lexicography. As dictionaries are sources for other people to consult, it is only natural to shift the focus to the ultimate goal of lexicographic production: to satisfy an information need.

FTL is underpinned by the assumption that dictionaries are objects of use, produced or should be produced to meet specific social needs. These necessities are to be viewed as concrete; they are connected to specific types of users in specific types of social situations. In order to satisfy these needs, different types of lexicographic data and methodologies are utilised and made available in specific types of dictionaries (Bergenholtz & Tarp 2002, Tarp 2008).

2.1.1 Dictionaries, Translators, and Needs

Translators, for example, are a particular case of language learners. During a period, bilingual dictionaries and translation dictionaries were perceived as one and the same (Tarp n.d., 2007, 2013, Fuertes-Olivera 2013, Giacomini 2018). This trend followed for many years, especially when dictionaries were either classified in terms of language, as monolingual, bilingual, or multilingual (ibid.), or in terms of usage aim (Kühn 1989), where translation purposes are seldom mentioned.

According to Giacomini (2018), "all bilingual dictionaries are designed for assisting translation" if translation means understanding or producing a text in L2. Still, bilingual dictionaries do not have translators as their primary users. There is, therefore, a strong need of detaching bilingual dictionaries from **Dictionaries for Translation** (DfTs), whose main purpose should be meeting the needs of students and professionals working in the field of translation.

However, it is particularly complex to produce dictionaries exclusively for translators since they compose a highly heterogeneous group with different needs, varying according to each phase of the translation process.

For each stage of the translation process, see (Bowker 2012, Tarp 2013), translators will need different solutions to satisfy each need. For instance, in the pre-translation phase, a monolingual solution in the L1 may suffice in supplying background information (general introduction to the subject and definitions for terms and unknown words) (Bowker 2012, Tarp 2013). For the translation phase, both bilingual and monolingual (in the L2) solutions will be required, given the need for background information, definitions, equivalents, collocations, expressions, register, grammar, and genre-specific conventions) (Bowker 2012, Tarp 2013). Finally, depending on who is working on the last phase, a monolingual solution will help more native speakers of the L1, a bilingual solution is preferred by non-native speakers, and both monolingual and bilingual solutions are required for comparing texts in L1 and L2 (Bowker 2012, Tarp 2013).

Given that dictionaries have existed now for centuries, it is surprising how little empirical user research has been carried out about dictionary use, especially in this day and age, when practicality is intrinsic to e-dictionaries. Overall, research conducted on dictionary use has heavily focused on foreign language learners or translation students (Béjoint 1981, Roberts 1992, Duvå & Laursen 1994, Dancette & Réthoré 1997, Mackintosh 1998, Varantola 1998, Hartmann 1999, Pastor 2001, Ramos & del Mar 2005, Bogaards 2005, East 2008). Professional translators, however, are usually underrepresented in these investigations (Durán Muñoz 2010).

Tarp (2013) overall review of translator's needs demystifies common misconceptions regarding dictionary use among translators. For instance, students and beginners usually need more general and specific background information about the subject field (Duvå & Laursen 1994, Varantola 1998, Tarp 2013). Rarely, however, experienced translators will use only bilingual dictionaries (Tomaszczyk 1989, Varantola 1998, Tarp 2013). In any case, both experienced and beginner translators will need background, terminological, linguistic, and/or grammatical information at least occasionally (Nord 2002, Tarp 2013).

These remarks are confirmed in Durán Muñoz (2010) comprehensive survey on professional translator's needs, where it was possible to shed light on what translators consider essential data for the resources they consult, including clear and concrete definitions, equivalents, derivatives and compounds, domain specification, examples, phraseological information, bilingual definitions, abbreviations, and acronyms.

That said, the needs of translators will shape the dictionary into a specific microstructure that cannot be taken for granted. From this perspective, providing monolingual, bilingual, and multilingual definitions is desirable, and it may be tackled with Definition Modelling.

2.1.2 Specialised Languages and Dictionaries

A Dictionary for Translation can cover the general language [1] or a specialised language [2]. The scientific and technical progress created the need for specific words to convey novel concepts and objects. Besides that, the intercultural exchange promoted by the Internet also forced languages to develop new words to coin such concepts created in other languages and countries.

These languages can be called LSP, sublanguages, special languages, or specialised languages. As Pearson (1998) points out, sublanguage seems to be preferred by researchers concerned with the computational processing of natural language. Sublanguage was first introduced by Harris in 1968 and further developed by Hirschman & Sager (1982) as a language of an area or specific domain (Pearson 1998, Kittredge 2003).

It is Cabré (1999) who introduces a broader concept for specialised languages under the scope of Communicative Theory of Terminology (CTT). For Castellví (1995), terminology has three different meanings: the discipline, the practice, and the product of the practice. The main object of terminology is the specialised unit – the term, which occurs in a specialised language.

For the author, specialised languages are contained in a general, natural language (ibid., p. 58); opposing to Harris' view of sublanguage, specialised languages are not homogeneous or a "monolithic set of grammar rules" (ibid.). In this perspective, three main features describe a specialised language: "subject field, type of user, and type of situation in which the communication takes place" (ibid., p. 65).

Under Cabré's view, specialised languages "are object of a specific learning process" (ibid.) They are a subset of natural language, constantly overlapping with general language and other specialised languages (ibid.). Communication among the different types of users in the specialised domain often follows scientific and professional guidelines, characterised by a formal register (ibid.). Finally, specialised languages are characterised by language-based and text-based features (ibid.).

2.1.3 Terminography or Lexicography

So far, we have discussed lexicography, terminology, and specialised communication. Some scholars might draw a straight line between the two fields and practices: terminography and lexicography may have been born with different objects of study, but technological advancements and the growing interdisciplinarity nature of the two pose an increasingly difficult challenge to distinguish one from the other. From a historical perspective, they have usually been described as the opposite.

Traditionally, terminography is concerned with compiling glossaries and termbases in the scope of a specialised language. On the other hand, lexicography seeks the production of dictionaries of general language. In theory, they do seem to grow apart; however, the daily practice suggests that they constantly overlap (Bowker 2017). The main reason for that is that the two areas aim at, ultimately, creating a "product that is useful for their target audiences" (ibid.).

Terminography emerges with Wüster (1898-1977) and his **General Theory of Terminology** (GTT), where he seeks to eliminate ambiguity in specialised language by standardising

[1] Frequently called Language for General Purposes (LGP).

[2] Or Language for Special Purposes (LSP).

terminology (Castells 2003). His main goal is to facilitate communication in specialised languages, mainly in hard sciences. GTT is a prescriptive approach, which does not consider polysemy and language change (ibid.).

Wüster's theory follows an onomasiological methodology, seeking to define a specialised unit from the concept to the term (Bowker 2017). He also proposes that experts should formulate definitions since terms should not be influenced by language change as natural language (Maciel 2001).

Later, Cabré challenges Wüster's approach to terminology by proposing the **Communicative Theory of Terminology** (CTT). As previously discussed, for this theory, terminology's object of study is the term, which occurs in a specialised domain to address specialised communication. CTT is considered a social approach to terminology (Maciel 2001), and one of its main contributions is to recognise polysemy in the scope of a specialised language (Oliveira Junior 2012).

On the other hand, lexicography is an ancient practice, appearing in 3200 BC, as seen in Durkin's comprehensive chronology of lexicography (2016). As it is widely known nowadays, modern lexicography is inaugurated by Samuel Johnson in 1747 with the Plan of a Dictionary of the English Language (ibid.).

Traditionally, lexicography is different from terminography in some respects. First, while terminography usually focuses on terms, lexicography concentrates on general language words (Bowker 2017). Second, terminography follows an onomasiological approach; it is mainly prescriptive (ibid.). Alternatively, lexicography adopts a semasiological approach (from word to meaning), mainly descriptive (ibid.). Finally, terminographic products are usually intended for experts and highly skilled professionals, while lexicographical products are made for non-experts and laypeople (ibid.).

Nevertheless, there are a number of factors influencing the recent overlapping between lexicography and terminography. First, the widespread tendency of creating crowdsourcing products included non-experts in the scope of both terminography and lexicography (Bowker 2017); however, products were only created by the terminographer or lexicographer, but currently, the audience can take part in the projects, share their sources and materials, blurring traditional boundaries between the two practices. Additionally, corpus-based methodologies made terminography adopt a more descriptive and semasiological approach to terms (ibid.).

If the focus is placed on user needs in the same fashion as the FTL does, there is no need to draw a straight line between lexicography and terminography. The main reason for that is that what matters, in the end, is how users, translators, will benefit from the product conceived.

As Bergenholtz et al. (2010) advise, “translation has a particular relevance for specialised dictionaries”, regardless of resulting from terminography or lexicography. In the end, "it is not important if the cat is white or black – it must be able to catch the mouse" (ibid., p. 36).

Above all, it does not matter to the final user the distinction between terminography and lexicography. In fact, translators are interested in both practices and products, especially in their overlap. In actual, daily produced natural language, terms, and general language are in constant contact, and translators need to have access to this information to produce more accurate and natural texts.

2.2 An Overview of Definitions

Definitions are crucial in the Information Society; they spread knowledge throughout different areas. However, whether compiling a dictionary or a glossary, formulating a definition is not a trivial task. Several scholars have debated the conceptualisation and characterisation of definitions from multiple points of view. By and large, a definition can be described as a set of information given about a word; an explanation, making it possible to understand the meaning behind a concept (Atkins & Rundell 2008, Hanks 2015).

First and foremost, it is essential to address the meaning of 'definition'. Atkins & Rundell (2008) argue that "'definition' is a misnomer" with a view to the misunderstanding that meanings can be isolated and described alone. Instead, many scholars in the area prefer 'explanation' (Atkins & Rundell 2008, Hanks 2013, 2015).

If definitions can be considered explanations, then what is the difference between glosses and definitions? While there is no formal distinction between the two, the term gloss usually designates a brief explanation or note of a foreign word or concept in a very restricted environment[3]. Definitions are meant to be complete explanations of the linguistic unit being described. In this thesis, we will call **definitions** the product of a complete, (human-made) lexicographic/terminographic process, whereas **glosses** will be reserved for brief explanations, including those generated by the machines, for simplification purposes.

Turning back to the nature of definitions, debates over definitions make a distinction between ***definiendum*** (the item that is to be defined) and ***definiens*** (the general explanation used to define the item) (Hanks 2008, 2015, Polguère 2003).

The history of definitions dates back to Aristotle, who provided guidelines on "stipulating the meaning of a concept" for scientific purposes (Hanks 2015). He propounds that objects and concepts (excluding people and people-related characteristics) can be defined in terms of a *genus* and *differentiae*; that is, by answering two questions: (i) what kind of thing it is? to identify the genus, and (ii) how this thing is different from the other members of the same group (the genus)? – to identify the differentiae (Hanks 2015).

Aristotle's *genus and differentiae* approach succeeded in systematising the knowledge for scientific concepts, but it was not effective when applied to abstract items. Later, Leibniz conceived the universal language of characterisation (Hanks 2015). He intended to create a language of "precisely defined concepts" to describe the scientific world (ibid.). He followed a top-down approach, focusing on the concepts – the terminology, instead of the general language.

Leibniz's work is clearly based on Aristotle's *genus* and *differentiae*, except for the employment of the substitutability approach for definition, which states: "two things are the same if one can be substituted for the other without affecting the truth" (Hanks 2015). Such a method is broadly adopted in many modern dictionaries nowadays.

The history of definitions shifts again in 1755 with Johnson's English Language dictionary, the first effort toward covering the entire standard general language (Pearson 1998, Hanks 2015). He states in the preface of his dictionary that defining is difficult because using synonyms and paraphrases does not work in all cases.

In 1987, Sinclair's COBUILD brought innovative contributions to lexicography. As the first corpus-based dictionary, COBUILD's lexicographers explain words in context using corpus

[3] See the difference made in the Lexicog Group and in Wiktionary

data and complete sentences (Pearson 1998, Hanks 2015).

COBUILD's pioneering approach raised a question: can authentic sentences (where definitions are displayed) be used raw as lexicographic definitions? The answer is no. Barros (2004) points out that "a definition should not be shared with two dictionaries, simply because each dictionary has its own specificities, which will determine the content and the organisation of the lexicographic definition". Nevertheless, these authentic sentences are valuable for lexicography as definitional contexts.

These different approaches to definitions resulted in taxonomies created to organise, understand, and collect definitions more efficiently. Before examining these taxonomies, it is crucial to distinguish between **definitional contexts (DC)**, and **definition markers (DM), lexicographic definitions (LD), terminological definitions (TD)**. Generally, LDs are definitions formulated for dictionaries; TDs are definitions created for and by specialists in the scope of specialised language; DCs are encountered in authentically produced texts, and DMs are linguistic or graphic indications suggesting the occurrence of a definition.

Overall, definitions are conveyed in the text through definitional contexts introduced by definition markers. In order to build lexicographic definitions, the (ideal) approach would be to retrieve definitional contexts from corpora and use a strategy to formulate the definitions. The following section will analyse how definitions are classified throughout the literature, and then we will explore how lexicographic definitions are created.

2.2.1 The Anatomy of Definitions

Due to the long-standing controversy over the nature of definitions, several scholars have proposed different taxonomies to classify definitions; Appendix A that definitions can be grouped in terms of formality (Trimble 1985), purpose (Robinson 1972), lexico-grammatical features (Westerhout & Monachesi 2007, Westerhout 2010), and semantics-based features (Sierra et al. 2006, Alacón Martínez et al. 2009).

2.2.2 Formulating Definitions

Different approaches explain the meaning of a word in a dictionary, depending on the user and the dictionary type. There are four methods to explain meaning: illustration, exemplification, discussion, and defining (Pearson 1998).

One practical method used for explaining meaning is **illustration**. Illustrations (pictures, figures, or tables) are commonly found in children's dictionaries, visual dictionaries, and specialised, technical dictionaries (Pearson 1998). Even though they cannot fit in every dictionary, illustrating is very useful for foreign-language users, as they can quickly associate the meaning with their L1; as commonly said, "a picture speaks more than a thousand words".

Another method used for lexicographic definitions is **exemplification**. Pearson (ibid.) shows that this method has a bifold purpose: either demonstrate meaning or usage. For the former purpose, examples of the word are cited, and in some cases, they may substitute the definition. As for the latter purpose, syntactic, selectional restrictions, collocations and stylistic levels are cited. The Oxford English Dictionary tends to employ examples to clarify both meaning and usage, while COBUILD resorts only to the latter (ibid.).

A third method for describing meaning is **discussion**. Discussion is mainly used for explaining function words, such as prepositions. Pearson (1998) does not analyse this method deeply, but this method seems to be related to the rule-giving definition type presented in Section 2.3.3.

Finally, the last method used to describe meaning is the **definition**. The most common and widely adopted method is the already mentioned Aristotelic pattern, i.e., "in terms of a lexical item's superordinate and a distinctive characteristic which distinguishes the lexical item from other members of the same group" (ibid.). There are two different explanatory strategies for defining meaning (Hanks 1987, 2008, 2015): the substitutable defining strategy and the COBUILD strategy.

As mentioned in Section 2.3, the substitutable defining strategy aims at creating definitions that "could be substituted in any context for the word being defined" (Hanks 1987). Surprisingly, this notion has been widely accepted and used by many lexicographers (Pearson 1998), even though it is known that perfect equivalence rarely exists. Hanks (1987) and Pearson (1998) emphasise that the persisting attempt to use this method, even when it is clear that it is not as helpful as it seems, has led to phrasing, which has continually led to distortions and "awkwardnesses".

The COBUILD defining strategy, as mentioned in Section 2.3, proposes definitions in regular prose. Explanations are constituted of two parts. The first part abandons the tradition and places the word being described (in bold) in a specific structure to show use; the second part follows to some degree the traditional approach to definitions to explain meaning (Hanks 1987, Pearson 1998, Hanks 2008, 2015).

In Pearson (1998) and Hanks (2008, 2015) view, the COBUILD approach is much superior as a defining method since it covers both meaning and usage. Pearson (ibid.) advocates using this strategy in specialised dictionaries, especially those in which usage examples are not provided.

In the previous sections, we discussed the traditional prescriptive nature of terminology. Béjoint (1997) draws attention to the fact that terminographic definitions are usually taken for granted without considering their content and structure. Faber (2002), however, affirms that terminological definitions are representations of knowledge. Therefore, there should be guidelines for building efficient definitions instead of simply copying and pasting from dictionaries or other sources.

Terminological definitions can be classified in terms of language (monolingual, bilingual, multilingual) or object (terms or concepts) (García de Quesada et al. 2001). Historically, terminological definitions have focused on defining concepts, which are units of knowledge represented and identified in reference to an object from interior and exterior world (Sager 1990, Cabré 1999). For García de Quesada et al. (2001), concepts are a "possible formalisation for a component of specialised knowledge"; they are a starting point for the term, which are defined based on one or more concepts used in the scope of a determined specialised language.

For Wright & Budin (1997), both conceptual systems and terminographic definitions can be classified into intensional and extensional definitions. Intensional definitions enumerate the set of characteristics that constitute a concept; these are the most common type of definitions in terminography (ibid.). Extensional definitions specify all objects to which the concept refers (ibid.). García de Quesada et al. (2001) emphasises that extensional definitions can only be formulated when the terminographer follows a well-structured and delimited taxonomy.

Broadly, the ISO 1087 proposes recommendations for formulating terminological definitions (Pearson 1998, García de Quesada et al. 2001). These recommendations are mainly related to the selection of superordinate (or hypernym) that will substitute the term in the definition (Pearson 1998), as terms should not be repeated in the definition (García de Quesada et al. 2001). García de Quesada advocates for improved and transparent guidelines for building terminological definitions, as the ISO's recommendations are not particularly useful and do not provide a basis for a methodology.

2.3 Machine Learning and Deep Learning

Machine Learning (ML) emerged from the AI field; from a broad perspective, it is concerned with teaching a machine to perform a task automatically by training it with real-world data (Müller & Guido 2017). Machine Learning employs supervised and unsupervised learning techniques.

Supervised learning techniques are used to predict an outcome from a given input (ibid.). In this task, the machine learns from a training set, usually prepared by humans, to teach the machine to perform the identification (ibid.). One example of supervised learning is the detection of fraud in credit card operations. The system learns from the habits of customers (time, place, amount) and then classifies the transaction as suspicious or not.

In unsupervised learning, the output is unknown since there is no training set. In this task, there is no training phase. Therefore, the machine is to extract knowledge from the given data (ibid.). An example of unsupervised learning is classifying blog posts by topics. The system does not know in advance how many topics or which topics are contained in the dataset; instead, the machine learns by itself to recognise the patterns and group the topics showing similarity.

2.3.1 Key Concepts of Deep Learning

Deep Learning (DL) is a subset of Machine Learning. The goal of DL is to learn by mimicking the human brain; that is, the machine processes data through nodes, called **neurons**, organised into different **layers**. The vertical stacking of multiple layers is called **neural networks** (NNs). Each layer can be responsible for tackling one activity, and this is why it is called **deep learning**. Figure 2.1 represents a typical deep learning architecture.

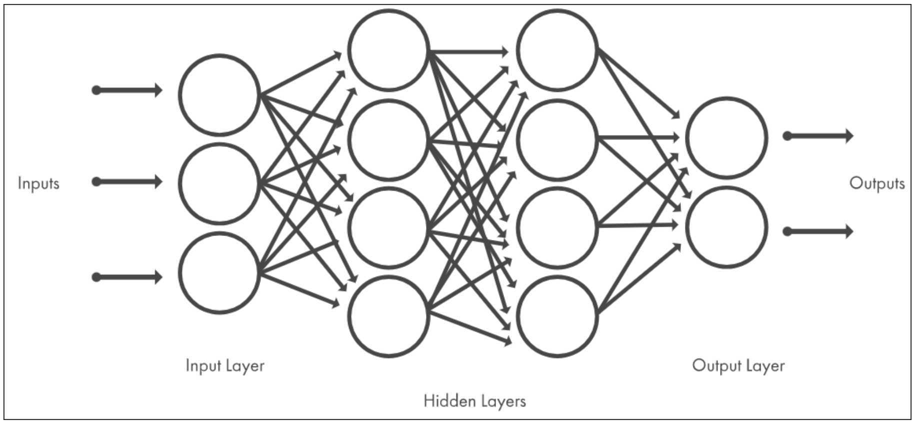


Figure 2.1: *A deep learning architecture*
Source Mathworks

By and large, to build a network with deep learning, the researcher should provide the system with a large, labelled dataset from which the machine will learn — for example, a dataset containing distinct parts of cats from different breeds. After the model is trained, a large unlabelled dataset is provided to the model on the **input** side. Each **layer** is then responsible for identifying one feature, and the **output** layer combines the results to present it later.

In the cat example, the input is a dataset containing several photos of different animals. The system's goal is to identify whether there is a cat in the image. Since the model is already trained, each layer is now responsible for predicting one feature of the cat: eyes, ears, legs, paws, fur texture, and tail. In the end, the output layer combines all the features and delivers only the images with a higher probability of containing an actual cat.

Each layer will pass the information on to the next layer from left to right, resulting in the so-called **Feed-Forward Network (FFN)**. Layers are constituted of **nodes**: neurons with one or more weighted input connections. In practice, when an input goes through a node, the input is multiplied by a weight, and the output of this operation is passed onto the following node in the next layer.

The next layer will process the input from the previous layer using an **activation function**: a mathematical function that will determine whether a neuron will be activated or not to calculate the output.

The main goal of an NN is to learn to map the set of inputs to the set of outputs from the training data. It is not possible to calculate the perfect value for the weights in advance, so the learning process turns into an optimisation problem, where the NN navigates the training set to learn to make predictions by trial and error. To know whether the model is making good predictions is crucial; this is done by employing a **loss function**, whose main purpose is to calculate the distance between the input and the output.

Once the model has the loss, the optimisation function will tell the model in which direction to go with a view to minimise the loss function and reduce the error. The **learning rate** will then, control how often the model should update the weights to find the lowest loss possible. This number is relatively small since big steps can mean missing the target and increasing the loss.

Finally, the model needs to learn from the experience and update the weights accordingly. This is done through **backpropagation**. At first, we assign random weight values, but through iterations, **epochs**, the model updates the loss from the previous iteration and retraces the search for the lowest loss value.

2.3.2 Different Types of Layers and Architectures

Deep learning models and their components are highly flexible. As a result, different layers and architectures have been devised, for example, Multi-Layer Perceptron, Recurrent Neural Networks, Long-Short Term Memory (Hochreiter & Schmidhuber 1997), Gated-Recurrent Units (Chung et al. 2014), and many more.

Different layers can be combined to form individual or hybrid architectures to tackle different problems. For example, one of the most essential frameworks in NLP is the **sequence-to-sequence** (seq2seq) framework, comprising any task that takes a sequence of characters as input and returns generated text. One of the most potent deep-learning models, the **encoder-decoder model**, was designed especially for seq2seq tasks (Sutskever et al. 2014).

The encoder-decoder model resonates with how humans communicate and the Saussurean notion of signified/signifier. For example, when we explain what a tree looks like, we create a mental representation of the tree (a pine tree), the signified; that is what the encoder does. The decoding process occurs when we talk about trees: we decode the mental representation by explaining it in different formats (the signifier), visually, a drawn of the tree, or verbally, the uttered word.

Formally, the encoder processes the input sequentially, representing the whole sequence as a **context vector**, which will be passed onto the decoder. The decoder, in its turn, will produce the output autoregressively, word by word, conditioned to the context vector.

Let us zoom in on each component. What does the encoder consist of? It is constituted of different layers. The first layer is in charge of encoding the input (a sentence) into a vector. Traditional seq2seq models have **Recurrent Neural Networks (RNNs)** (Cho et al. 2014) in their core.

RNNs read one sequence after the other. They have three main components: input, output, and recurrent. They process the input into steps called **time steps**. The recurrent connections are the essence of an RNN; they ensure the flow of information in-between time steps. The input comes in, it is processed with an activation function, and the output, called **hidden state**, comes out. The hidden state becomes an input. Then, it is added to the representation of the next word, and the process goes on.

In the end, the NN will produce the context vector, a compact representation of the entire input passed on to the decoder. However, compacting all information into one final hidden state can be problematic because it creates a bottleneck[4]. Moreover, larger datasets and longer inputs are also unwieldy since each hidden state depends on the output of the previous hidden state, resulting in the vanishing gradient problem [5]. Therefore, other RNN-based architectures have been proposed to tackle these challenges.

Long-Short Term Memory layers (LSTMs) (Hochreiter & Schmidhuber 1997) solve this problem by employing special memory cells, cell states, that hold information in memory for longer time steps. These cells have a set of gates (other layers) responsible for controlling the information being added or removed from the memory. The supplementary gates enable LSTMs to deal with longer sentences. Another variety of LSTMs is the **Gated Recurrent Units (GRUs)** (Cho et al. 2014). They are a simplified version of LSTMs with a reset gate and an update gate.

Even though LSTMs fare well with textual data, they cannot handle complex sequences well. Nevertheless, an innovative idea paved the way for state-of-the-art (SOTA) models in the field. If we enable the components to pay attention to the most important parts of a sentence, it can alleviate the bottleneck problem. This is known as the **attention mechanism** (Vaswani et al. 2017). It enables all hidden states to be passed on to the decoder instead of forwarding them only the final state proposed in traditional seq2seq models.

The attention mechanism describes the relation between two states and provides a soft

[4] It occurs when you have much information from one side being compacted on another side; it makes you forget things or lose details.

[5] Multiplying smaller vector numbers leads to even smaller numbers. The more we approximate to zero, the higher the chance of having zero as a derivative for the loss, which means that if you cannot move, the network stops learning.

alignment between them through an attention score. The most prominent way to calculate the scores is employing the dot product. What if, instead of looking only for an input-output sequence association, we can look for the association of words within a sentence? That is the **self-attention mechanism**. This mechanism is the building block of an even more impressive architecture.

Google introduced the Transformer architecture with the seminal paper **Attention Is All You Need** (Vaswani et al. 2017). It is also an encoder-decoder based architecture, but it substitutes RNNs for Attention layers and FFNNs. This architecture enables the encoder and decoder to simultaneously see the entire input sequence using self-attention.

Let us examine the transformers' architecture in detail. In contrast to RNNs and LSTMs that process the input sequentially, Transformers architectures process inputs as a set, losing the notion of word order. **Positional encoding** is applied before the first self-attention layer to solve this issue. Thus, if the same word appears in a different position, the actual representation will be slightly different depending on where it appears in the input sentence.

Another important aspect is the key-value-query concepts. For example, when you search **(query)** for a word in Google, the search engine will map your query against a set of **keys** (website title, description, and more) associated with possible stored websites. Then the algorithm will present to you the best-matched websites **(values)**. This is the foundation of content/feature-based lookup adopted in Transformers.

Instead of choosing where to look according to the position, we can now look to the content we want to look at so that inputs can be split into key-value pairs. The keys define the attention weights (how important something is), and the values are the information itself. We calculate the scaled dot-product to obtain the attention weights to describe how two words correlate. To get the final value as a probability distribution, the SoftMax function is applied.

The main idea underpinning Transformers is that we can pay attention to multiple things simultaneously. This is called the **multi-head attention mechanism**. The intuition behind multi-head attention is that it allows us to attend to different sequence parts differently each time.

Finally, as Transformers do not have RNN-based layers, the bottleneck problem is alleviated and dealing with long term sequences becomes feasible.

2.3.3 Low-Resource Languages and DL Paradigms

While neural networks are certainly powerful, they are also data-hungry. Because they learn through examples, especially for NLP, enormous amounts of high-quality data are required to represent the richness of human languages.

However, most languages do not have sufficient resources to build NLP applications. One of the main reasons for that is related to the presence of languages on the web (Pimienta et al. 2009); it is much easier to obtain data already available over the Internet through crawling. Another reason for that has to do with political and economic reasons: the more economic incentive available, the more research and resources are made available, which is indisputably not the case for all languages.

These languages are called **Low-resource languages (LRLs)**. LRLs is an umbrella term that encompasses less studied, resource-scarce, less computerised, and/or less privileged languages (Singh 2008, Cieri et al. 2016, Tsvetkov 2017, Magueresse et al. 2020). In this thesis, the term will be used in the case of resource-scarce languages; more precisely, for Portuguese in relation to Definition Modelling.

Building fundamental NLP resources for LRLs impacts many human activities and brings positive change. Undeniably, language technologies have commercial value and strengthen the economy. Moreover, these resources can foster language preservation, knowledge expansion, monitoring of demographic and political processes, and emergency response applications.

In order to mitigate the lack of resources for LRLs, distinct DL learning techniques were devised for different scenarios. In **Few-shot Learning (FSL)**, models learn from only a restricted number of examples in the training set (Wang et al. 2020). When there is only one example of a phenomenon available in the training set, FSL becomes **one-shot learning** (ibid.). Likewise, when there are zero examples of a phenomenon in the training set, FSL is called **zero-shot learning** (ibid.).

FSL is typically combined with **transfer-learning**: transferring knowledge from a rich-resource domain/task to a low-resource domain/task (ibid.). It is also possible to employ **cross-lingual learning**, which transfers knowledge from a rich-resource language to a low-resource one (Adams et al. 2017). Finally, **multilingual learning** enables the transference of learning in leveraging rich-resource languages (ibid.).

Although not perfect, these techniques have yielded promising results across different tasks, for example, Automatic Question Generation (Kumar et al. 2019), Multilingual Quality Estimation (Sun et al. 2020, Ranasinghe et al. 2021*a*), Cross-lingual Quality Estimation (Ranasinghe et al. 2021*b*), Multilingual Offensive Language Identification (Ranasinghe & Zampieri 2020), and much more.

2.4 Definition Modelling

Definition modelling (DM) is a reasonably recent language modelling task introduced by Noraset et al. (2017) to generate a definition of a word and its embedding. The task's main goal was to evaluate word embeddings. DM has proven to be a quite challenging task for various reasons.

To start with, there is a critical issue with the notion of **sense**. Roughly, word senses are meanings conveyed in a given context. Now more than ever, in such an interconnected, multilingual world, not only do people interpret meanings differently, but it can be difficult even for linguists and lexicographers to agree on the senses of a word (Hanks 2000, Kilgarriff 2007).

Furthermore, delimitating sense boundaries deems to be an unattainable task (Kilgarriff 2007), casting doubt on the very existence of word senses (Kilgarriff 1997, Hanks 2000). Undoubtedly, the main reason for this dispute comes from the creative and productive features of natural languages, allowing for **polysemy**.

Apart from this, computers lack the so-called real-world (also called encyclopaedic or general) knowledge that is very difficult to convey with a small amount of data.

Lastly, another issue is representing words in computational form. Word embeddings are a fixed-length vector representation of words. Here, the main challenge comes from representing words as infinite real numbers with a finite number of bit patterns. Early approaches adopted frequency-based embeddings, as in the window-based co-occurrence matrix (Haralick 1979) and singular value decomposition matrix (Rohde et al. 2006), but these approaches become problematic when documents grow bigger. In addition, it is computationally expensive and difficult to include new words.

Then, it would be possible to represent words in a binary scheme, as accomplished in one-hot encoding[6]. However, this approach generates orthogonal vectors, consequently missing the notion of similarity and context. In order to obtain similar words, one could draw information from Wordnet, but this approach has proven to be computationally expensive. Besides that, Wordnet has many weaknesses itself, such as the absence of precision (nuance) between senses, the lack of new words, and the requirement of human labour to update the whole structure, to name a few.

The idea, however, of disambiguating meaning through context is not new in lexical semantics. The most notable contribution on this matter is drawn from Firth (1957):"you shall know a word by the company it keeps" His contributions resonated through many areas of language, but most importantly, in computational linguistics.

Although not the first to suggest dense-vector representations, Mikolov, Chen, Corrado & Dean (2013) made history by proposing word2vec. This model learns word vectors by translating a word into a one-hot encoding vector and mapping it to the surrounding words (the co-text) using neural networks. Word2vec produced insightful results indeed and inspired many other approaches to follow: Skip-gram vectors (Mikolov, Sutskever, Chen, Corrado & Dean 2013), CBOW (Pennington et al. 2014), Char2vec (Cao & Rei 2016), FastText (Bojanowski et al. 2017*a*), and many more.

Even though word2vec generates static word embeddings, these vectors still capture relational meaning and similarities to some extent. In contrast, new advances in deep learning architectures, such as the attention mechanism and transformers, paved the way for contextualised word representations that are much more powerful than static ones.

Contextualised word representations enabled easier disambiguation between homographs by still changing vector representations after training. The disambiguation of "the muddy river *bank*" and "the *bank* investors" is the most classic case in point.

Finally, transforming words into numbers facilitated a myriad of vector operations that make machine learning algorithms so powerful. The cornerstone of these operations is **similarity measures (SMs)**. As it was covered, usually in ML problems, we want to classify, group, and generate data. Therefore, SMs are employed to calculate the distance between vectors in the representation space.

There are, in fact, different types of measures: Euclidean distance, Manhattan distance, Minkowski distance and Jaccard similarity. For this work, the most important one is the **cosine similarity** - the normalised dot product of two vectors. Cosine similarity has been shown to work very well with comparable corpora and NLP tasks.

2.4.1 Related Work

Definition Modelling emerged as a deep learning problem due to the challenging nature of the task. Most DP approaches have relied on supervised conditional generation (Noraset et al. 2017), while some approaches have introduced different conditioning features, such as context (Gadetsky et al. 2018, Ishiwatari et al. 2019), sememes (Yang et al. 2019), or semantic decomposition (Li et al. 2020).

[6] Assigning 0 when a word does not occur, and 1 when it occurs in the sentence

Noraset et al. (2017) present an RNN-based model with an update function inspired by GRU gates to tackle word-to-sequence definition in their seminal work. They make available the first English dataset for the task, composed of definitions extracted from the Oxford Dictionary. The main weakness of their system is the use of static word embeddings, which disables the generation of polysemic definitions. Besides that, they do not address multi-word units.

Later, it became clear that the pioneer system could not learn polyseme semantics having a single word as an input. To tackle this problem, Gadetsky et al. (2018) put forth two models that include contextual information for the first time. One of the models is an adaptive skip-gram model that generates different context-based word embeddings. The second model employs an attention-based mechanism to extract only relevant information. Results suggest that the use of context yields good results, at least in the automatic metrics.

Yang et al. (2019) propose to incorporate sememes to address Chinese DM. They employ an encoder-decoder framework with both adaptive and self-adaptive attention models (popularly known as transformers) so that the model can decide when and to which sememes to pay attention. In addition, they conduct an ablation study to verify which components impact the quality of results. Adding a module for incorporating sememes proved to have the highest impact on the quality of the generated definitions.

Returning to adopting context, Ishiwatari et al. (2019) use local context (co-text) and global context (external information) to generate unknown definitions. To minimise ambiguity, the researchers employ an LSTM-based encoder-decoder model with a gated unit. In addition, they conduct manual and automatic evaluations, error annotation and BLEU, respectively. While findings indicate that the adoption of local context helps in both disambiguating polysemous words and inferring the meaning of unknown expressions, it is shown that global context only helps when the local context is extremely short (for example, in Rihanna is a singer). Finally, the study confirms that the generation task becomes harder and harder the more ambiguous and polysemic the words are.

Chang & Chen (2019) reframe definition modelling as a definition ranking (classification) problem. Their model is to select the best-fitting definition given a lemma and its context. Their approach avoids the problem of generating disfluent sequences, the exposure bias problem (generating sequences without a target in test time) and evaluation problems. To carry out the task, they leverage pre-trained models - ELMo and BERT. Instead of having the classifier map discrete labels, the model learns a mapping function to encode the definitions in the embedding space. After that, an embedding is generated given a context. The last stage applies the k-nearest neighbour (kNN) algorithm to predict multiple definitions for a given word by leveraging the embeddings. Their experiment shows that contextualised word embeddings do represent sense-informative cues.

By contrast, Mickus et al. (2019) recast DM as a sequence-to-sequence task rather than a word-to-sequence task; that is, context should be given as an input instead of the lemma. This is highly beneficial because it enables the incorporation of multi-word expressions in future tasks. Furthermore, they include a marker embedding to highlight the *definienda* in context, allowing them to achieve excellent results in contextual and non-contextual definition modelling.

So far, definition modelling has been addressed as a context-dependent task. Kabiri & Cook (2020), on the other hand, introduce a context-agnostic approach to DM by exploiting multi-sense word embeddings. The authors devise the first multilingual study; nine languages are included. They train their model on sense-definitions pairs, where they represent words as an average of all words and associate with the sense by measuring pair-wise cosine similarity. Their approach can generate definitions for polysemic words across nine languages.

Alternatively, Li et al. (2020) propose decomposing word meanings into semantic components with explicit semantic decomposition. This information is then modelled with the discrete latent variable in an encoder-decoder bidirectional-LSTM model. Reported results show that the method leads to both qualitative and quantitative improvements, culminating in more specific and accurate definitions.

Bevilacqua et al. (2020) leverage the power of a pre-trained transformer-based model, BART, in a sequence-to-sequence model. Generationary tackles not only definition modelling but also Word-Sense Disambiguation and Word-in-Context tasks. Moreover, the model is the first one to handle multi-word expressions.

At the end of 2021, Mickus et al. (2022) proposed the Comparing Dictionaries and Word Embeddings (CODWOE) shared task. Their main goal is to generate glosses from vectors (in the definition modelling track) and reconstruct embeddings given a gloss (in the reverse dictionary track). They make available the first multilingual dataset for the task in English, French, Spanish, Italian, German, and Russian to achieve these goals. The authors encourage participants to investigate whether multilingual and cross-lingual learning can improve results.

2.4.2 Evaluation

Over the past decade, the scientific community could witness exponential growth and innovation in NLP tasks. It is still unknown the full impact of AI technologies on human activities, and now more than ever, evaluation plays a key role.

Evaluating results can be difficult due to the open-ended nature of natural language generation tasks and the multiplicity of gold standards. Thus, it is important to report different criteria to understand the system's usefulness.

Broadly, the literature classifies evaluation into two types: intrinsic versus extrinsic (Passonneau & Mani 2014). While the former focuses on the system's performance in achieving its goals, the second concentrates on assessing the system's usefulness in an operational context (in another task, or the real world, for example).

Similarly to machine translation, DM evaluation oftentimes takes a definition reference (a gold standard) created by humans and compares them to the automatically generated definitions. Papers report a variety of automatic metrics.

Machine learning studies usually report traditional metrics: **precision** and **recall**. The first one calculates the proportion of correctly identified words in the output sentence. The last one refers to the sensitivity of the system: it counts the number of correctly identified words in the pool of references available. DM studies occasionally report precision, while recall is used as a basis for other automatic metrics.

Despite not capturing semantic features, the **Bilingual Evaluation Understudy (BLEU)** (Papineni et al. 2002) has been widely reported in DM studies. BLEU scores are precision-based and calculate the n-gram overlap between output and reference. BLEU scores were reported in all studies conducted in DM so far. The higher the BLEU scores, the better.

Another important metric is **Perplexity** (Jurafsky & Martin 2021). It measures how difficult it is for the language model to generate a sentence. It measures the amount of entropy in a probability distribution and normalises it by sentence length. The lower the perplexity, the better.

Following BLEU in n-gram overlapping, the **Recall-orientated Understudy for Gisting Evaluation** (ROUGE) (Lin 2004) also measures the overlap of n-grams. A variation of the metrics, ROUGE-L, takes the longest sequence of tokens shared between reference and output to measure similarity. The higher the ROUGE score, the higher the overlap between input and output.

In contrast to BLEU and ROUGE, which performs string matching, studies also report the **Metric for Evaluation of Translation with Explicit Ordering** (METEOR) (Banerjee & Lavie 2005, Denkowski & Lavie 2011). This metric uses stemming and WordNet synonyms to do unigram matching between the reference and the output by employing a harmonic mean of precision and recall.

BERTScore **(Bidirectional Encoder representations from Transformers)** (Zhang et al. 2019) computes the similarity score for each token in both reference and output using contextual embeddings. Since it is an embedding-based metric, BERTScore seems to capture semantic and similarity features, crucial for definition modelling.

Finally, Zhao et al. (2019) propose the **MoverScore**, a monolingual measure designed to evaluate the similarity between two sentences in the same language. It measures the semantic distance between input and reference in order to correlate better with human judgement. The lower, the better.

When the model produces fluent and meaningful definitions, qualitative experiments are devised. Usually, researchers employ a Likert Scale to evaluate whether the generated definition fits the target context (which is given to the model in order to generate definitions). Moreover, error analysis is also generally employed, given the nature of the task.

Despite the employment of automatic metrics to conduct intrinsic evaluation, so far, no study has focused on extrinsic evaluation or the usefulness of the generated definitions in a real-world context.

While post-editing has become a fairly widespread practice in translation, in lexicography, the term refers to a much broader collection of entry components[7] (lemma lists, collocations, colligations, usage examples, relations between lexemes, and much more). Only in the latest few years attention has been drawn to the automatic generation of definitions, mainly, as already highlighted, due to the complexity of the task at hand.

2.5 Conclusions

Throughout this section, key concepts to DM were presented. First, an overview of the field of lexicography was provided, with a special focus on the **Function Theory of Lexicography**, which places the user in the centre of the lexicographical project. Although lexicography has been thriving due to technological advances, dictionaries frequently fail to deliver necessary information to translators. This is partly due to the lack of research on translator dictionary use and also to the labour-intensive nature of lexicographic projects. Therefore, automating the generation of definitions can facilitate the work of lexicographers and pave the way for producing news resources for translators.

[7] More information available on the Proceedings on the eLex 2021 Conference, available in: `https://elex.link/elex2021/wp-content/uploads/eLex_2021-proceedings.pdf`

Section 2.2 showed that formalising the concept of definition is complex, and several taxonomies have been proposed in an effort to systematise the variety of naturally occurring definition types. This thesis adopts a more flexible taxonomy since the goal is to exploit what is already available and not classify it. Lexicographic definitions will be used then to refer to definitions taken from reference resources and glosses will be used to designate automatically generated definitions, for simplification purposes.

Concerning Section 2.3, machine-learning techniques have indeed increased the availability of annotated data. Surprisingly, this is not true for Portuguese, even though it is the 10th most spoken language globally[8]. Hence, in this thesis, since no studies have been conducted in Portuguese, this language will be considered a low-resource language. Moreover, Section 2.4.1 shows that no study has attempted so far to generate definitions attending translators demands, combining monolingual L1 and L2 definitions, or generating bilingual definitions, which can be very useful for translators, as pointed out in the detailed review in Section 2.1.

Furthermore, in light of the deep-learning concepts presented, this thesis seeks to leverage transformer-based models, provided that they deal better with longer sequences of definitions, and they have proven to achieve SOTA results in DM. Besides that, cross-lingual and multilingual learning techniques are employed with a view to exploring whether these techniques can aid generating definitions for Portuguese, a low-resourced language for Definition Modelling.

Finally, in addition to evaluating outputs with standard metrics of the field, human evaluation is conducted with translators in light of the Function Theory of Lexicography, in order to bring translators to the centre of the lexicographic process.

[8] More information available in: `https://www.ethnologue.com/guides/ethnologue200`

Chapter 3

Methodology

Overview – This chapter describes the methodology adopted to automatically generate definitions with two deep learning models. Section 3.1 explains the process of data acquisition for all the languages. Next, the description of the datasets after cleaning and pruning is given. Then, the models adopted to perform DM are explained in 3.3. After that, the experimentation phase is examined, along with the required pre-processing steps for each experiment. Finally, Section 3.5 brings details about evaluation.

3.1 Data Collection

In general, machine learning systems rely on three main components to be trained: data, features, and algorithms (Jordan & Mitchell 2015). Data availability is one of the most significant challenges in Definition Modelling since the task carries out language generation (Adadi 2021). Besides that, DM is quite recent, and there is a shortage of accessible resources for languages other than English (Tsvetkov 2017, Magueresse et al. 2020, Haddow et al. 2021). Moreover, to write definitions, lexicographers hinge heavily on their general, real-world knowledge, which can be one of the most challenging tasks even for the most sophisticated AI systems available worldwide (Fjelland 2020).

Given that deep learning models require large amounts of data to be trained, it would be feasible at this stage to generate definitions for a specific domain. The main reason for that is the scarcity of data for Portuguese (and the other languages). Moreover, in line with Corpus Linguistics principles (McEnery & Hardie 2011), it is also good practice to account for different resources to avoid idiosyncratic usage, which increases the difficulty of obtaining more data. These limitations set the course of this research in the sense that it will not focus on a specific domain rather than on the languages themselves.

In Definition Modelling, dictionary data is preferred. However, given the time constraints and challenges associated with accessing dictionary resources, exploiting available resources is always advisable. Since there are high-quality datasets already available for English and other romance languages, the following section will first focus on the acquisition of data for Portuguese and then proceed to the collection for the other languages.

3.1.1 Collecting Definitions for Portuguese

As seen in Chapter 2, two kinds of resources are central to DM: definitions and contexts. Normally, both can be easily found in dictionaries. Thankfully to recent advances in technology, electronic dictionaries (e-dictionaries) should now be easily accessible, eliminating the need to digitalise printed resources.

There are two main ways of collecting data from online resources: obtaining the data directly from the publisher or scraping the data. The first method is usually preferred since data scraping is laborious and time-consuming. Besides that, it is much easier to process data in a computational-tractable format. Hence, the first step adopted in data acquisition was to survey the available Portuguese e-dictionaries to check whether it is possible to obtain data directly from publishers.

A Landscape of E-Dictionaries in Portuguese

Concerning the Portuguese language, there are at least seven free monolingual e-dictionaries candidates available for online consultation. They are: Michaelis, Houaiss, Aulete, Priberam, Portuguese Oxford, Dicio, and Wikcionário. The primary reason for surveying these resources is that they are free and open, enabling translators to access them easily.

First and foremost, if we place monolingual, general language dictionaries under the lens of the Function Theory of Lexicography, these resources typically do not have translators as main users. Unfortunately, this problem cannot be easily circumvented due to both the lack of investment in the area and the amount of effort needed to undertake the dictionary compilation task.

Furthermore, Portuguese lexicography has its own problems. For example, the most critical issue is the systematic lack of authentic examples for each entry in all the dictionaries[1]. While this problem seems to have been already overtaken by prominent English dictionaries (see Lexico Oxford, the American Heritage, Cambridge, Collins Dictionary, Macmillan, to name a few), Portuguese dictionaries present manufactured, or random examples from old literary sources[2], if any.

Some dictionaries do bring random examples retrieved from newspapers, making the user accountable for connecting senses to examples. Unfortunately, the lack of examples does not really assist their users, whoever these may be, which makes us wonder whether these dictionaries are genuinely based on corpora[3].

Table 3.1 summarises the main features of well-known e-dictionaries available to the Portuguese community. Wikcionário – an encyclopaedia - is not a dictionary per se, but its free crowdsourcing character makes it a valuable resource for research.

[1] It was quite challenging to find examples in the online dictionaries available in Table 3.1. One example is amarelo (yellow) in Aulete, which only presents examples for metonymic use.

[2] That is the case of Dicio, presenting literary examples from translated books.

[3] None of the dictionaries makes officially clear whether they use corpora to create their entries.

Dictionary	N. of Senses	Allows scraping	Has context	Research use
Dicionário Michaelis	350,000	No	For some entries	No
Dicionário Houaiss	376,500	No	For some entries	No
Dicionário Aulete	818,000	No	For some entries	No
Dicionário Priberam	100,000	No	Unordered examples	Yes
Oxford Português	146,000	Unknown	For some entries	Yes
Dicio	400,000	Under request	Unordered examples	Yes
Wikcionário	at least 270,501	Yes	For some entries	Yes

Table 3.1: *Summary of lexicographic resources and their features.*

Selecting which resource to include was challenging because the quality of the data will reflect on the output of the system later. Michaelis, Houaiss, and Aulete dictionaries are maintained by private companies, which does not allow for uses other than personal. Priberam dictionary is a product from a language service provider (LSP) that allows for research but does not enable caching or storing the definitions, precluding the use for this project.

Oxford Dictionaries do enable the distribution of their lexicographic corpora for research. However, they were contacted and never replied. Given that Wikcionário is open and crowdsourced, it was essential to have samples from a lexicographer-made dictionary. Dicio provides its content for research use, given that it is distributed under the same license. They do not provide an application (API) to retrieve its content, so one should be created for this research.

As for Wikcionário, its use in research has been widely known for other languages. This is mainly because there are already-made resources by the scientific community. Surprisingly, the so-called dumps (Wiktionary's database) are not parsed in Portuguese, revealing then the need to scrape the website.

Available APIs for Scraping

Given that Dicio does not provide an API for retrieving its contents, a quick survey on the available APIs for Wikcionário was made in order to optimise time since developing a scraper can be laborious and time-consuming.

Wiktionary parser[4] is a python package that retrieves and parses Wiktionary's contents. Although the API claims to work with languages other than English, in practice, it did not work for Portuguese. Even though there is an option to change the language and the main URL for the language, after some attempts, it was clear that changing the language makes the API retrieve content from the translation fields in the entries. Hence, it does not fit for Portuguese.

Other tools were also tested. Wiktextract[5] is a python package that extracts information from a Wiktionary dump. Some tests were carried out to check whether the tool could work

[4] Available in: https://github.com/Suyash458/WiktionaryParser

[5] Available in: https://github.com/tatuylonen/wiktextract

for the Portuguese dump without extensive adjustments. It turned out that English and Portuguese HTML pages do not share the same structure, limiting the use of the tool.

Nonetheless, exploiting the dumps directly was an alternative that proved to be not fruitful since dumps are provided in SQL, RDF, or Lime Ontology format. Since dealing with these three formats requires advanced knowledge of data structures and other programming languages, other options were explored.

Scraping Dicio and Wikcionário

Provided that previous tools do not work for Portuguese, two different scrapers were used for each website due to distinct structural features in each HTML documentation.

The first step was to inspect their HTML structure. Unexpectedly, this search showed that overall HTML structures were not as machine-readable as it was thought. Whereas Dicio's page follows a more organised structure, Wikcionário's structure is quite chaotic, provided that entries can be edited and formatted by anyone. These inconsistencies in the structure posed multiple obstacles for both retrieval and pre-processing of the content.

Another issue encountered in the pre-scraping phase was that none of the dictionaries clarified how many and which entries they contained, creating the need for wordlists to adjust the URL for each entry. Thus, to obtain wordlists, two approaches were adopted. First, exploiting machine-readable dumps supplied by the Kaikki project[6] project was explored. The project's goal is to facilitate access to Wiktionary's data in a machine-readable manner in JSON files, which opened the possibility of extracting glosses for Portuguese.

A major drawback of Kaikki's project is that entries are also English-centred. Even though the data was said to be in Portuguese, all the information in the entries (definitions, etymology, synonyms, and more) was in English. Hence, only lemmas were extracted from the dumps using python with the JSON library.

The second approach adopted to build wordlists was using Dicio's own wordlists. The dictionary makes available a list of 1-thousand words for each letter of the alphabet. These wordlists were manually gathered and stored in plain text files.

Once the wordlists were collected, they were concatenated so that we could have a better shot at collecting as many definitions as possible. The links for the repository containing all scripts developed for this thesis are available in Appendix A.

The next step consisted of developing the scrapers themselves. The idea was to first create a URL by concatenating the root address of the website to each word from the wordlist. Then, the program verified whether the page for that word existed on the website by sending a head request to the server. Should the URL not exist or be moved, the address would be discarded. To manage this process, the urllib library was used with python.

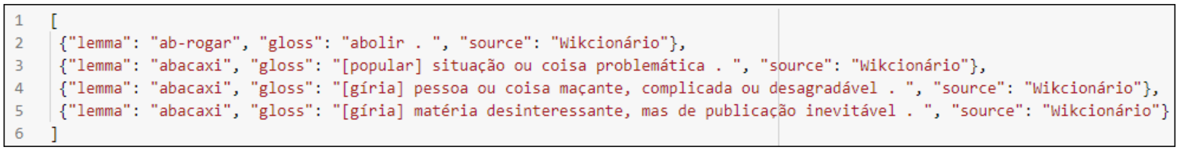
After validating the URLs, the program proceeded to retrieve the page's contents with the Beaultiful Soup library[7]. Requests to the websites were sent in a random interval from ten to fifteen seconds to avoid being blocked by the server. From the pages, lemmas and definitions were retrieved. The contents were stored in batches of JSON files containing 100 definitions each. Exception handling was employed when an error in retrieving the page was found so that the routine was not interrupted in between. In the end, an error report was written. In the collection phase, words were grouped by the first letter to avoid errors and loss of all

[6] Available in: https://kaikki.org/

[7] Available in: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

contents retrieved. It took approximately four weeks using three computers to retrieve the full extent of the wordlist between scraping and re-scraping routines.

In the end, 812 files were retrieved for Dicio, totalling 81.200 items. For Wikcionário, 457 files were retrieved, adding up to 45.700 items. An item consists of a python dictionary having lemma, gloss, and source as fields. Figure 3.1 shows a list containing four items. Each file contained a list of 100 python dictionaries. The main reason for choosing this format is that it is both human and machine-readable, so that I could inspect the material and correct errors manually if necessary.



```
[
 {"lemma": "ab-rogar", "gloss": "abolir . ", "source": "Wikcionário"},
 {"lemma": "abacaxi", "gloss": "[popular] situação ou coisa problemática . ", "source": "Wikcionário"},
 {"lemma": "abacaxi", "gloss": "[gíria] pessoa ou coisa maçante, complicada ou desagradável . ", "source": "Wikcionário"},
 {"lemma": "abacaxi", "gloss": "[gíria] matéria desinteressante, mas de publicação inevitável . ", "source": "Wikcionário"}
]
```

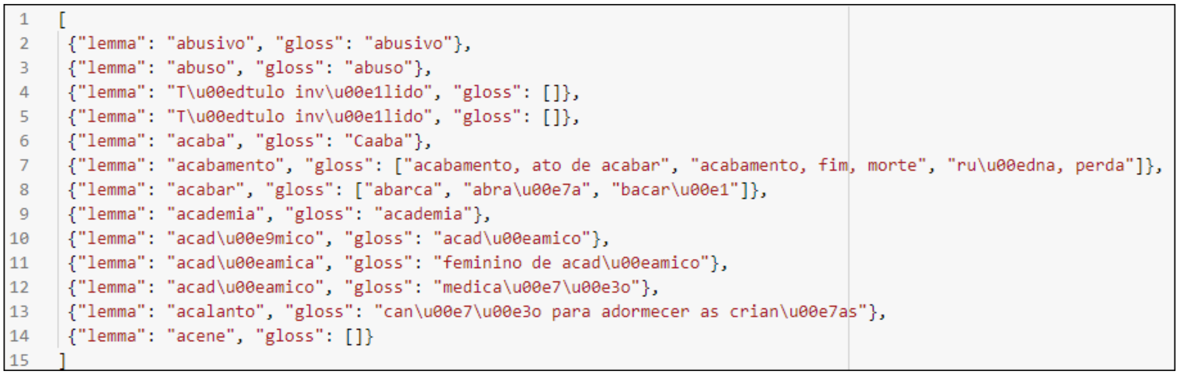
Figure 3.1: *An example of how the glosses were stored.*

Given the well-organised nature of Dicio's HTML files, there was no need to perform any cleaning or pruning[8] steps after scraping. On the other hand, contents retrieved from Wikcionário needed to be carefully cleaned. Many issues resulted from the irregular structures of the pages. For example:

1. Wikcionário provides valid links for blank entries. So, when the content of a blank page was retrieved, the lemma field became "T0̆0edtulo inv0̆0e1lido", and the gloss was empty.

2. Sometimes, polysemic lemmas lead to multiple definitions stored in a single list.

3. There is no HTML tag or CSS element to distinguish between definitions and examples (when they occurred) on the HTML page. So, oftentimes, the program would fetch wrong information from different fields (synonyms, anagrams) because the formatting either resembled the definition position or a structural markup.



```
[
 {"lemma": "abusivo", "gloss": "abusivo"},
 {"lemma": "abuso", "gloss": "abuso"},
 {"lemma": "T\u00edtulo inv\u00e1lido", "gloss": []},
 {"lemma": "T\u00edtulo inv\u00e1lido", "gloss": []},
 {"lemma": "acaba", "gloss": "Caaba"},
 {"lemma": "acabamento", "gloss": ["acabamento, ato de acabar", "acabamento, fim, morte", "ru\u00edna, perda"]},
 {"lemma": "acabar", "gloss": ["abarca", "abra\u00e7a", "bacar\u00e1"]},
 {"lemma": "academia", "gloss": "academia"},
 {"lemma": "acad\u00e9mico", "gloss": "acad\u00eamico"},
 {"lemma": "acad\u00eamica", "gloss": "feminino de acad\u00eamico"},
 {"lemma": "acad\u00eamico", "gloss": "medica\u00e7\u00e3o"},
 {"lemma": "acalanto", "gloss": "can\u00e7\u00e3o para adormecer as crian\u00e7as"},
 {"lemma": "acene", "gloss": []}
]
```

Figure 3.2: *Example of a file with raw glosses from Wikcionário.*

[8] This thesis will use cleaning and pruning as steps adopted to treat raw data, whereas pre-processing will be reserved for steps adopted to adjust the data for the experiments.

Figure 3.2 shows an example of the contents of a file with raw glosses. The initial approach was to check and correct errors manually to guarantee outstanding data quality and have a better understanding of the data. However, throughout the time, it became clear that this task was unattainable given the limited time, so semi-automatic pre-processing was applied. First, invalid titles and empty glosses were automatically removed. Second, lists containing more than one definition per lemma were automatically split. Third, Unicode symbols were converted into UTF-8 letters. Finally, tags with extra information (such as register and domain) were encapsulated in brackets.

Furthermore, unretrieved glosses (where lemma was equal to definition) were excluded when it was not possible to retrieve manually or re-scrape. Finally, a double-check script was employed to check for anagrams and symbols (arrows, chevrons, and any other unusual characters) and remove them. An in-depth description of the final dataset after cleaning is presented in Section 3.2.

Since the Portuguese dataset is rather small in comparison to the English dataset (explained in Section 3.2.2), an attempt was made to extend the number of samples by exploiting BabelNet[9] (Navigli & Ponzetto 2012). BabelNet is a multilingual encyclopedic dictionary containing more than 200 million entries combined from Wiktionary, Omega Wiki, Wordnet, and many other resources) in various languages, summing up to more than 40GB of data.

In order to exploit it, researchers can request the download of its database by signing an agreement (attached to Appendix B) containing indices meant to be exploited using their API. Unfortunately, the API is only available in Java. Although attempts have been made to create a script to extract the contents of the database, the process still depends on a wordlist to query and retrieve entries. Besides that, Babelnet entries for Portuguese do not include examples either. For these reasons, in addition to the difficulty of dealing with Java, this resource was discarded.

3.1.2 Collecting Contexts for Portuguese

In the beginning, it was thought that dictionaries would contain contexts that would be easily linked to the definitions; in reality, that did not hold true. Nevertheless, granted the possibility of using pre-trained language models for definition modelling, having contexts is essential to verify whether the model can generate a definition for a word-in-context.

Preferably, the best way to go would be to obtain an already pre-processed corpus in Portuguese. The resource Corpus do Português e Modelos Diversos[10] is available on GitHub. The project aims at providing a cleaned and pre-processed version of Wikipedia dumps for Portuguese. The only problem, however, was that the entire corpus comes together in a single file, which would require information extraction.

As an alternative, the Wikipedia[11] python package was utilised due to its simplicity and practicality. The package retrieves content directly from the website, given a word to the entry. All lemmas collected for Portuguese in the previous section were given as a seed list, and two sentences containing the lemma were retrieved. The disambiguation function was also leveraged to collect multiple contexts for a given lemma if they existed in the platform. At first, 124,800 items were collected.

[9] Available in: https://babelnet.org/

[10] Available in: https://github.com/fabiocmazzo/corpusportugues

[11] Available in: https://pypi.org/project/wikipedia/

The next step was cleaning and pruning. This step entailed converting upper-case letters into lowercase, removing line and paragraph breaks, and erasing parentheses from lemmas resulting from the automatic disambiguation. Finally, sentences where the lemma did not occur were removed. A script was run to remove duplicates and randomise the samples in the end.

A significant drawback of the automatic disambiguation is that frequently the package retrieved named entities articles. In order to remove these, the Spacy[12] library for Portuguese was employed. Even so, the tool did not recognise many named entities. Then, manual cleaning was set off to remove some of the most critical cases. Unfortunately, the limited time did not allow for perfection. However, given that this dataset would be mainly used for testing purposes, the issue was not as aggravating as if used for training. After cleaning, 78,887 contexts remained.

Finally, an attempt was made to connect definitions to contexts by employing sentence embeddings and the cosine measure. However, it was not successful, leaving room for future research on this topic.

Overall, the data collection phase for Portuguese was long and laborious, given the multiple obstacles found in the process. Data collection for English, nevertheless, was very straightforward, as described below.

3.1.3 Collecting Definitions for English

Initially, the idea was to collect definitions from the COBUILD Dictionary[13], given the contextual nature of its definitions. COBUILD is now a product of the Collins Dictionary company. Although they granted access to their API and enabled collecting a thousand definitions for this project, it was forbidden to store and/or cache their contents. Therefore, it was not viable to use COBUILD's definition for this research.

The next alternative was to exploit the other datasets available for English. As discussed in the background knowledge section, several datasets have been already proposed for English definition modelling due to the high-resource nature of the language.

For this thesis, datasets containing both lemma-definition and definition-context were considered. Table 3.2 summarises the datasets available from previous investigations.

Investigation	Dataset	N. of definitions	Contents
(Noraset et al. 2016)	GCIDE and Wordnet	162,925	def
(Gadetsky et al. 2018)	Oxford Dictionary	122,319	def + cont
(Kabiri & Cook 2020)	Wiktionary, Omega and Wordnet	869,700	def + cont[14]
(Bevilacqua et al. 2020)	Hei ++ - free MWEs	713	def
(Mickus et al. 2022)	Codwoe Dataset	63,596	def + cont

Table 3.2: *List of English datasets proposed in previous investigations on DM, where def stands for definition, and cont for contexts.*

[12] Available in: https://spacy.io/

[13] Available in: https://www.collinsdictionary.com/dictionary/english

As discussed previously, deep-learning systems are data-hungry, so ideally, the more data, the better. Therefore, all datasets in Table 3.2 were collected and converted into JSON files, except for the HEI++ dataset (more details below). After that, a simple script was applied to remove duplicated content since it is difficult to check each definition manually.

The HEI++ dataset contains free multiword expressions such as clear sky and starry night. Because it does not contain traditional lemmas, it was set aside for evaluation purposes only. The dataset was created by lexicographers for Generationary (Bevilacqua et al. 2020), and, for Portuguese, it was translated by a professional translator and reviewed by me for this thesis. Section 3.2.2 describe the figures for the English dataset.

3.1.4 Collecting Definitions for French, Spanish and Italian

Finally, cross-lingual experiments demanded a dataset comprising other languages. Romance languages (Italian, Spanish, and French) were chosen due to the similarity between them, as it may be possible that the model learns to define by leveraging the knowledge from the other languages, given that the Portuguese dataset is small.

In order to optimise and facilitate the collection for these languages, the CodWoe (Mickus et al. 2022) dataset was chosen, given that the data is already pre-processed in the format required for the experiments, which is time-wise beneficial. Table 3.3 outlines the features of the CodWoe (Mickus et al. 2022) dataset per language.

Language	Number of definitions	Content
Italian	62,465	lemma – definition
French	200,880	lemma – definition – context
Spanish	75,057	lemma – definition
Total	338,402	-

Table 3.3: *Number of definitions and overall content per language in the CodWoe dataset.*

As shown in Table 3.3, neither the number of sentences nor the content of the dataset is balanced across languages, probably due to the difficulty of compiling a dataset for such a specific genre. As the dataset is meant to be used all together, the sum of all sentences is also shown in Table 3.3.

3.2 Data Description

This section provides a detailed description of the data after cleaning and pruning. For organisation purposes, definitions received id numbers in the pre-processing phase, consisting of the following pattern: [language].defmod.[number of the definition].json, resulting in pt.defmod.235, for example. Languages were written in the two-letter code following the ISO 639-1 standard: Portuguese became ‘pt’, English became ‘en’, French became ‘fr’, Spanish became ‘es’, and Italian became ‘it’. Files were also labelled according to the following pattern: [language].defmod.[batch number].json. Let us now examine the figures for the Portuguese dataset.

3.2.1 The DEFMOD_PT Dataset

The Portuguese dataset can be divided into two sub-datasets: one containing lemma and glosses collected from Dicio and Wikcionário, and another one containing lemma and contexts collected from Wikipedia. The first subset (henceforth **PT subset 1**) includes 103,019 raw definitions in total. The second subset (henceforth **PT subset 2**) has 19,737 items on the whole. Subset 1 will be described first. Figure 3.3 shows the number of glosses by source for the first subset.

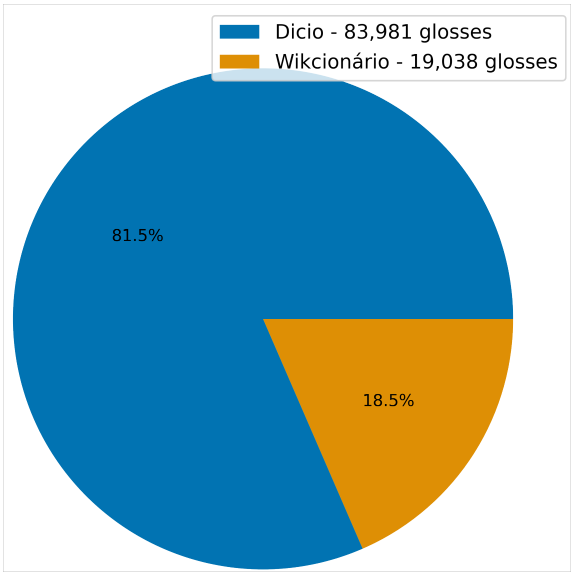


Figure 3.3: *Number of glosses per source for PT subset 1.*

PT subset 1 encompasses 27,978 lemma types (or unique lemmas). Figure 3.3 shows the number of senses per lemma by sources. Most lemmas from Wikcionário have one-word definitions, that is, glosses. This is expected because the website is open to anybody who wishes to edit it. Surprisingly, the word with most senses is *bater* with 76 senses.

Given that this number seems a bit inflated, a closer inspection revealed that the scraper ended up collecting some definitions from a previous version of the website, which was updated right when the scraping stage was ongoing. It is not a problem because the definitions are not duplicated. On average, there is a mean of 5.43 senses per lemma, with a standard deviation of 2.72.

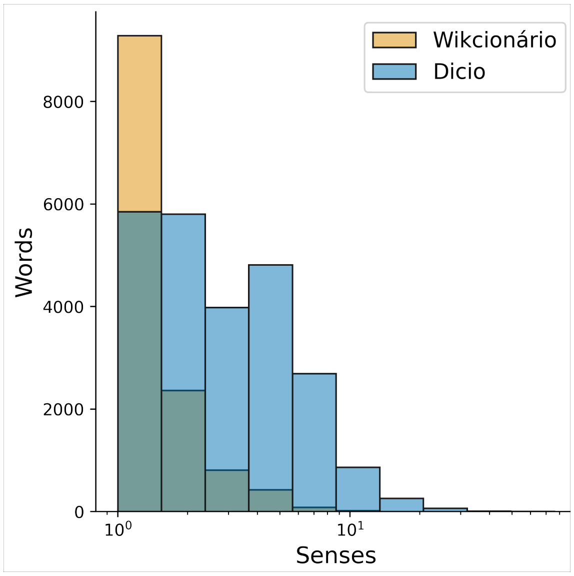


Figure 3.4: *Number of senses/words by source for PT subset 1 in log-scale.*

Among the 103,019 lemmas, there are 2,104 multiword expressions. Curiously, all of them are monosemic. Figure 3.5 shows the distribution of multiword expressions[15] by source.

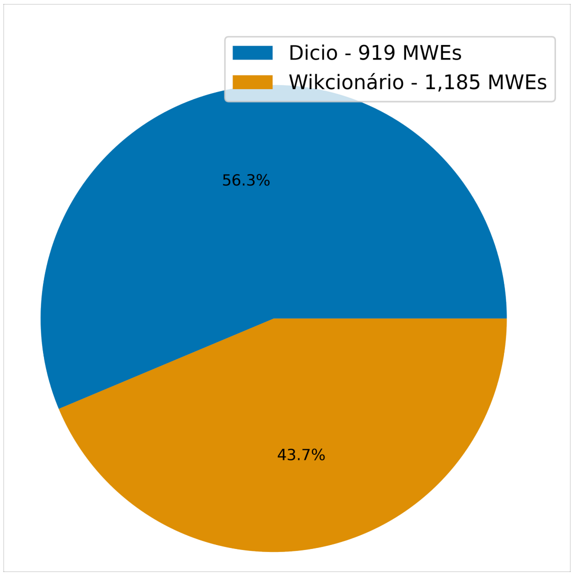


Figure 3.5: *Number of multiword expressions by source for PT subset 1.*

An important feature for transformer-based deep-learning models is the size of the sentences. Figure 3.6 shows that Dicio's definitions are longer than Wikcionário's. The main reason for that is that Wikcionário usually contains glosses or synonyms as their preferred method for

[15] Note that the MWEs were simply collected as provided by the dictionaries. It is not in the scope of this thesis to discover MWEs.

defining, making sentences smaller. Typically, sentences are 72.38 characters long or 11.38 words long, with a standard deviation of 48.81 characters or 7.60 words.

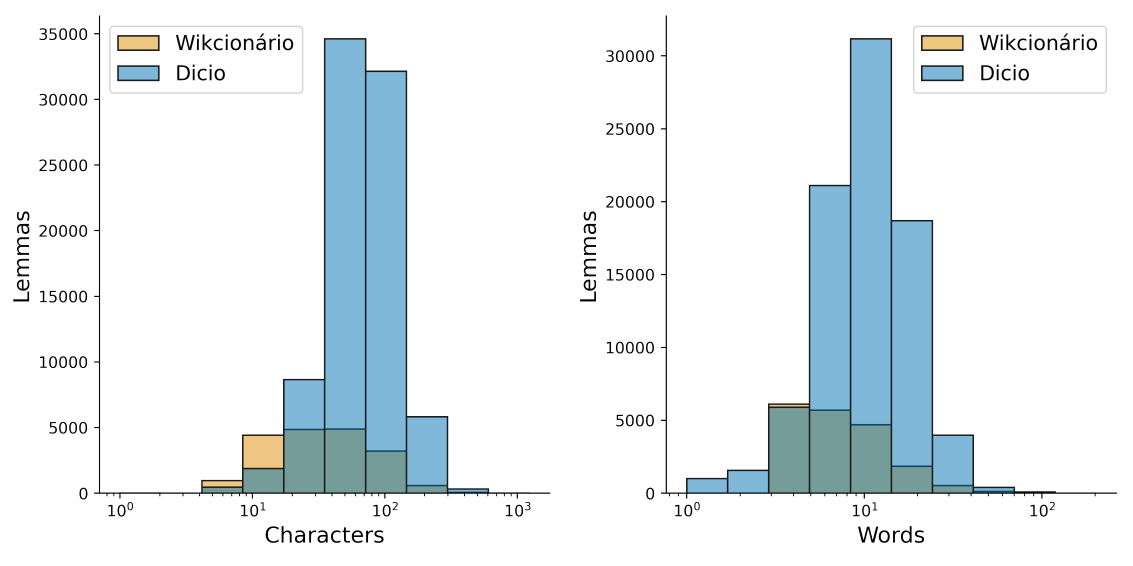


Figure 3.6: *Length of sentences in char and words by source for PT subset 1.*

Now moving on to PT subset 2, this subcorpus contains 19,737 contexts with 18,292 unique lemmas. Sentences are 167.63 characters long, or 27.99 words long, with a standard deviation of 83.10 characters or 12.94 words. Figure 3.7 displays the length distribution by character and words in PT subset 2.

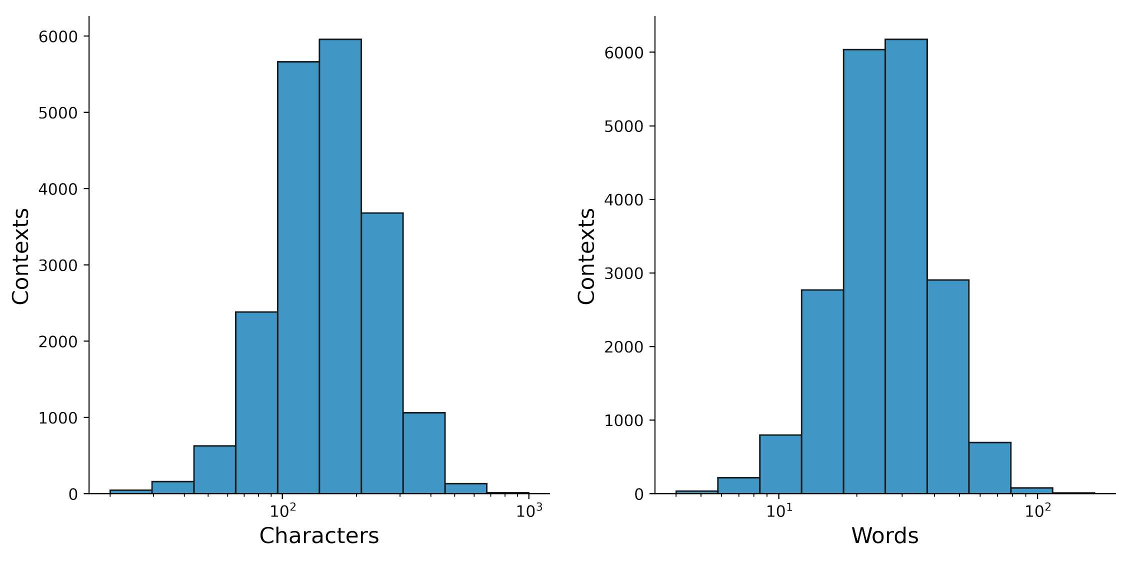


Figure 3.7: *Length of sentences in char and words for PT subset 2.*

Finally, as mentioned in Section 3.1.3, the HEI++ Dataset was translated from English into Portuguese. It contains 702 monosemic multiword expressions with their respective definitions.

Sentence length contains, on average, 51.66 characters or 8.5 words with a standard deviation of 16.80 characters or 2.75 words. Figure 3.8 summarises the length of the sentences in the HEI_PT Dataset.

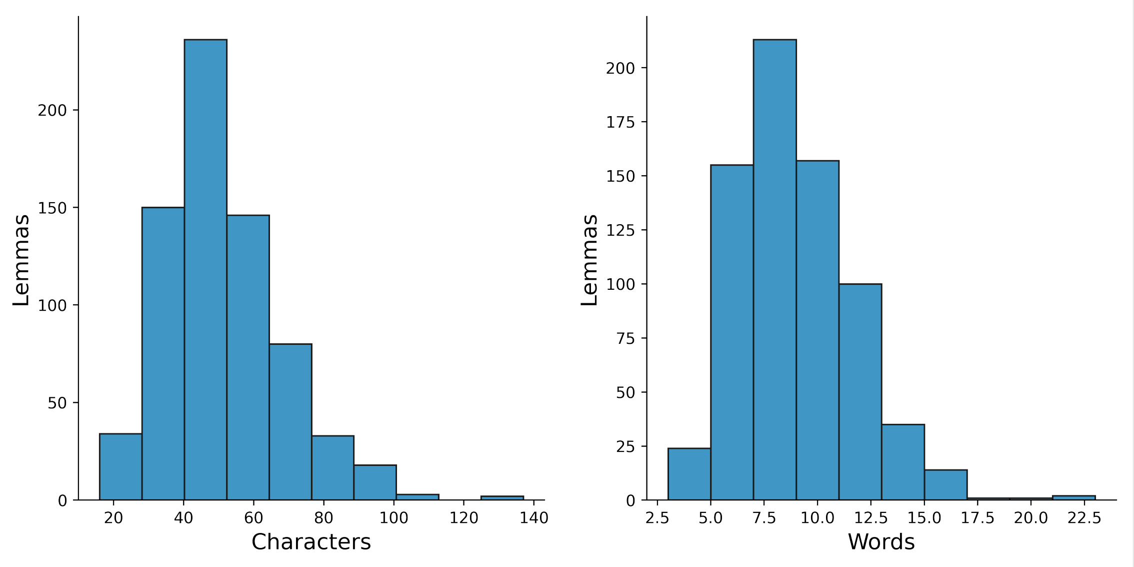


Figure 3.8: *Sentence length in char and words in the HEI_PT Dataset.*

3.2.2 The English, French, Spanish, and Italian Datasets

Given that the following datasets have already been analysed in depth in their original papers, Table 3.4 outlines the main features of all the other datasets.

Content	types	senses/lemma	avg len - char	avg len - words
877,001 EN def + cont	509,944	1.72	56.04	9.46
200,880 FR def + cont	55,068	6.2	75.63	14.30
75,057 ES def	33,860	2.12	80.84	14.75
62,465 IT def	35,987	1.73	78.97	13.61

Table 3.4: *A summary of essential information for English, French, Spanish and Italian datasets.*

The sum of all English datasets from the previous sections resulted in 1,064,995 items. Nonetheless, when duplicates were removed, 877,001 remained with 509,944 types.

3.3 Definition Modelling with Deep Learning

In order to generate definitions, two models were employed in this thesis. Here, Definition Modelling is cast as a sequence-to-sequence task, as proposed by (Mickus et al. 2020). Model 1, henceforth **baseline model**, tackles the generation of glosses for a word given the embeddings of its definition (a vector to sequence task – vec2seq), as shown in Figure 3.9.

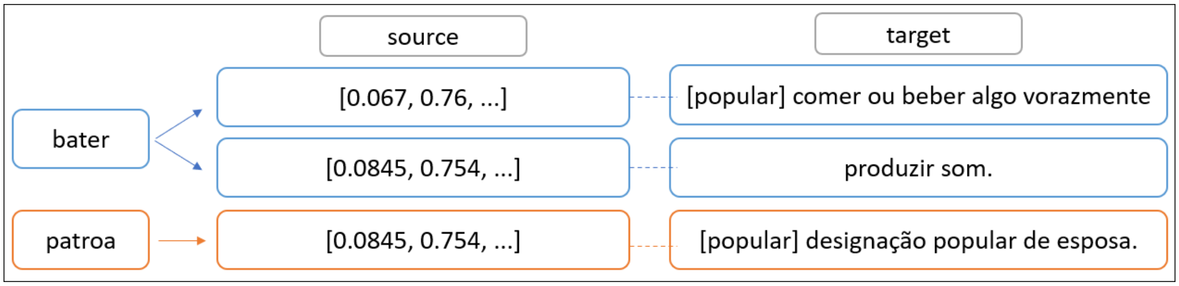


Figure 3.9: *Schema of the baseline model vec2seq processing.*

In a similar fashion, Model 2, henceforth **MT5 model**, addresses the generation of glosses provided a context sentence (the source) and its definition (the target) as simplified in Figure 3.10. In what follows, more details are given regarding the setup of the models.

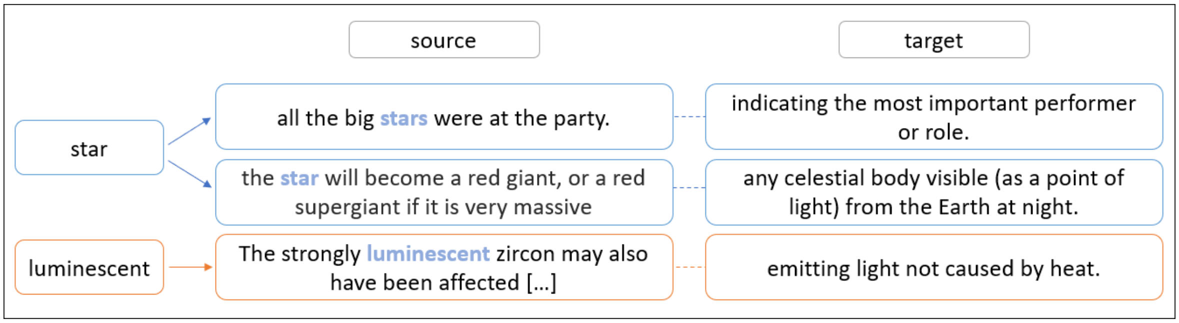

source
target
star
all the big stars were at the party.
indicating the most important performer or role.
the star will become a red giant, or a red supergiant if it is very massive
any celestial body visible (as a point of light) from the Earth at night.
luminescent
The strongly luminescent zircon may also have been affected [...]
emitting light not caused by heat.

Figure 3.10: *Schema of the MT5 model seq2seq processing.*

3.3.1 Baseline Model

The baseline model was proposed in the CodWoe Sharedtask (Mickus et al. 2022) at the end of 2021. Given that it is a vec2seq model, it takes different embeddings as input and learns how to map embeddings to definitions, reducing it to the auto-regressive[16] generation of glosses. The model consists of a multi-head, attention-based encoder-decoder architecture with four layers.

By default, the baseline architecture accepts three types of embeddings separately: character-based embeddings (henceforth **char embeddings**), skip-gram with negative sampling embeddings, pre-trained word2vec (henceforth **sgns embeddings**), and transformer-based contextualised embeddings, electra embeddings (henceforth **transformer embeddings**).

Furthermore, the model employs (1) Bayesian optimisation to handle hyperparameter optimisation, (2) sentence-piece retokenisation so that vocabulary has the same size throughout languages, and (3) beam-search decoding to generate different definitions.

In order to have the model up and running, the first step is to have files in the JSON format (following the formatting explained in the data collection). After that, glosses should be vectorised separately. For these experiments herein proposed, three types of vectors were also employed.

[16] One word after the other, conditioned on the previous sequence.

Originally, the plan was to adopt only Flair embeddings (Akbik et al. 2018). These embeddings are contextual string embeddings, and they bring multiple advantages. The main benefit of implementing this model is that words and contexts are modelled as a sequence of characters, preserving contextual and polysemous meaning, besides being more robust to misspelt words and hapax legomena.

Furthermore, the flair library also enables stacking – merging forward and backwards embeddings, in addition to combining pre-trained transformed embeddings, making them extremely powerful and versatile. However, the main drawback of employing Flair embeddings is their enormous size – 4096 dimensions.

Indeed, the size of the embeddings turned out to be an enormous problem. Recall that the dataset is divided into multiple batches of 100 definitions. After vectorising with stacked Flair embeddings, each file ended up with 9,57MB in size. Adding the size of all files would result in more than 80GB for the entire English dataset only.

Given that the RGCL server has 126GB of RAM memory, it was not possible to concatenate all files and hold them in memory to perform train/test split. An alternative was performing this operation in Google Colab Pro+, but it was also unsuccessful due to memory issues.

Then, files were manually separated into train/dev/test folders to split. Although splitting was successful at this stage, the baseline model loads train, validation, and testing sets simultaneously, which becomes unwieldy and computationally expensive. As a result, the kernel kills the process, and the experiments do not even start.

To surmount this obstacle, multiple tests were performed. First, I tried to decrease the number of decimal digits of vectors, demanding re-processing the whole dataset every time. In the first test, numbers were rounded from 14 to 9. It did not work; the final file was still too heavy.

The next attempt reduced decimal places to six, which still yielded heavy files (10GB for the training set). Lastly, reducing vector size to 4 decimal points resulted in a 4GB training dataset. Of course, this is highly detrimental to the experiments since information richness is already lost, but an attempt to execute the model again showed that it would not be possible to start experimenting with these embeddings either way.

A final attempt was made to reduce even more the size of the embeddings by removing transformer embeddings from the stacks. This test revealed that, in fact, standard transformer embeddings are not as heavy as it seems. Moreover, it was clear that it would not be possible to utilise flair embeddings in the other settings that require concatenating the datasets, making the 4GB become 15GB, for example. For these reasons, the vectorisation process restarted with different types of embeddings.

First, static character-based embeddings (Bojanowski et al. 2017*b*) from the FastText library were employed. The main reason for choosing this model is that embeddings are light and available in multiple languages. FastText embeddings have 300 dimensions by default. Main disadvantages include the fact that these embeddings are not contextualised; they do not account for polysemy. Besides that, unfortunately, FastText embeddings are not trained in multiple languages simultaneously; hence for cross-lingual and multilingual embeddings, we will combine the embeddings for the sake of experimenting, although it is expected to obtain poor results.

Finally, transformer-based embeddings are generated. For monolingual embeddings, BERTimbau (Souza et al. 2020) and RoBERTa (Liu et al. 2019) vectors were produced for Portuguese and English, respectively. These embeddings have 512 dimensions by default, although it is possible to specify larger (and richer) ones. For cross-lingual and multilingual experiments, XLM-RoBERTa (Conneau et al. 2020) was harnessed since the model was trained in more than 100 languages, including, of course, the working languages of this thesis. All embeddings are generated from the base version of the models as a result of the previous size constraints.

Having observed that preliminary results from the baseline were not satisfactory as desired, another model was devised to overcome the baseline model limitations.

3.3.2 MT5 Model

Departing from the hypothesis that pre-trained models can provide encyclopaedic knowledge that will (hopefully!) enable or enhance the model's power to define, in order to conduct seq2seq experiments, a large encoder-decoder model should be used.

Ideally, a monolingual model should be used to generate the monolingual definitions, but for Portuguese, only BERTimbau is widely accessible. The model consists of an encoder only, precluding language generation. Then, at first, the plan was to take advantage of Multilingual BART (Liu et al. 2020). However, the model was not trained in Portuguese, which is a major drawback given that no research has been conducted for this task and language to the best of my knowledge.

Considering all these limitations and that experiments should be conducted in different settings to answer the research questions, Multilingual T5 (MT5) (Xue et al. 2021) from Google Research is harnessed. This model is based on the Text-to-Text Transfer Transformer model (Raffel et al. 2020), achieving state-of-the-art results in multiple generative tasks.

MT5 diverges little from T5 in the sense that both share powerful ideas behind their architecture. First, the model builds upon the idea of unifying all NLP tasks under the Text-to-Text umbrella, which is highly beneficial for generative tasks such as machine translation, abstractive summarisation, and, as I hypothesise, definition modelling.

Similarly to the baseline model, MT5 comprises a standard encoder-decoder architecture, trained on "masked language modelling span-corruption objective" (Raffel et al. 2020). The model was trained on an 800GB colossal dataset in 701 languages with up to 13 billion parameters (Xue et al. 2021).

An obvious shortcoming of such a massive model is that it is relatively expensive to train and transfer learning. Therefore, small MT5 is employed for the seq2seq experiments in this thesis for computational power and storage limitations.

Concerning the experimental task, definition modelling with MT5 is formally taken as a typical sequence-to-sequence conditional generation. The model learns to map the probability of a gloss occurring given a context-definition pair by calculating it auto-regressively.

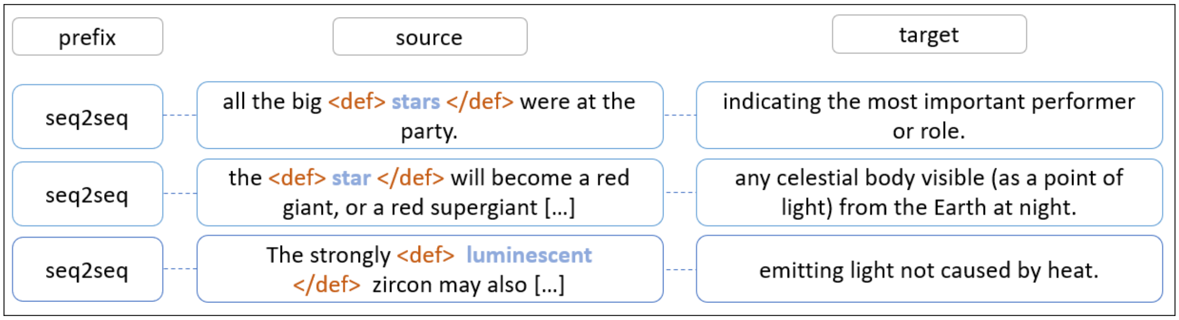
To finetune the model, the Simpletransformers[17] library is employed with the same number of parameters as small MT5, i.e., 300M. Hyperparameter details and reproducibility features are disclosed in the E.

The first experiment with MT5 revealed that the model does not learn at first which word should be defined in the context-definition pair. To solve this issue, the *definienda* is wrapped in tags, such as <def> </def>.

[17] Available in: https://simpletransformers.ai/

Another important feature of MT5 is that the adoption of the Text-to-Text paradigm enabled the model to multi-task. To leverage this power, a prefix specifying the task should be added to the input.

In the same fashion as the baseline model, training consists of feeding the model with source-target pairs besides adding the prefix, resulting in prefix + source + target triples, as shown in Figure 3.11.



prefix	source	target
seq2seq	all the big <def> stars </def> were at the party.	indicating the most important performer or role.
seq2seq	the <def> star </def> will become a red giant, or a red supergiant [...]	any celestial body visible (as a point of light) from the Earth at night.
seq2seq	The strongly <def> luminescent </def> zircon may also [...]	emitting light not caused by heat.

Figure 3.11: *An illustration of MT5 training input.*

Finally, the MT5 model deals with encoding the sentences into vectors so that it is not necessary to vectorise the input in advance. The following section examines the mechanics of the experimentation phase.

3.4 Experiments

Machine learning experiments take three iterative phases: training, validation, and predictions. First, following best practices in the field, datasets were divided into training and testing sets following the 75:25 split. Next, the training set was re-divided into 80:20 split for training and development (validation) datasets.

As pointed out in previous sections, experiments were also conducted in three different settings: monolingual, cross-lingual, and multilingual. Table 3.5 shows how experiments were organised and their main features for the baseline model.

At first, in order to finetune and find the best configuration in hyperparameters for embeddings and data size, each experiment ran for ten epochs, then for 25 epochs, and finally reaching 50-55 epochs at most due to the time it takes to train (models running with English data could take more than five days to finish). Time constraints and data size deeply affected the execution of baseline experiments; more details are given in the results section).

Table 3.6 summarises the setup for experimentation for the MT5 model. MT5 experiments, likewise baseline experiments, were also affected by time constraints and data size (details are provided in the results section). The model can take more than one week to run for ten epochs depending on the size of the dataset.

3.4.1 Aligning Definitions

At the start of this project, the intention was to align glosses to produce bilingual glosses. Nevertheless, the data collection procedure revealed that it would not be possible to build

Experiment label	Setting	Language	Input	Output
Experiment 1.1	Monolingual	PT	Char embed	PT glosses
Experiment 1.2	Monolingual	EN	Char embed	EN glosses
Experiment 1.3	Monolingual	PT	Flair embed	PT glosses
Experiment 1.4	Monolingual	EN	Flair embed	EN glosses
Experiment 1.5	Monolingual	PT	Transf embed	PT glosses
Experiment 1.6	Monolingual	EN	Transf embed	EN glosses
Experiment 2.1	Cross-lingual	PT	Char embed	PT glosses
Experiment 2.2	Cross-lingual	EN	Char embedd	EN glosses
Experiment 2.3	Cross-lingual	PT and EN	Char embed	Both PT and EN glosses
Experiment 2.4	Cross-lingual	PT	Cross-lingual transf embed	PT glosses
Experiment 2.5	Cross-lingual	EN	Cross-lingual transf embed	EN glosses
Experiment 2.6	Cross-lingual	PT and EN	Cross-lingual transf embed	Both PT and EN glosses
Experiment 3.1	Multilingual	PT	Char embed	PT glosses
Experiment 3.2	Multilingual	EN	Char embed	EN glosses
Experiment 3.3	Multilingual	PT and EN	Char embed	Both PT and EN glosses
Experiment 3.4	Multilingual	PT	Multilingual transf embed	PT glosses
Experiment 3.5	Multilingual	EN	Multilingual transf embed	EN glosses
Experiment 3.6	Multilingual	PT and EN	Multilingual transf embed	Both PT and EN glosses

Table 3.5: *A blueprint for the experiments and their features with the baseline model.*

a truly comparable corpus for Portuguese in terms of size or variety of resources, given the multiple limitations of where to obtain the source in the first place. In addition to that, there is a scarcity of contexts that preclude the generation of adequate definitions.

However, should glosses be generated for Portuguese, the alignment is done by employing sentence embeddings. Instead of vectorising word by word and averaging the embeddings, Multilingual Sentence-BERT (Reimers & Gurevych 2020) is employed to generate a proper embedding to the entire definition. To implement it, the Sentence Transformer library is used. Firstly, sentences are tokenised, and then cosine measure is employed to measure how similar the sentences are. Similar sentences should have a small cosine similarity between them.

3.5 Evaluation

In this section, the procedures adopted to evaluate the system's performance and the generated definitions are described. As described in the first chapter, the evaluation of DM is typically conducted through both intrinsic and extrinsic evaluation.

Experiment label	Setting	Language	Input	Output
Experiment 4.1	Monolingual	English	tagged contexts-def	EN glosses
Experiment 4.2	Monolingual	Portuguese	tagged lemma-def	PT glosses
Experiment 4.3	Cross-lingual	Portuguese	tagged cross-lingual contexts-def	PT glosses
Experiment 4.4	Multilingual	Portuguese	tagged cross-lingual contexts-def	PT and EN glosses

Table 3.6: *A summary of essential information for English, French, Spanish and Italian datasets.*

Concerning the automatic evaluation, first, automatic metrics were measured. Following the literature, this thesis reports BLEU(Papineni et al. 2002), BERTScore(Zhang et al. 2019), and MoverScore(Zhao et al. 2019). Since the number of generated definitions is quite high and it would be impossible to assess all definitions manually, this thesis also employs Quality Estimation (QE) to understand the quality of the glosses better.

Likewise machine translation, definition modelling is an open-ended task in the sense that there might be various valid outputs to the same input. This characteristic makes it harder to develop automatic metrics that correlate well with human judgement. On top of that, there is no metric specifically created for definition modelling.

Moreover, it is also hard to classify outputs as good or bad, especially when lexicographic users are diverse and possess many different needs. Lastly, human evaluation is costly; it is challenging to find suitable candidates willing to evaluate definitions voluntarily.

Another challenging issue regarding definition modelling, mainly for low-resourced languages, is the lack of references. QE aims to fill in this gap by supplying an "estimate of how good or reliable" (Specia et al. 2018) are the generated results regardless of a reference.

Furthermore, there is also an accountability issue. Now more than ever, more and more NLP systems are being incorporated into the daily life of workers, but increasingly in the LSP industry. Indeed, translators and post-editors have to justify their choices whenever required, which oftentimes means being able to find high-quality resources that can underpin the linguistic choices made for their translations.

From this perspective, having access to quality-related information is key. Considering that the Functional Theory of Lexicography is all about producing useful content for its users, adopting Quality Estimation in Definition Modelling can be particularly useful for translators and interpreters. QE has never been employed in DM to the best of my knowledge. Meanwhile, automatic-generated dictionaries are already being released without any indication of their automatic nature[18].

Since DM is similar to machine translation as a seq2seq task, in order to carry out QE, the TransQuest framework (Ranasinghe et al. 2020*a*) was adopted to perform sentence-level QE. The model achieved SOTA results for QE in machine translation (Ranasinghe et al. 2020*a*,*b*). It harnesses the power of cross-lingual transformer models to predict the quality of a sentence.

At first, in order to avoid cherry-picking in human evaluation, QE scores were to be used to determine which glosses were going to be evaluated by humans. However, since each experiment generated an enormous number of glosses, it was impossible to evaluate everything

[18] This is the case of LitMind, an English-Chinese dictionary for language learners entirely created with definition modelling.

and then start human evaluation. Nevertheless, QE were still generated in order to have more information regarding the quality of the glosses.

Human Evaluation

Human evaluation is a key aspect of natural-language generation tasks. For DM is also essential but hardly employed, first because glosses need to be at least readable in order to be evaluated, and secondly because it is hard to find suitable evaluators.

Nevertheless, the perception of the final users is of the utmost importance since they are the ones dealing with the outcome daily. Because of that, human evaluation with translators is for the first time implemented through this thesis to the best of my knowledge.

In order to carry out human evaluation, a minimum threshold was established: glosses need to be at least readable to be eligible for evaluation. In addition, glosses should achieve at least a 50% score in QE to be included in the evaluation. Finally, glosses were randomly selected by id with a python script to avoid selection bias.

To carry out this task, the online surveying method is used. Questionnaires were prepared well in advance before results were available and can be found online . A copy of all materials is attached to Appendix D of this thesis. Questionnaires were created using both Microsoft Forms[19] and Typeform[20]. The first was used to collect the participants' expressions of interest, and the second was used for the remainder questionnaires. Thus, three questionnaires were devised: (1) an expression of interest questionnaire (henceforth **Q1**), (2) a pre-questionnaire (henceforth **Q2**), and a (3) post-questionnaire (henceforth **Q3**).

The expression of interest serves the purpose of offering people a chance to participate in the study anonymously and freely. Once people indicated they wanted to take part in the study, they would receive an identification code (henceforth **ID**) that would follow their responses throughout the entire process. Since the questionnaires for participation were devised well in advance, there was time to circulate them to different people and gather feedback to avoid participants facing difficulties or misunderstandings.

By and large, the objective of applying the questionnaires was twofold: first, to understand whether the glosses can be used in a translation assignment (extrinsic evaluation). Then, to investigate the quality of the glosses. Table 3.7 outlines the structure of the questionnaires and their questions.

[19] Available in: https://forms.office.com/

[20] Questionnaire 1 available in https://annafurtado.typeform.com/to/u5b70rz2, and Questionnaire 2 available in https://annafurtado.typeform.com/to/gWq5FFD0

Questionnaire's section	Topic
Expression of interest	Demonstrating the wish to participate in the project
About you	Participants' age, qualifications and working languages
Lexicographic resources	Useful features for translators
Translation feedback	Use of the provided glosses and external materials
Usefulness of definition	Evaluation of provided glosses
Quality of the glosses	Evaluation of the glosses

Table 3.7: *A summary of the questions asked in the three questionnaires.*

The three-fold questionnaire study worked as follows: first, questionnaires for expressing interest were circulated on social media and via e-mails. Anyone interested in taking part in the study could simply read the recruiting form and provide any e-mail as they wished.

After they expressed their interest in participating, they received the ID, following the pattern DEFMOD_GROUP_PERSON, resulting in DEFMOD_1_3, for example. The ID assigned the participant to a three-individual group receiving a specific set of materials.

Each set of material comprised English/Portuguese texts that should be translated into/from Portuguese/English. Texts were constituted of 20 paragraphs extracted from scientific articles containing the words whose definitions were provided. Along with the material for translation, participants received a glossary covering definitions for the assignment, the link for the pre-questionnaire, and the link for the post-questionnaire.

The DEFMOD Glossary

The first step in crafting the translation glossary was to collect the definitions composing the resource. To do so, as explained previously, the idea was to exploit QE scores to establish a threshold for selecting the glosses. However, scoring did not go as fast as planned; thus, the selection of glosses for evaluation was carried out as follows: first, ids were randomly sorted with a python script, and should the gloss be readable, then it was selected to compose the glossary. Finally, ten glosses were selected per language.

To provide more consistency, two control glosses were added in each batch of 20 glosses: two golden references from a human-made dictionary. Altogether, each group should evaluate 22 glosses.

Also, at first, I thought of investigating the usefulness of the QE scores by adding the score assigned to the gloss so that participants could agree on the assigned score or not. However, this idea was not put forward since it was too risky. In addition, it could influence participants' perception of either the usefulness or the quality of the glosses they were evaluating.

Once chosen the glosses, initially, the idea was to collect ten paragraphs from only scientific journals to integrate the translation task. However, this task was proven to be quite challenging, given that the glosses are more inclined towards general language than specialised languages. Thus, to solve the issue quickly, journalistic texts were also considered.

It can be argued that selecting paragraphs was not a good choice, given that translators usually translate a complete text with full context. However, at this stage, the goal was not to evaluate the translation skills of the participants; instead, the focus was placed on the usefulness of the glosses, and for that, a paragraph should be enough, provided that there are no anaphorical uses.

Appendix D reports all materials and sources used for this task.

3.5.1 Evaluation Scale

First and foremost, it is important to mention that the target audience for this study was primary translators working with Portuguese as a native language. Obviously, the study will not preclude the participation of people working in the opposite direction (EN-PT); however, the entire study was designed for Portuguese-speaking groups, given the lack of research on DM for this audience.

Before starting the translation assignment, participants were asked to fill in the pre-questionnaire to express their consent to participate in this research and obtain more information about them.

Since the goal of this project is not to evaluate the use of CAT tools, participants received the raw text and were allowed to use CAT tools if they wished to do so. Additionally, participants were instructed to use the given definitions should they have any doubts. Finally, in case of definitions would not satisfy their needs, they were allowed to consult external resources.

After they finished the translation assignment, they were redirected to the post-questionnaire, where they would first upload the translation they made. After that, they were asked whether they used the glossary or different resources and which, if any.

The next step consisted of evaluating how useful the glosses were by applying a five-level Likert scale:

1. **not useful** - this gloss is not useful, and it does not help me understand the word.
2. **requires adjustment** - the gloss is understandable, but I would still prefer to consult other sources.
3. **good** - the gloss is vague and generalist, I would still prefer to consult other sources.
4. **good** - the gloss is good, but it misses some details.
5. **useful** - fit for purpose.

After that, given a context, they were required to evaluate whether the gloss is adequate to describe the word in context, following (Bevilacqua et al. 2020). The context is contained in the paragraph given in the translation assignment, for Portuguese. For English, the context is collected from the gold reference dataset, where the context was given to generate the gloss. To evaluate, the following five-level Likert scale was employed:

Context: *Was he going to be saddled with a creep for a **bar-buddy**.*[21]
target word: **bar-buddy**.

- 1. **Wrong gloss** - may refer to a homonym of the target.
- e.g.: a heating element in an electric fire.

[21] Example extracted from (Bevilacqua et al. 2020).

- 2. **Wrong gloss** - captures the domain of the target.
- e.g.: a counter where you can obtain food or drink.
- 3. **Correct gloss** - Utterly vague and generic.
- e.g.: a person with whom you are acquainted.
- 4. **Correct gloss** - Fits the context, but misses some details.
- a close friend who accompanies his buddies in their activities
- 5. **Correct gloss** - Perfectly describes the target in its context
- e.g.: a friend who you frequent bars with.

Finally, translators could also make comments if they wished to do so.

3.5.2 Questionnaire Data Analysis

In order to perform the data analysis, quantitative analysis was performed for the closed questions. To do so, answers received a code; for example, in the question *"how old are you?"* the answer *"18 to 25 years old"* received the number 1, and so on.

Encoding the questions allowed to calculate the inter-evaluator agreement between the parties for the adequacy-related questions. Moreover, it did not make sense to calculate inter-evaluator agreement for usefulness-related questions. The main reason for that is that usefulness is highly subjective and can be subject to language-level, for example.

The inter-evaluator agreement can be calculated by using the Kappa Statistic (McHugh 2012). Ideally, the resulting score should be above 0.6.

The next chapter shows the results obtained in this thesis.

Chapter 4

Results

Overview - In this chapter, the results of the experiments are presented. This chapter is organised as follows: first, results are reported by settings and model. Then, results concerning the alignment are presented. Finally, human evaluation is reported.

4.1 Results for the baseline model

As described in the previous section, a critical issue for all experiments and settings in the training phase was the memory errors resulting from vector size. Given that there was no time to re-conduct all the experiments with the optimal dataset size, first, an initial train/test (resulting in train, dev, and test sets) split was made.

Then, if required, the data was split again using slicing for each experiment. This is certainly not the best approach to go; however, given that sometimes it would take more than four days to train a single experiment, splitting using the indexes was the only approach that could ensure reproducibility later and avoid data contamination.

Another problem is that, given the amount of data, it was impossible to run experiments for more than 50 epochs to achieve the minimal loss function or the best scores possible. The main reason was time constraints – 50 epochs could take six days for English with the baseline model. MT5, on the other hand, could only run for ten epochs, meaning it was running for an entire week. Unfortunately, transformer-based experiments draw heavily on computational power and still leave an enormous carbon footprint (Strubell et al. 2020).

Table 4.1 shows the training loss and number of epochs for the baseline model.

Experiment	Language/Embedding	Training Loss	Epochs
Experiment 1.1	PT – char	3.80	50
Experiment 1.2	EN – char	2.75	50
Experiment 1.3	PT - flair	2.44	50
Experiment 1.4	EN - flair	2.30	50
Experiment 1.5	PT - trans	2.67	50
Experiment 1.6	EN - trans	2.77	50
Experiment 2.1	PT - char	3.28	50
Experiment 2.2	EN - char	3.28	50
Experiment 2.3	PT and EN - char	3.28	50
Experiment 2.4	PT - trans	4.22	50
Experiment 2.5	EN - trans	4.22	50
Experiment 2.6	PT and EN - trans	4.22	50
Experiment 3.1	PT - char	4.45	50
Experiment 3.2	EN - char	4.45	50
Experiment 3.3	PT and EN – char	4.45	50
Experiment 3.4	PT – trans	4.7	50
Experiment 3.5	EN – trans	4.7	50
Experiment 3.6	PT and EN trans	4.7	50

Table 4.1: *Training information for all settings.*

Ideally, each experiment should have been run at least three times to verify the consistency of the predictions. However, given the time constraints, it was not possible to run multiple times for the same experiment.

The same is valid for scoring the automatic metrics. Unfortunately, given the vast number of generated glosses, scoring took much longer than foreseen. For this reason, best results are reported for the baseline model in Table **??**.

Experiment	Language/Embedding	Training Loss	Epochs
Experiment 1.1	PT – char	3.80	50
Experiment 1.2	EN – char	2.75	50
Experiment 1.3	PT - flair	2.44	50
Experiment 1.4	EN - flair	2.30	50
Experiment 1.5	PT - trans	2.67	50
Experiment 1.6	EN - trans	2.77	50
Experiment 2.1	PT - char	3.28	50
Experiment 2.2	EN - char	3.28	50
Experiment 2.3	PT and EN - char	3.28	50
Experiment 2.4	PT - trans	4.22	50
Experiment 2.5	EN - trans	4.22	50
Experiment 2.6	PT and EN - trans	4.22	50
Experiment 3.1	PT - char	4.45	50
Experiment 3.2	EN - char	4.45	50
Experiment 3.3	PT and EN – char	4.45	50
Experiment 3.4	PT – trans	4.7	50
Experiment 3.5	EN – trans	4.7	50
Experiment 3.6	PT and EN trans	4.7	50

Table 4.2: *Results for the baseline model.*

4.2 Results for the MT5 Model

The MT5 model ran for ten epochs in each experiment. Experiments with the English dataset were conducted, with only half of the datasets sliced in half by the index. In Appendix X, exact numbers for train/test split are given for each dataset. In Table 4.3, training information is provided for the MT5 model.

Experiment	Training Loss	Epochs
Experiment 4.1	1.050	10
Experiment 4.2	1.050	10
Experiment 4.3	1.050	10

Table 4.3: *Results for the baseline model.*

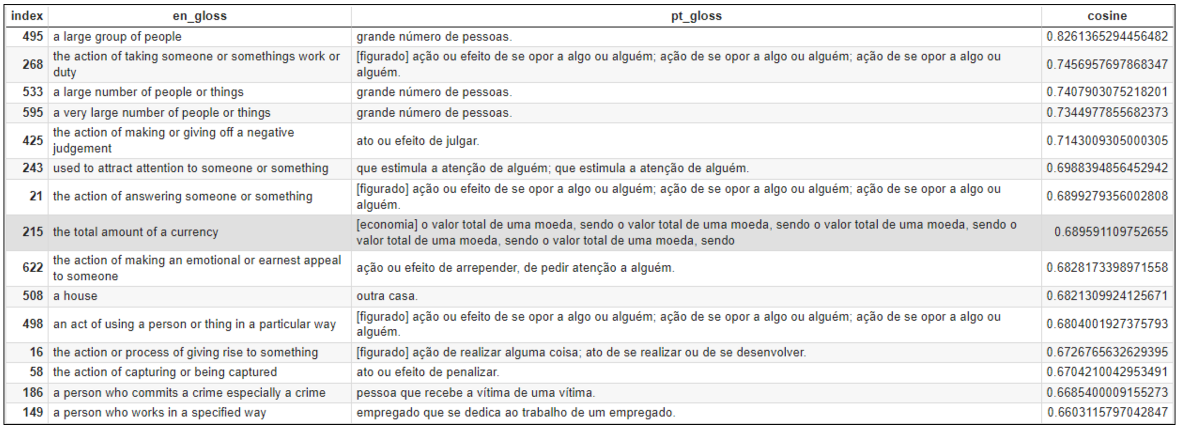
Given that experiments with MT5 seem to have produced more readable glosses, evaluation is also conducted with both EN and PT HEI datasets. Table 4.4 summarises automatic metrics for the MT5 model.

Experiment	BLEU	Precision	Recall	F1	QE
Experiment 4.1	0.2035	0.8889	0.8836	0.8860	0.6434
Experiment 4.2	0.0081	0.3305	0.3290	0.3197	0.2314
Experiment 4.3	0.1570	0.5209	0.4901	0.5109	0.5942
EN HEI ++	0.0076	0.8928	0.8861	0.8899	0.7234
PT HEI ++	0.0054	0.8578	0.8755	0.8664	0.7189

Table 4.4: *Results for the MT5 model.*

4.3 Aligning glosses

Since baseline results proved to be quite unreadable, aligning those glosses was unfeasible. However, the MT5 model yielded moderate results and the EN and PT HEI ++ datasets are comparable; thus, the aligning was applied only to them. Unfortunately, provided that there were no aligned references in order to compare to the automatically aligned one, from Figure 4.1, it is possible to see that alignment is more effective when the cosine measure is higher than 0.5.



index	en_gloss	pt_gloss	cosine
495	a large group of people	grande número de pessoas.	0.8261365294456482
268	the action of taking someone or somethings work or duty	[figurado] ação ou efeito de se opor a algo ou alguém; ação de se opor a algo ou alguém; ação de se opor a algo ou alguém.	0.7456957697868347
533	a large number of people or things	grande número de pessoas.	0.7407903075218201
595	a very large number of people or things	grande número de pessoas.	0.7344977855682373
425	the action of making or giving off a negative judgement	ato ou efeito de julgar.	0.7143009305000305
243	used to attract attention to someone or something	que estimula a atenção de alguém; que estimula a atenção de alguém.	0.6988394856452942
21	the action of answering someone or something	[figurado] ação ou efeito de se opor a algo ou alguém; ação de se opor a algo ou alguém; ação de se opor a algo ou alguém.	0.6899279356002808
215	the total amount of a currency	[economia] o valor total de uma moeda, sendo o valor total de uma moeda, sendo o valor total de uma moeda, sendo o valor total de uma moeda, sendo o valor total de uma moeda, sendo	0.689591109752655
622	the action of making an emotional or earnest appeal to someone	ação ou efeito de arrepender, de pedir atenção a alguém.	0.6828173398971558
508	a house	outra casa.	0.6821309924125671
498	an act of using a person or thing in a particular way	[figurado] ação ou efeito de se opor a algo ou alguém; ação de se opor a algo ou alguém; ação de se opor a algo ou alguém.	0.6804001927375793
16	the action or process of giving rise to something	[figurado] ação de realizar alguma coisa; ato de se realizar ou de se desenvolver.	0.6726765632629395
58	the action of capturing or being captured	ato ou efeito de penalizar.	0.6704210042953491
186	a person who commits a crime especially a crime	pessoa que recebe a vítima de uma vítima.	0.6685400009155273
149	a person who works in a specified way	empregado que se dedica ao trabalho de um empregado.	0.6603115797042847

Figure 4.1: *The first 10 glosses aligned.*

4.4 Human Evaluation

Three people participated in the human evaluation study. Two of them are 25–34-year-old, while one is 45-54-year-old. All of them have worked as professional translators for more than five years. Their main area of expertise is technical translation. They all translate from and into Portuguese-English. All participants shared that they consult both monolingual and bilingual dictionaries to tackle their translation assignments. As for translation needs, all of them report that monolingual definitions are useful, while one of them said it would be desirable to have bilingual definitions.

Concerning the post-questionnaire (Q3), all translators used external resources, and all of them used the glossary provided for the task. They also used corpora and different dictionaries to complete the task.

For the questions regarding the usefulness of the definitions, answers vary considerably. Glosses can be grouped as follows:

Group A - full agreement on **not useful**:

1. octave – synthetic organic compound which is a fossil which can be a constituent of many kinds of copper.
2. plasma – the colourless fluid part of a room or of a planet containing the suns apparent suspension of the candelabra.
3. deanery – the territory of an ancient roman dean.

Group B - full agreement on **very useful**:

1. displicência – falta de cuidado, de atenção; negligência, desatenção.
2. florear – [figurado] tornar elegante; enfeitar: florear a música.
3. corrode – of metal or other materials be destroyed or damaged slowly by chemical action.
4. dendritic – relating to or denoting a cell which causes progressive weakness in the cell walls of a tissue.
5. landfill – the disposal of waste material by burying it especially as a method of filling in and reclaiming excavated pits.
6. froth – a mass of small bubbles in liquid caused by agitation fermentation or salivating.

For adequacy-related questions, the inter-annotator agreement was calculated with the NLTK implementation[1] using the Fleiss Kappa, with which is possible to measure the agreement between two or more people. The score obtained was 0.67, which can be interpreted as a substantial agreement between the annotators. Furthermore, a complete agreement was reached in the gold glosses and also in:

1. Full agreement – **correct gloss** – fit for purpose:
   (a) Lesivo – que é capaz de lesar, de prejudicar, de causar danos, prejuízos; nocivo.
2. Full agreement – **correct gloss** – utterly vague and generic:
   (a) Abyss – a catastrophic situation seen as likely to occur.
3. Full agreement - **wrong gloss** – may refer to a homonym of the target:
   (a) Octave – synthetic organic compound which is a fossil which can be a constituent of many kinds of copper.

In the next section, results are discussed in light of the research questions.

[1] Available in: https://www.nltk.org/api/nltk.metrics.agreement.html

Chapter 5

Discussion

Let us now proceed to the discussion of the results in light of the research questions and objectives.

5.1 RQ 1: How effective are deep-learning techniques in generating glosses for Portuguese?

To answer RQ1 and sub-questions, nine experiments were conducted in Portuguese: Experiment 1.1 (monolingual char), Experiment 1.3 (monolingual flair), Experiment 1.5 (monolingual transformers), Experiment 2.1 (cross-lingual char), Experiment 2.4 (cross-lingual transformers), Experiment 3.1 (multilingual char), Experiment 3.4 (multilingual transformers), Experiment 4.2 (cross-lingual MT5), Experiment 4.3 (monolingual MT5).

From the figures provided in both Table 4.2 (baseline model) and Table 4.4, it is possible to see that the models are capable of generating definitions, even though they may have shown to be of poor quality (in the case of the baseline model) or moderately acceptable (in the case of the MT5).

In fact, results show that casting definition modelling as a sequence-to-sequence task, and employing a pre-trained model, yields better results. However, it can be argued that, at the end of the day, a sequence-to-sequence task can become a vector-to-sequence task since inputs need to be encoded in order to be processed.

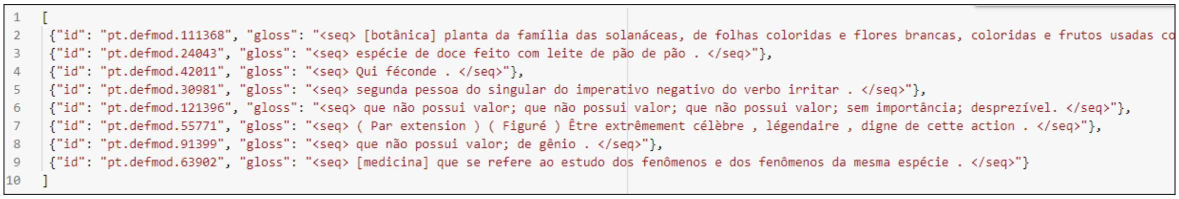
Moving on to the next sub-question, yes, indeed, adopting a pre-trained model produced better results in both quality and automatic metrics. Scores increase dramatically for all metrics when a pre-trained model is adopted.

As for the different types of embeddings, it is surprising how transformer embeddings fare worse than character-based embeddings for Portuguese, while flair embeddings perform better. One of the possible reasons for this may be related to the number of dimensions of the embeddings. Flair embeddings are far richer than both the other types, which may explain the increase in scores.

5.2 RQ2: Can we apply cross-lingual learning techniques to generate definitions for Portuguese?

Research question RQ2 and sub-questions led to the execution of two experiments: Experiment 2.1 (char) and Experiment 2.4 (transformer). Both experiments showed that cross-lingual learning techniques did not bring the best results for generating definitions for Portuguese. In fact, using the romance language dataset made the model confuse the languages and generate even poorer definitions. One main reason for these poor results is data balancing.

Experiment 4.3 with MT5 showed that cross-lingual and multilingual learning techniques could work when there is a fine balance between the proportion of languages in the dataset. Admittedly, the romance language training set for the baseline experiment was much larger than the Portuguese dataset, which definitely interfered with the model's learning, as illustrated in Figure 5.1.



```
[
 {"id": "pt.defmod.111368", "gloss": "<seq> [botânica] planta da família das solanáceas, de folhas coloridas e flores brancas, coloridas e frutos usadas co
 {"id": "pt.defmod.24043", "gloss": "<seq> espécie de doce feito com leite de pão de pão . </seq>"},
 {"id": "pt.defmod.42011", "gloss": "<seq> Qui féconde . </seq>"},
 {"id": "pt.defmod.30981", "gloss": "<seq> segunda pessoa do singular do imperativo negativo do verbo irritar . </seq>"},
 {"id": "pt.defmod.121396", "gloss": "<seq> que não possui valor; que não possui valor; que não possui valor; sem importância; desprezível. </seq>"},
 {"id": "pt.defmod.55771", "gloss": "<seq> ( Par extension ) ( Figuré ) Être extrêmement célèbre , légendaire , digne de cette action . </seq>"},
 {"id": "pt.defmod.91399", "gloss": "<seq> que não possui valor; de gênio . </seq>"},
 {"id": "pt.defmod.63902", "gloss": "<seq> [medicina] que se refere ao estudo dos fenômenos e dos fenômenos da mesma espécie . </seq>"}
]
```

Figure 5.1: *Glosses generated in cross-lingual settings.*

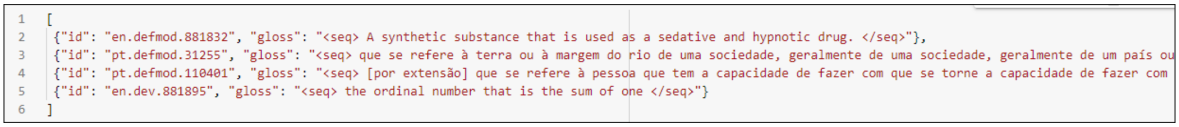
Unexpectedly, when fed with Portuguese contexts in the cross-lingual experiments, the MT5 model generated French glosses that fit Portuguese contexts, explaining the low scores obtained for this task. The model learned to translate Portuguese contexts into French and generate glosses in French.

As expected, transformer-based embeddings fare better than character-based embeddings in cross-lingual learning settings. Since the character-based embeddings were not multilingual, it was thought that the embeddings would perhaps overlap in the vector space, making the model lose the nuance between the languages. More experiments are required to investigate this issue rigorously.

5.3 RQ3: Can we apply multilingual-learning techniques to definition modelling?

With the aim of answering RQ3 and sub-questions, six experiments were devised (Experiments 3.1 (PT char), 3.2 (EN char), 3.3. (PT and EN char), 3.4 (PT transformer), 3.5 (EN transformer), and 3.6 (PT and EN transformer).

These experiments showed that it is theoretically possible to generate definitions for both languages with the same model, but the scores are relatively low. A possible issue is like in cross-lingual settings, the EN dataset was much bigger than the Portuguese one. Figure 5.2 shows examples of glosses generated simultaneously. An issue with these glosses is that, in general, the length of Portuguese glosses is two or three times larger than the English ones.



```
[
 {"id": "en.defmod.881832", "gloss": "<seq> A synthetic substance that is used as a sedative and hypnotic drug. </seq>"},
 {"id": "pt.defmod.31255", "gloss": "<seq> que se refere à terra ou à margem do rio de uma sociedade, geralmente de uma sociedade, geralmente de um país ou
 {"id": "pt.defmod.110401", "gloss": "<seq> [por extensão] que se refere à pessoa que tem a capacidade de fazer com que se torne a capacidade de fazer com
 {"id": "en.dev.881895", "gloss": "<seq> the ordinal number that is the sum of one </seq>"}
]
```

Figure 5.2: *The first 10 glosses aligned.*

As for the embeddings, again, the cross-lingual transformer-based embeddings fare much better in generating the definitions than the character-based ones.

Overall, English definitions are better scored than the Portuguese ones, probably because of the size of the English dataset.

5.4 RQ4: is it possible to align Portuguese-English glosses automatically?

Surprisingly, the small experiment with transformer-based sentence embeddings showed that it is indeed possible to align the glosses by measuring the distance between the vectors with the cosine measure.

However, it is obviously clear that the alignment will depend on the quality of the generated glosses. Hence, at this point, it is safe to say that it is possible to do so, but post-editing is required.

5.5 RQ5: Are the automatically generated glosses useful for translators?

Given the enormous number of glosses generated for this project, it was impossible to evaluate all of them. Nevertheless, the small-scale study conducted with professional translators revealed that some of the glosses are indeed useful, although there was no perfect agreement regarding their quality.

In this case, it is safe to assume that these glosses require post-editing before being available to the public, especially if the audience concerns language learners.

5.6 Limitations of this thesis

Overall, the entire experimentation process was rather laborious and challenging to manage. Although the glosses proposed by the baseline model were of inferior quality, the vec2seq model can provide some insight into the quality of the embeddings and establish a baseline upon which future work can be built.

Indeed, cross-lingual and multilingual learning techniques may be beneficial but finetuning the amount of data is absolutely necessary in order to avoid illegal translations[1]. Nevertheless, monolingual settings still render the best results.

Pre-trained models do indeed provide better-quality glosses, but so far, this comes at the expense of large amounts of data and electricity. Unfortunately, given these limitations, it

[1] Illegal Translation occurs when a model translates outputs without being explictly taught or required to do so

was not possible to explore all the capabilities of the models neither with the entire English dataset nor by running for MT5 for 50 epochs, for example.

Chapter 6

Conclusions

The ultimate goal of this thesis was to generate definitions that were useful for translators in light of the Function Theory of Lexicography. To do so, this project explored monolingual, cross-lingual, and multilingual learning techniques with two transformer-based encoder-decoder models and three types of embeddings with a view to generating definitions for Portuguese and English and verifying their usefulness for translators.

Results show that pre-trained models fare better in all settings. Even though flair embeddings provide better results in monolingual settings for the baseline model, it was not possible to employ them in the experiments for the other settings. Therefore character-based and transformer-based embeddings were employed.

It was thought that transformer-based embeddings would provide better results; however, using the lighter version (and consequently not as rich as the version with more dimensions) showed that character-based embeddings can, in fact, render better results.

Finally, the human-evaluation study with translators showed that some of the glosses are, in fact, useful in a real-world assignment. However, it is clear that further and larger studies are required on how the glosses could be post-edited, for example.

Overall, the entire process of experimentation was rather difficult given the issues with the memory and the vectors. In addition to that, time constraints did not allow for fully exploring the models' capabilities, which could probably have increased the quality of the glosses with the MT5 model.

Nonetheless, this thesis paves the way for future research in all areas related to this project. First, to what extent can automatically generated glosses be incorporated into lexicographical products nowadays? For translation: would it be possible to generate domain-specific dictionaries for English and Portuguese? For definition modelling: is it possible to exploit the illegal translation capability of MT5 in favour of multilingual dictionaries? What is the best proportion of data in cross-lingual and multilingual learning in order to prevent illegal translations?

Lastly, in the consecution of the objectives of this thesis, all scripts and datasets are made available to the scientific community.

Bibliography

Adadi, A. (2021), 'A survey on data-efficient algorithms in big data era', *Journal of Big Data* **8**(1), 1–54.

Adams, O., Makarucha, A., Neubig, G., Bird, S. & Cohn, T. (2017), Cross-lingual word embeddings for low-resource language modeling, *in* 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers', pp. 937–947.

Akbik, A., Blythe, D. & Vollgraf, R. (2018), Contextual string embeddings for sequence labeling, *in* 'Proceedings of the 27th international conference on computational linguistics', pp. 1638–1649.

Alacón Martínez, R. et al. (2009), *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*, Universitat Pompeu Fabra.

Atkins, B. S. & Rundell, M. (2008), *The Oxford guide to practical lexicography*, Oxford University Press.

Banerjee, S. & Lavie, A. (2005), Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, *in* 'Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization', pp. 65–72.

Barros, L. A. (2004), *Curso básico de terminologia*, Vol. 54, Edusp.

Béjoint, H. (1981), 'The foreign student's use of monolingual english dictionaries: A study of language needs and reference skills', *Applied linguistics* **2**(3), 207–222.

Béjoint, H. (1997), 'Regards sur la définition en terminologie', *Cahiers de lexicologie* **70**(1), 19–26.

Béjoint, H. (2016), Dictionaries for general users, *in* 'The Oxford handbook of lexicography'.

Bergenholtz, H. & Tarp, S. (2002), 'Die moderne lexikographische funktionslehre. diskussionsbeitrag zu neuen und alten paradigmen, die wörterbúcher als gebrauchsgegenstände verstehen', *Lexicographica: International annual for lexicography* (18), 253–263.

Bergenholtz, H., Tarp, S. et al. (2010), 'Lsp lexicography or terminography? the lexicographer's point of view', *Specialized dictionaries for learners* pp. 27–37.

Bevilacqua, M., Maru, M. & Navigli, R. (2020), Generationary or "how we went beyond word sense inventories and learned to gloss", *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 7207–7221.

Bogaards, P. (2005), 'Dictionaries and productive tasks in a foreign language', *Kernerman Dictionary News* **13**, 20–23.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017*a*), 'Enriching word vectors with subword information', *Transactions of the association for computational linguistics* **5**, 135–146.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017*b*), 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics* **5**, 135–146.
URL: *https://aclanthology.org/Q17-1010*

Bowker, L. (2012), Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities., *in* S. Granger & M. Paquot, eds, 'Electronic Lexicography', Oxford University Press.

Bowker, L. (2017), Lexicography and terminology, *in* 'The Routledge handbook of lexicography', Routledge, pp. 138–151.

Cabré, M. T. (1999), *Terminology: Theory, methods, and applications*, Vol. 1, John Benjamins Publishing.

Cao, K. & Rei, M. (2016), 'A joint model for word embedding and word morphology', *arXiv preprint arXiv:1606.02601* .

Castells, M. (2003), *A Galáxia Internet: reflexões sobre a Internet, negócios e a sociedade*, Zahar.

Castellví, M. T. C. (1995), 'On diversity and terminology', *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* **2**(1), 1–16.

Chang, T.-Y. & Chen, Y.-N. (2019), What does this word mean? explaining contextualized embeddings with natural language definition, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', pp. 6064–6070.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014), 'Learning phrase representations using rnn encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078* .

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014), 'Empirical evaluation of gated recurrent neural networks on sequence modeling', *arXiv preprint arXiv:1412.3555* .

Cieri, C., Maxwell, M., Strassel, S. & Tracey, J. (2016), Selection criteria for low resource language programs, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 4543–4549.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020), Unsupervised cross-lingual representation learning at scale, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 8440–8451.
URL: *https://aclanthology.org/2020.acl-main.747*

Dancette, J. & Réthoré, C. (1997), 'La mémoire du commerce: aspects étymologiques et phraséologiques', *Cinquièmes Journées scientifiques du réseau Lexicologie, terminologie, traduction de l'AUPELF-UREF (Tunis, Tunisie* .

Denkowski, M. & Lavie, A. (2011), Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems, *in* 'Proceedings of the sixth workshop on statistical machine translation', pp. 85–91.

Duda, W. (1986), *Zu einer Theorie der zweisprachigen Lexikographie: Überlegungen zu einem neuen russisch-deutschen Wörterbuch*, Vol. 142, Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft.

Durán Muñoz, I. (2010), 'Specialized lexicographical resources: a survey of translators' needs', *Granger, Sylviane & Magali Paquot (eds.)* pp. 55–66.

Duvå, G. & Laursen, A.-L. (1994), 'Translation and lsp lexicography: A user survey', *Fachlexikographie. Fachwissen und seine Repräsentation in Wörterbüchern. Tübingen: Narr* pp. 247–267.

East, M. (2008), *Dictionary use in foreign language writing exams: Impact and implications*, Vol. 22, John Benjamins Publishing.

Faber, P. (2002), 'Terminographic definition and concept representation', *Training the language services provider for the new millennium* pp. 343–354.

Firth, J. R. (1957), 'A synopsis of linguistic theory, 1930-1955', *Studies in linguistic analysis* .

Fjelland, R. (2020), 'Why general artificial intelligence will not be realized', *Humanities and Social Sciences Communications* **7**(1), 10.
URL: *https://doi.org/10.1057/s41599-020-0494-4*

Fuertes-Olivera, P. A. (2013), 'The theory and practice of specialised online dictionaries for translation', *Lexicographica* **29**(1), 69–91.

Gadetsky, A., Yakubovskiy, I. & Vetrov, D. (2018), Conditional generators of words definitions, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 266–271.
URL: *https://aclanthology.org/P18-2043*

García de Quesada, M. et al. (2001), 'Estructura definicional terminográfica en el subdominio de la oncología clínica'.

Giacomini, L. (2018), Dictionaries for translation, *in* 'The Routledge Handbook of Lexicography', Routledge, pp. 284–299.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J. & Birch, A. (2021), 'Survey of low-resource machine translation', *arXiv preprint arXiv:2109.00486* .

Hanks, P. (1987), 'Definitions and explanations', *Sinclair, JM (Ed.)* **1987**, 115–136.

Hanks, P. (2000), 'Do word meanings exist?', *Computers and the Humanities* **34**(1/2), 205–215.

Hanks, P. (2008), 'The lexicographical legacy of john sinclair', *International Journal of Lexicography* **21**(3), 219–229.

Hanks, P. (2013), *Lexical analysis: Norms and exploitations*, Mit Press.

Hanks, P. (2015), 'Definition'.

Haralick, R. M. (1979), 'Statistical and structural approaches to texture', *Proceedings of the IEEE* **67**(5), 786–804.

Hartmann, R. R. K. (1999), 'Case study: The exeter university survey of dictionary use', *Thematic Network Projects, Sub-project* **9**, 36–52.

Hausmann, F. J. (2017), Kollokationen im deutschen wörterbuch. ein beitrag zur theorie des lexikographischen beispiels, *in* 'Lexikographie und Grammatik', Max Niemeyer Verlag, pp. 118–129.

Hirschman, L. & Sager, N. (1982), 'Automatic information formatting of a medical sublanguage', *Sublanguage: studies of language in restricted semantic domains* pp. 27–80.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Ishiwatari, S., Hayashi, H., Yoshinaga, N., Neubig, G., Sato, S., Toyoda, M. & Kitsuregawa, M. (2019), Learning to describe unknown phrases with local and global contexts, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', pp. 3467–3476.

Jordan, M. I. & Mitchell, T. M. (2015), 'Machine learning: Trends, perspectives, and prospects', *Science* **349**(6245), 255–260.

Jurafsky, D. & Martin, H. J. (2021), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Third Edition draft.

Kabiri, A. & Cook, P. (2020), Evaluating a multi-sense definition generation model for multiple languages, *in* 'International Conference on Text, Speech, and Dialogue', Springer, pp. 153–161.

Kilgarriff, A. (1997), 'I don't believe in word senses', *Computers and the Humanities* **31**(2), 91–113.

Kilgarriff, A. (2007), Word senses, *in* ‘Word Sense Disambiguation’, Springer, pp. 29–46.

Kittredge, R. I. (2003), Sublanguages and controlled languages, *in* ‘The Oxford Handbook of Computational Linguistics 2nd edition’.

Kudashev, I. (2007), ‘Terminography vs. lexicography opposition revisited’, *Oversattningsteori, facksprak och flersprakighet. Publikationer av VAKKI* (4), 157–166.

Kühn, P. (1989), ‘Typologie der wörterbücher nach benutzungsmöglichkeiten’, **Part 1**, 111–127.
URL: *http://dx.doi.org/10.1515/9783110095852.1.2.111*

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G. & Jyothi, P. (2019), Cross-lingual training for automatic question generation, *in* ‘Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics’, Association for Computational Linguistics, Florence, Italy, pp. 4863–4872.
URL: *https://aclanthology.org/P19-1481*

Landau, S. (n.d.), ‘1.2001. dictionaries: The art and craft of lexicography’.

Li, J., Bao, Y., Huang, S., Dai, X. & Chen, J. (2020), Explicit semantic decomposition for definition generation, *in* ‘Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics’, pp. 708–717.

Lin, C.-Y. (2004), Rouge: A package for automatic evaluation of summaries, *in* ‘Text summarization branches out’, pp. 74–81.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M. & Zettlemoyer, L. (2020), ‘Multilingual denoising pre-training for neural machine translation’, *Transactions of the Association for Computational Linguistics* **8**, 726–742.
URL: *https://aclanthology.org/2020.tacl-1.47*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), ‘Roberta: A robustly optimized bert pretraining approach’, *arXiv preprint arXiv:1907.11692* .

Maciel, A. M. B. (2001), ‘Para o reconhecimento da especificidade do termo jurídico’.

Mackintosh, K. (1998), ‘An empirical study of dictionary use in l2-l1 translation’, *Using dictionaries: Studies of dictionary use by language learners and translators* pp. 123–149.

Magueresse, A., Carles, V. & Heetderks, E. (2020), ‘Low-resource languages: A review of past work and future challenges’, *arXiv preprint arXiv:2006.07264* .

Mann, M. (2011), ‘Estructuras lexicográficas. aspectos centrales de una teoría de la forma del diccionario, herbert ernst wiegand and ma teresa fuentes morán’, *Lexikos* **21**(1), 384–390.

McEnery, T. & Hardie, A. (2011), *Corpus linguistics: Method, theory and practice*, Cambridge University Press.

McHugh, M. L. (2012), ‘Interrater reliability: the kappa statistic’, *Biochemia medica* **22**(3), 276–282.

Meier, H. H. (2003), 'Lexicography as applied', *Lexicography: Lexicography, metalexicography and reference science* **3**, 307.

Mickus, T., Bernard, T. & Paperno, D. (2020), What meaning-form correlation has to compose with: A study of MFC on artificial and natural language, *in* 'Proceedings of the 28th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 3737–3749.
URL: *https://aclanthology.org/2020.coling-main.333*

Mickus, T., Paperno, D. & Constant, M. (2019), 'Mark my word: A sequence-to-sequence approach to definition modeling', *arXiv preprint arXiv:1911.05715* .

Mickus, T., Paperno, D. & Constant, M. (2022), Semeval-2022 task 1: Codwoe – comparing dictionaries and word embeddings, *in* 'Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)'.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), 'Distributed representations of words and phrases and their compositionality', *Advances in neural information processing systems* **26**.

Müller, A. C. & Guido, S. (2017), *Introduction to machine learning with Python: a guide for data scientists*, " O'Reilly Media, Inc.".

Navigli, R. & Ponzetto, S. P. (2012), 'Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network', *Artificial Intelligence* **193**, 217–250.
URL: *https://www.sciencedirect.com/science/article/pii/S0004370212000793*

Noraset, T., Liang, C., Birnbaum, L. & Downey, D. (2016), 'Definition modeling: Learning to define word embeddings in natural language'.
URL: *https://arxiv.org/abs/1612.00394*

Noraset, T., Liang, C., Birnbaum, L. & Downey, D. (2017), Definition modeling: Learning to define word embeddings in natural language, *in* 'Thirty-First AAAI Conference on Artificial Intelligence'.

Nord, C. (2002), 'Manipulation and loyalty in functional translation', *Current writing: Text and reception in Southern Africa* **14**(2), 32–44.

Oliveira Junior, C. D. d. (2012), 'Extração automática de contextos definitórios em textos acadêmicos da ciência da informação'.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), 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.

Passonneau, R. & Mani, I. (2014), 'Definition'.

Pastor, G. C. (2001), 'Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada', *TRANS. Revista de traductología* (5), 155–184.

Pearson, J. (1998), *Terms in context*, Vol. 1, John Benjamins Publishing.

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.

Pimienta, D., Prado, D. & Blanco, Á. (2009), 'Twelve years of measuring linguistic diversity in the internet: balance and perspectives'.

Polguère, A. (2003), *Lexicologie et sémantique lexicale: notions fondamentales*, Pum.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020), 'Exploring the limits of transfer learning with a unified text-to-text transformer', *Journal of Machine Learning Research* **21**, 1–67.

Ramos, S. & del Mar, M. (2005), 'Research on dictionary use by trainee translators', *Translation journal* **9**(2).

Ranasinghe, T., Orasan, C. & Mitkov, R. (2020*a*), TransQuest at WMT2020: Sentence-level direct assessment, *in* 'Proceedings of the Fifth Conference on Machine Translation', Association for Computational Linguistics, Online, pp. 1049–1055.
URL: *https://aclanthology.org/2020.wmt-1.122*

Ranasinghe, T., Orasan, C. & Mitkov, R. (2020*b*), TransQuest: Translation quality estimation with cross-lingual transformers, *in* 'Proceedings of the 28th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 5070–5081.
URL: *https://aclanthology.org/2020.coling-main.445*

Ranasinghe, T., Orasan, C. & Mitkov, R. (2021*a*), 'An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers', *arXiv preprint arXiv:2106.00143* .

Ranasinghe, T., Orasan, C. & Mitkov, R. (2021*b*), An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)', Association for Computational Linguistics, Online, pp. 434–440.

Ranasinghe, T. & Zampieri, M. (2020), 'Multilingual offensive language identification with cross-lingual embeddings', *arXiv preprint arXiv:2010.05324* .

Reimers, N. & Gurevych, I. (2020), Making monolingual sentence embeddings multilingual using knowledge distillation, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 4512–4525.
URL: *https://aclanthology.org/2020.emnlp-main.365*

Roberts, R. P. (1992), 'Translation pedagogy: strategies for improving dictionary use', *Traduction, Terminologie et Rédaction* **5**(1), 49–76.

Robinson, R. (1972), *Definitions*, Cambridge University Press.

Rohde, D. L., Gonnerman, L. M. & Plaut, D. C. (2006), 'An improved model of semantic similarity based on lexical co-occurrence', *Communications of the ACM* **8**(627-633), 116.

Sager, J. C. (1990), *Practical course in terminology processing*, John Benjamins Publishing.

Ščerba, L. V. (1995), 'Towards a general theory of lexicography', *International Journal of Lexicography* **8**(4), 315–350.

Sierra, G., Alarcón, R., Aguilar, C. & Barrón, A. (2006), Towards the building of a corpus of definitional contexts, *in* 'Proceeding of the 12th EURALEX International Congress, Torino, Italy', pp. 229–40.

Singh, A. K. (2008), Natural language processing for less privileged languages: Where do we come from? where are we going?, *in* 'Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages'.

Souza, F., Nogueira, R. & Lotufo, R. (2020), Bertimbau: pretrained bert models for brazilian portuguese, *in* 'Brazilian Conference on Intelligent Systems', Springer, pp. 403–417.

Specia, L., Scarton, C. & Paetzold, G. H. (2018), 'Quality estimation for machine translation', *Synthesis Lectures on Human Language Technologies* **11**(1), 1–162.

Strubell, E., Ganesh, A. & McCallum, A. (2020), Energy and policy considerations for modern deep learning research, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 13693–13696.

Sun, S., Fomicheva, M., Blain, F., Chaudhary, V., El-Kishky, A., Renduchintala, A., Guzmán, F. & Specia, L. (2020), An exploratory study on multilingual quality estimation, Association for Computational Linguistics.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *Advances in neural information processing systems* **27**.

Tarp, S. (2007), 'Lexicography in the information age', *Lexikos* **17**(1), 170–179.
URL: *https://journals.co.za/doi/abs/10.10520/EJC60599*

Tarp, S. (2008), Lexicography in the borderland between knowledge and non-knowledge, *in* 'Lexicography in the Borderland between Knowledge and Non-Knowledge', Max Niemeyer Verlag.

Tarp, S. (2013), 'What should we demand from an online dictionary for specialized translation?', *Lexicographica* **29**(2013), 146–162.
URL: *https://doi.org/10.1515/lexi-2013-0010*

Tarp, S. (2018), 'Lexicography as an independent science', *The Routledge handbook of lexicography* pp. 19–33.

Tarp, S. (n.d.), 'Basic problems of learner's lexicography: on learners' dictionaries', *Lexikos* **14**(1).

Tomaszczyk, J. (1989), L1-l2 technical translation and dictionaries, *in* 'Translation and Lexicography', John Benjamins, p. 177.

Trimble, L. (1985), *English for science and technology: A discourse approach*, Cambridge University Press.

Tsvetkov, Y. (2017), 'Opportunities and challenges in working with low-resource languages', *Slides Part-1* .

Varantola, K. (1998), 'Translators and their use of dictionaries: User needs and user habits', *Using dictionaries: Studies of dictionary use by language learners and translators* pp. 179–192.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. (2020), 'Generalizing from a few examples: A survey on few-shot learning', *ACM computing surveys (csur)* **53**(3), 1–34.

Westerhout, E. (2010), *Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch*, Netherlands Graduate School of Linguistics.

Westerhout, E. & Monachesi, P. (2007), 'Extraction of dutch definitory contexts for elearning purposes', *LOT Occasional Series* **7**, 219–234.

Wiegand, H. E. (1998), *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*, Walter de Gruyter.

Wiegand, H. E. (2013), 'Lexikographie und angewandte linguistik', *Zeitschrift für angewandte Linguistik* **58**(1), 13–39.

Wright, S. E. & Budin, G. (1997), *Handbook of terminology management. Vol. 1, Basic aspects of terminology management*, John Benjamins publishing company.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. & Raffel, C. (2021), mT5: A massively multilingual pre-trained text-to-text transformer, *in* 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 483–498.
URL: *https://aclanthology.org/2021.naacl-main.41*

Yang, L., Kong, C., Chen, Y., Liu, Y., Fan, Q. & Yang, E. (2019), 'Incorporating sememes into chinese definition modeling', *arXiv preprint arXiv:1905.06512* .

Yang, L., Kong, C., Chen, Y., Liu, Y., Fan, Q. & Yang, E. (2020), 'Incorporating sememes into chinese definition modeling', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 1669–1677.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2019), 'Bertscore: Evaluating text generation with bert', *arXiv preprint arXiv:1904.09675* .

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M. & Eger, S. (2019), MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 563–578.
URL: *https://aclanthology.org/D19-1053*

Appendices

Appendix A

A Summary of the Taxonomies for Definitions

Taxonomies for definitions

Taxonomy	types of definitions	description
1. Formality-based definitions (Trimble, 1985)	Formal definitions	complete definitions that follow the *genus* + *differentiae* pattern
	Semi-formal definitions	incomplete definitions that follow *term* + *differentiae*
	Informal definitions	incomplete definitions with synonym or characteristics of the term
2. Purpose-based definitions (Robinson, 1972)	lexical definitions	dictionaries definitions
	stipulative definitions	context-based definitions
3. Lexico-syntactic, pattern-based definitions (Westerhout and Monachesi, 2007, 2010)	is definitions	*definiendum* is connected to *definiens* with the verb "be"
	punctuation definitions	punctuation marks are the connectors
	layout definitions	definitions identified through layout features of a genre
	verb definitions	*definiendum* is connected to *definiens* with a verb or verb phrase
	pronoun definitions	definitions in which pronouns make the connection or refer to a previous *definiendum* or *definiens*
	other or unclassifiable	definitions following a different pattern
4. Method-based definitions (Robinson, 1972; Borsodi, 1967; Westerhout, 2010)	ostensive definitions	exploit user encyclopaedic knowledge to define by unorthodox methods (drawing, pointing, miming)
	synonymous definitions	provide synonyms instead of a phrased definition
	analytical definitions	the traditional *genus* and *differentiae* Aristotelic pattern
	contextual definitions	place the lemma in context to explain its meaning
	reference definitions	point to secondary sources of information (quotes, historical and/or descriptive data) to define
	relational definitions	introduce meaning in terms of the relationship among words (mainly antonymy and meronymy)
	exemplifying definitions	use examples to delineate meaning
5. Semantics-based definitions (Sierra et al, 2008; Alarcón, 2009)	analytic or Aristotelic definitions	the traditional Aristotelic pattern
	synonymic definitions	exclude the *differentiae* and present synonyms of the *genus*
	functional definitions	exclude the *genus* and bring the *differentiae*
	extensional definitions	exclude the *genus* and bring a *differentia* specifying the entity's integrating parts
6. Larivière purpose-based definitions (1996)	lexicographic definitions	definitions for dictionaries that disambiguate senses and describe usage
	encyclopaedic definitions	definitions for dictionaries and encyclopaedias providing real-world knowledge
	terminological definitions	definitions for describing terms in their specialised contexts

Appendix B

License to Use BabelNet

Sapienza NLP license agreement

This is a license agreement for the resources requested below and made available for research purposes by the Sapienza NLP Group of the Sapienza University of Rome.

User information

First name:	Anna Beatriz
Last name:	Dimas Furtado
Organization:	University of Wolverhampton
Country:	United Kingdom
Email address:	[e-mail address redacted]

Description of planned use

The resource will be used for definition modelling with deep learning in Portuguese. The idea is to use the definitions to enrich the training set, since there is no dataset available for this language yet.

Requested resources

- BabelNet 5 5.0 Full version

By signing the following pages, you confirm that you have read and agree to the requested resources' licenses of use.

BabelNet 5 5.0 Full version

BabelNet is an innovative multilingual encyclopedic dictionary, with wide lexicographic and encyclopedic coverage of terms, and a semantic network/ontology which connects concepts and named entities in a very large network of semantic relations, made up of about 20 million entries.

License

1. Definitions

"Adaptation" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original.

"Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or other works or subject matter which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined above) for the purposes of this License.

"Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

"License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, Noncommercial, ShareAlike.

"Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

"Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i)in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

"Work" means BabelNet, that is, the links between heterogeneous lexical-semantic resources, each released with a separate license.

"You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

"Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

"Reproduce" means to make digital or paper copies of the Work by any means including without limitation by textual, sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage in digital form or other electronic medium.

2. Fair Dealing Rights

Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant

Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

to Reproduce the Work, to incorporate the Work into one or more Collections, and to

Reproduce the Work as incorporated in the Collections provided the Work is clearly identifiable, linked to babelnet.org and made accessible only to research institutions; to create, Reproduce, Distribute and Publicly Perform Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, an Adaptation could be marked "This data is a processed version of BabelNet [APPROPRIATE_VERSION_HERE, e.g. v4] downloaded from babelnet.org, made available with the BabelNet Non-Commercial License (see babelnet.org/license)" and made accessible only to research institutions. Alternatively, a link to the BabelNet website can be provided for download of the official data and code for the creation of the Adaptation can be provided to any user.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights described in Section 4(e).

4. Restrictions

The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) babelnet.org/license for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(d), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(d), as requested.

You may Distribute or Publicly Perform an Adaptation only under: (i) the terms of this License; (ii) a later version of this License with the same License Elements as this License. You must include a copy of, or the URI babelnet.org/license, for Applicable License with every copy of each Adaptation You Distribute or Publicly Perform. You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License. You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

You may not exercise any of the rights granted to You in Section 3 above if You are not a research institution or if in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation.

If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work (BabelNet); (iii) the URI (babelnet.org); and, (iv) consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "This data is a processed version of BabelNet [APPROPRIATE_VERSION_HERE, e.g. v4] downloaded from babelnet.org, made available with the BabelNet Non-Commercial License (see babelnet.org/license)"). The credit required by this Section 4(d) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the

Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.
For the avoidance of doubt:

Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;
Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License if Your exercise of such rights is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c) and otherwise waives the right to collect royalties through any statutory or compulsory licensing scheme; and,
Voluntary License Schemes. The Licensor reserves the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License that is for a purpose or use which is otherwise than noncommercial as permitted under Section 4(c).
Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer
UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING AND TO THE FULLEST EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO THIS EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability
EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination
This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

I hereby declare to have read, agree and to comply with the license of use of the requested resource.

Sincerely,

Anna Beatriz Dimas Furtado

[Signature redacted]

20.3.22

(Date)

(Signature)

Appendix C

Important Links

Link to Github: https://github.com/annafurtado/defmod

Appendix D

Human Evaluation - Materials

European Master's in Technology for Translation and Interpreting

Ethics proposal form for (Level Seven) Master's dissertations

You should read the University Ethics Guidance before filling in this form. This is available from: https://www.wlv.ac.uk/research/research-policies-procedures--guidelines/ethics-guidance/
[*For your information, the ACL currently adopts the ACM Code of Ethics and Professional Conduct: https://www.acm.org/code-of-ethics. This may be helpful when considering the ethical issues raised by your proposal. Another useful document for large-scale projects is: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf*].

Once completed fully, please submit this ethics proposal form to your supervisor for the project you are undertaking. Once the supervisor has approved and signed it, you must append a copy to your project placed as appendix A. Please note that this permission is only concerned with ethical issues and does not indicate anything about the intellectual merit of your project.

Please type details into the form.

1.	**Name**	Anna Beatriz Dimas Furtado
2.	**Student number**	2018321
3.	**Email address** (this must be your University email address)	[e-mail address redacted]

If this is a group project, please list ALL other students involved. (Full names and student numbers – number the new tables as 1.a., 2.a., 3.a.; 1.b.; 2.b, 3.b.; etc.):

4.	**Subject to which the study will contribute:**
Natural Language Processing, Lexicography, and Translation	

5.	**Name of supervisor(s)**	Dr. Maria Rosario Bautista Zambrana and Prof. Dr. Ruslan Mitkov

6.	**Module code and title**	7LN015 – Dissertation II

7.	**Project title:**
Generating Definitions in Portuguese and English from Comparable Corpora	

8.	**I confirm that I have:** (Tick to confirm)		
	a.	**Discussed my research with my supervisor.**	☒
	b.	**Read the Guide to Ethics and consulted the Ethics Guidance Web pages** (https://www.wlv.ac.uk/research/research-policies-procedures—guidelines/ethics-guidance/)	☒

9.	**Project research category:** (Tick to as applicable)		
	Category 0	Research that does not involve human subjects or raise any ethical concerns.	□
	Category A	Research that involves human subjects but is considered not to cause any physical or psychological harm.	☒
	Category B	Research that: • may be considered likely to cause physical or psychological harm. • may be contentious and/or risks bringing the University into disrepute. • requires accessing confidential data. • involves individuals considered to be vulnerable. **Undergraduate and Taught Masters students are not normally permitted to undertake Category B research.**	□

10.	**The project involves:** (Please tick all that apply)	
	Making video	□
	Making audio recording	□
	Observation of human subjects	□
	Opinion surveys	□
	Participant observation	□
	Telephone and/or Email contact with individuals or organisations	□
	Interviews (structured/semi-structured/unstructured) [*Delete as appropriate*]	□
	Questionnaires (including on-line questionnaires)	☒
	Access to confidential information	□
	Contact with minors (anyone under the age of 18)	□
	Contact with other vulnerable people (e.g. victims of crime, the recently bereaved, low IQ)	□
	Eye tracking	□
	Keystroke logging	□
	Creation of a new corpus from online data	☒
	Manual annotation of corpora	□
	Research requiring access to user generated content, including product reviews, forum posts, and data from social media	□
	Research about a controversial issue	□
	Research in which project outputs will be used to help provide public services or services to people beyond the EMTTI consortium	□
	Collection of personal data	□
	Other [Please specify. E.g. Will project research outputs be open access? Datasets and scripts will be open access.	☒

11. Brief outline of the project

Definitions are key for many areas of knowledge; they convey meaning, refer to product conceptualization and naming, facilitate communication, provide clarity, and pervade all areas of human activity. Hence, having access to definitions is essential for many professions, but it is crucial especially for translation and interpreting. In order to carry out their daily work, translators and interpreters rely on different supporting materials, but dictionaries, glossaries, and ontologies are best friends with these professionals.

Crafting high quality definitions, however, is a rather labour-intensive and time-consuming activity, posing a significant challenge to lexicographers and terminographers. Therefore, the automatic generation of definitions can optimise the work of lexicographers and scholars involved in this process so that final users can benefit in a timely manner.

Definition Modelling (DM) is a quite recent natural-language generation task concerned with the automatic generation of definitions from word representations (vectors) (Noraset et al., 2017). Initial approaches focused on generating monosemic, English definitions, and latest approaches concentrate on polysemic definitions. Nevertheless, investigations aiming at generating definitions from comparable corpora have never been carried out and can be key in developing bi- and multilingual resources for translators and interpreters.

Therefore, the overall goal of this project is to explore whether it is possible to automatically generate definitions for both Portuguese (a low-resourced language for DM) and English by employing deep learning algorithms. This goal is pursued by investigating the following research questions:

RQ1: How effective are deep learning techniques at generating definitions for Portuguese?

RQ2: Can we apply cross-lingual contextual word-embeddings to definition modelling for Portuguese?

a) how effective are cross-lingual word embeddings at generating definitions for Portuguese?

RQ3: Can we apply multilingual contextual word-embeddings to definition modelling?

a) how effective are multilingual contextual word-embeddings at generating definitions for Portuguese and for English?

RQ4: Can we generate bilingual definitions in Portuguese-English?

a) how effective is applying the cosine measure to map Portuguese definitions to English ones?

RQ5: Are the definitions generated automatically useful for translators?

Results will be evaluated in terms of precision, recall, F1 score, BLEU, and BertScore. Should the system produce meaningful definitions, qualitative experiments will be devised with a view to assess whether automatic generated definitions are useful for translators and whether they prove to be of good quality. For these, questionnaires will be applied, and the inter-annotator agreement will be calculated.

12. **Methodology**

The description should also indicate:

- Your objective in gathering primary data from participants.
- How participants will be identified (including sampling method if doing questionnaires).
  [*For GDPR compliance, online surveys should be conducted using either JISC Online Surveys (JISC OS) or Microsoft Forms (via MS Teams). The University of Wolverhampton has site licenses for both. All EMTTI students can access them.*]
- The number of research participants.
- A sample of questions (if conducting either interviews or surveys).
- A sample of materials used to recruit research participants.

[*Expand as necessary.*]

1. **Data Collection**

The main challenge is to gather a large dataset for Brazilian Portuguese composed exclusively of definitions, since the task has never been conducted in this language before. Hence, to do so, we exploit and scrape *Wikcionário*, the Portuguese version of Wiktionary (a crowdsourced dictionary) and *Dicio* (a lexicographer-made dictionary). For the English language, we use available datasets produced in previous investigations – the English version of Wiktionary and the Oxford definitions dataset.

In order to investigate whether cross-lingual or multilingual learning can facilitate the generation of definitions we will exploit pre-existing available datasets for romance languages: French, Spanish and Italian, as we hypothesize that combining romance languages can improve the generation of definitions for Portuguese. These datasets have already been released for the CodWoe Shared Task.

To investigate whether we can generate zero-shot (unseen) definitions for free, complex multiword expressions, we will use the HEI++ dataset (available for English) and translate it to Portuguese (to allow for comparability).

2. **Definition modelling**

In order to generate the definitions, we will use a sequence-to-sequence model fine-tuned to BERT (a bidirectional transformer-based model) to produce glosses auto-regressively. The dataset is split into training (75%) and testing (25%) sets. Experiments are conducted in three settings to answer the research questions.

a. Monolingual setting

The model is trained in each language (PT, EN) separately, the glosses are generated, then mapped together with cosine measure to produce bilingual glosses.

b. Cross-lingual setting

The model is trained with the cross-lingual dataset (FR, ES, IT, PT), then glosses are generated, and finally mapped with their counterparts in English.

c. Multilingual setting

The model is trained with a multilingual dataset (all languages), then glosses are generated for PT, and for EN. Subsequently, glosses are mapped into one another with cosine measure.

3. **Evaluation**

a. **Intrinsic evaluation** – classical metrics are employed to the quantitative aspect of this task, such as Perplexity, BLEU, Rouge, Meteor and BertScores. For the qualitative aspect, we will employ questionnaires.

b. **Extrinsic evaluation** – here we investigate how useful the definitions are in a translation assignment. For this, a questionnaire will be applied, and the inter-annotator agreement will be computed.

Experiment – Verifying the quality and usefulness of definitions
Goals: assess the ability of the model to accurately generate definitions for complex, free multiword expressions rarely found in dictionaries but potentially valuable for translators, such as *unclear action* and *fresh air*. Besides that, verify whether automatically generated definitions can be used in a translation task.

Recruitment and participants: at least three volunteer annotators with a background in Linguistics/Translation and operational proficiency in both Portuguese and English will participate in the study.

Participants will be recruited in online language/translator groups where they can freely choose to take part in the study. In order to express their interest in the experiment, participants will fill in a survey with their full name and email. Once they confirm their interest, an identifier will be assigned to each participant for ensuring their anonymity status.

Recruitment material:
Title: Assessment and usefulness of Portuguese/English definitions
Description: You are being invited to participate in an experiment to assess the quality of definitions. This integrates a dissertation project within the framework of the European Master's in Technology for Translation and Interpreting (EM TTI).

This is a two-step study. In this first part, you are asked to answer a short pre-questionnaire and translate (from English into Portuguese and vice versa) two short excerpts from two different technical texts (around 300 words), and it should not take you more than 60 minutes to finish it.

In the second part, you will be asked to assess the quality of the glosses you were given to the translation assignment following a five-level Likert scale. More details will be provided in the task form.

The deadline to finish all three parts is May 5th, 2022.

Your participation in this study is voluntary, and there is no direct benefit or risk for you. Confidential information will not be shared anywhere else. If, for any reason during the study, you do not feel comfortable, you can discontinue your participation.

Upon request, when this study is complete, you will be provided with the results of the experiment. Should you have any questions, you may send them to [e-mail address redacted]

Task:
Pre-questionnaire: Background information will be asked in this step. The goal is to profile the participants by asking their majors, gender, profession, age, and for how long they have been working in the industry.

Part 1: Participants will be given two different technical-scientific texts to translate from and into Portuguese and English. They are only allowed to use our definitions.

Part 2: Participants will assess the glosses used in their translation assignment by employing a four-level Likert scale. A gloss annotation sheet will be provided with examples, which will be drawn from the automatically generated glosses.

1) not useful - this gloss is not useful, and it does not help me understand the word.
2) requires adjustment - the gloss is understandable, but I would still prefer to consult other sources.
3) good - the gloss is good, but it misses some details.
4) fit for purpose.

13.	**Ethical Issues** Give a brief indication of the ethical issues (e.g anonymity, data protection etc) raised by your proposal and how you intend to address these issues. [*Expand as necessary.*]

1. Data Copyright

Obtaining high quality definitions may be challenging due to data copyright. Thus, we use pre-existing datasets with CC Attribution-ShareAlike licence from the CodWoe Shared task and the Oxford Dictionary Dataset offered for research by previous investigations.

Collecting definitions for Portuguese has proven to be more challenging because there is no dataset available. Hence, Wiktionary is the main source due to its CC licence, which allows for the distribution of both raw and experimental data.

For lexicographer-made definitions, we rely on Dicio, a dictionary produced by the 7Graus company. They allow the use of their content for research. However, data collected from their website will only be available in the form of experimental data in order to avoid copyright issues in the future.

2. Anonymity

For the qualitative experiments, with a view to ensuring anonymity of the participants, once they express interest on participating in the study, an identifier will be assigned to them so that they are not identified in the background questionnaire, or in the experiment phase.

14.	**Is ethical approval required by an external agency/parents of participants?** [*If yes, please provide further details*]	□

EM TTI

European Master's in Technology for Translation and Interpreting

Student Signature:	[Signature redacted]
Date:	20.3.22
Name of Supervisor 1	Dr. Maria Rosario Bautista Zambrana
Approval Signature of Supervisor 1	[Signature redacted]
Date:	25.3.22
Name of Supervisor 2	Prof. Dr. Ruslan Mitkov
Approval Signature of Supervisor 2	[Signature redacted]
Date:	21.3.22

Realizar a contratação de um serviço ou de um produto junto a um banco é fácil. Mas, em muitos casos, rescindir o contrato depois pode dar muita dor de cabeça.

Neste estudo, conclui-se que a flexibilidade pode ter tido influência no histórico lesivo do praticante e essa relação é diferente entre modalidades de resistência.

Essa interpretação apoia as interpretações de que a mecânica quântica que se alinham com os princípios científicos clássicos. “A interpretação é objetiva e realista, e ao mesmo tempo a mais simples possível. Nós gostamos de clareza e preferimos remover todo o misticismo”, diz Liukkonen.

O estudo evidencia que houve uma deterioração da condução da política fiscal a partir de 2009, e, também que a rigidez orçamentária deve ser levada em consideração ao se desenhar um regime fiscal capaz de estabilizar o produto e reduzir a vulnerabilidade do país a crises.

Antes de começar a trabalhar no comportamento canino, “eu acreditava que essa ideia era diferente da realidade”, afirma Kathleen Morrill, geneticista de cães na Faculdade Médica Chan da Universidade de Massachusetts.

Algumas pessoas receberam com perplexidade a notícia de que o ator francês Alain Delon decidiu que chegou seu momento de partir para sempre. Em outras palavras e sem minimizar ou florear o relato: ele decidiu que chegou seu momento de morrer e anunciou aos fãs sua despedida.

Em Pernambuco, região Nordeste do país, estado que congrega uma significativa população indígena, os territórios sagrados são o alvo de conflitos sangrentos entre produtores rurais, latifundiários, garimpeiros, madeireiros e os povos tradicionais, sob a displicência ingênua, colonial e, cada vez mais, permissiva do Estado brasileiro.

A relação do casal Tarsiwald com a alta-costura francesa, especialmente a *maison* Paul Poiret, indica os trânsitos entre tradição e modernidade, e as negociações, com os agentes das vanguardas artísticas, em torno do gosto conservador das elites brasileiras.

Na América Latina, o agronegócio brasileiro equivale, neste mesmo ano, a 86,57% do PIB da Argentina, segunda maior economia da região. Trata-se, portanto, de um segmento importante para a economia brasileira, justificando que mais estudos sejam feitos para melhor compreendê-lo.

Com isso, este estudo espera *i.* esboçar um panorama sobre o perfil dos artigos; *ii.* identificar como estes se distribuem entre os periódicos do estrato A1 em sociologia e entre instituições de pesquisa. *iii.* examinar tendências, sobretudo em relação às temáticas e aos assuntos nelas mais estudados; e *iv.* traçar paralelos com o campo mais amplo da sociologia brasileira.

A epidemia trouxe ao cenário o nascimento de crianças com necessidades desconhecidas e inesperadas, com marcas e resultados distintos dos padrões, inclusive das microcefalias já conhecidas e resultantes de outras doenças congênitas.

Quando as plantas crescem na direção da luz, não se estiram verticalmente, mas vão se torcendo, descrevendo um movimento de saca-rolha, na direção da luz. A circum-nutação, concluía Darwin, era uma disposição universal das plantas, antecessora de todos os outros movimentos giratórios nos vegetais.

Although great strides have been made in the development of new alloys and production processes, the industry still has unresolved issues related to untimely corrosion, which limits anode life and may lead to higher contaminant levels in the metal being produced. Lead anodes corrode because of the difference in the chemical/electrochemical potential across the microstructural features of an anode.

The effect of flotation operational parameters on froth stability and froth recovery was studied. Froth stability was measured using a special column. To determine the froth recovery, the froth height change model and froth height exponential model were used. It was found that since the interactions between the pulp and froth zones affect the time of froth formation, the exponential model is more suitable than the froth height change method for determining the froth recovery. The results showed that superficial air velocity and collector dosage have, respectively, the highest and lowest effect on the froth recovery, while froth recovery decreases sharply with increasing froth height.

This experiment demonstrated that the use of an anthelmintic strategically in the management of captive reared agoutis had no statistical effect ($p > 0.05$) on the reproductive parameters. Therefore, these animals can be kept in captive conditions without being dewormed and produce efficiently with proper feeding and housing management.

The celebration took place in the Deanery, a terribly proper oasis on the Barnard campus, filled with needlepointed chairs and Oriental rugs. The advocates of higher education for women looked down darkly from the pale walls, frozen in oil paint.

The researchers placed coils carrying electric current inside the experiment to simulate sunspots, magnetically active regions on the sun where current flows beneath the surface. On top of these coils, two copper electrodes serve as anchor points for a loop of plasma that mimics the arcs of plasma over the solar surface, or photosphere.

This article aims to define and discuss three interconnected concepts and practices, but, at the same time, irreducible: big data, datification and dataism. These are recent concepts that have been imposed due to the vertiginous increase of data that run at a unique speed on the internet. Discussions are developed to lead to more emphasis on dataism today criticized as a new form of religion. Its ambivalences, paradoxes and contradictions are highlighted, especially the paradox of the uninterrupted use of networks and the new contradictions of capitalism that are fueled by this use. The balance that oscillates between the user's desire and the harmful effects on the most recent logic of capitalism opens an abyss that this article places under reflection.

It’s the contrast between face and voice that does it. The face is round, pure, with two dimples holding her smile in place—it is the face of childhood yearning, Juliette Gréco EPs and moon-gazing through suburban windows. But the voice is something else: about half an octave lower than you expect, luxuriantly so, with unexpected notes of sanguinity and self-amusement—it is unambiguously the voice of a woman, if not fully grown, then bearing a secret apprehension of the oncoming battle between dreams and their disappointment.

Based on previous human and experimental AAA studies, several exogenous immune cells, including lymphocytes, macrophages, neutrophils, natural killer (NK) cells, and dendritic cells, have been found to infiltrate into the aneurysmal tissues, evoking a series of inflammatory reactions by releasing a wide range of pro-inflammatory cytokines that contribute to the direct structural protein degradation of the abdominal aorta

The fast economic development in recent years has resulted in a huge generation of municipal and industrial solid waste.It is necessary to dispose the waste generated in the cities in suitable places. According to the study by **Zhang**, landfills are the most common final disposal techniques for urban solid waste and are found in several locations around the world. Other

alternatives are needed for sustainable disposal of the industrial wastewater..
The meteorological conditions in 2016 showed favorable conditions for the growth and development of perennial crops. The annual amount of precipitation in 2016 exceeded the average annual normal and amounted to 559.9 mm. In 2017, the annual precipitation exceeded the average annual precipitation rate and amounted to 425.9 mm, which favorably influenced the growth and development of perennial grasses.
. In Earthsea, magic is predicated on naming, on knowing the “true name” of a person, place or thing. Her wizards attempt a sort of disclosure of the world as it is, rather than a flight from it. Le Guin has observed that “wizardry is artistry. The trilogy is, in this sense, about art, the creative experience, the creative process.”
This article investigates whether the introduction of the secret ballot in 1932 constitutes a change to the model adopted in the First Republic, where parties organized voters, mobilizing and controlling them in the act of voting. Based on an analysis of the newspapers at that time and the electoral results of the 1934 election, I show that the effect of the reform is not significant. Even though voters cast their votes in an isolated environment, ballots were printed and distributed by parties. Candidates and parties managed to bring together and mobilize voters in elections by organizing the preparation and distribution of ballots.

Dear participant,

Your Identification Code for this study is DEFMOD0101.

Please enter this code in your pre-questionnaire and post-questionnaire. This code is used to link your answers in both questionnaires to the tasks anonymously.

Your links to each of the three parts are the following:

1 — Pre-questionnaire:
https://annafurtado.typeform.com/to/u5b70rz2
Please, answer the questions accordingly.

2 — Translation Task:
Please, translate the short excerpt from English into Portuguese, and from Portuguese into English. Your final text will not be assessed, but please translate it as if it would be published.
Please, use the glossary attached to this email as your main reference. You may use other supporting tools if you wish to do so.

Important: you will be asked to upload your translation files in the post-questionnaire.

3 — Post-questionnaire:
https://annafurtado.typeform.com/to/gWq5FFD0
Please answer the questions accordingly considering the whole process and make as many comments as you would like.

Each part can be completed in different moments, just make sure to finish everything by May 10th. Feel free to email me if you have any questions as well. Thank you again for your participation!

Best wishes,

Anna 🙂

1	rescindir - [jurídico] anular o que foi alvo de cancelamento; cancelar: rescindir um contrato.	Realizar a contratação de um serviço ou de um produto junto a um banco é fácil. Mas, em muitos casos, rescindir o contrato depois pode dar muita dor de cabeça.	Source
2	lesivo - que é capaz de lesar, de prejudicar, de causar danos, prejuízos; nocivo.	Neste estudo, conclui-se que a flexibilidade pode ter tido influência no histórico lesivo do praticante e essa relação é diferente entre modalidades de resistência.	Source
3	mecânica - ciência que estuda os movimentos e os movimentos de um corpo.	Essa interpretação apoia as interpretações de que a mecânica quântica que se alinham com os princípios científicos clássicos. "A interpretação é objetiva e realista, e ao mesmo tempo a mais simples possível. Nós gostamos de clareza e preferimos remover todo o misticismo", diz Liukkonen.	Source
4	deterioração - [figurado] ação de se tornar ruim; em que há ou demonstra malvadeza; mal-estar.	O estudo evidencia que houve uma deterioração da condução da política fiscal a partir de 2009, e, também que a rigidez orçamentária deve ser levada em consideração ao se desenhar um regime fiscal capaz de estabilizar o produto e reduzir a vulnerabilidade do país a crises.	Source
5	trabalhar - realizar alguma coisa com perfeição: trabalhar a música.	Antes de começar a trabalhar no comportamento canino, "eu acreditava que essa ideia era diferente da realidade", afirma Kathleen Morrill, geneticista de cães na Faculdade Médica Chan da Universidade de Massachusetts.	Source
6	florear - [figurado] tornar elegante; enfeitar: florear a música.	Algumas pessoas receberam com perplexidade a notícia de que o ator francês Alain Delon decidiu que chegou seu momento de partir para sempre. Em outras palavras e sem minimizar ou florear o relato: ele decidiu que chegou seu momento de morrer e anunciou aos fãs sua despedida.	Source
7	displicência - falta de cuidado, de atenção; negligência, desatenção.	Em Pernambuco, região Nordeste do país, estado que congrega uma significativa população indígena, os territórios sagrados são o alvo de conflitos sangrentos entre produtores rurais, latifundiários, garimpeiros, madeireiros e os povos tradicionais, sob a displicência ingênua, colonial e, cada vez mais, permissiva do Estado brasileiro.	Source
8	gosto - o que se faz com o intuito de agradar; o que se agrada: o gosto do chefe.	A relação do casal Tarsiwald com a alta-costura francesa, especialmente a *maison* Paul Poiret, indica os trânsitos entre tradição e modernidade, e as negociações, com os agentes das vanguardas artísticas, em torno do gosto conservador das elites brasileiras.	Source

9	estudo - ação de adquirir conhecimento sobre algo ou alguém; ação de realizar alguma coisa: estudo de ciências.	Na América Latina, o agronegócio brasileiro equivale, neste mesmo ano, a 86,57% do PIB da Argentina, segunda maior economia da região. Trata-se, portanto, de um segmento importante para a economia brasileira, justificando que mais estudos sejam feitos para melhor compreendê-lo.	Source
10	traçar - [figurado] levar uma coisa a outra; levar: traçar a viagem ao parque.	Com isso, este estudo espera *i.* esboçar um panorama sobre o perfil dos artigos; *ii.* identificar como estes se distribuem entre os periódicos do estrato A1 em sociologia e entre instituições de pesquisa. *iii.* examinar tendências, sobretudo em relação às temáticas e aos assuntos nelas mais estudados; e *iv.* traçar paralelos com o campo mais amplo da sociologia brasileira.	Source
11	epidemia - [medicina] doença infecciosa e contagiosa que se espalha ou ataca com rapidez um grande número de pessoas, numa determinada região.	A epidemia trouxe ao cenário o nascimento de crianças com necessidades desconhecidas e inesperadas, com marcas e resultados distintos dos padrões, inclusive das microcefalias já conhecidas e resultantes de outras doenças congênitas.	Source
12	nutação - [botânica] propriedade que têm certas plantas de inclinar ou contrair as suas folhas e flores em certas horas do dia .	Quando as plantas crescem na direção da luz, não se estiram verticalmente, mas vão se torcendo, descrevendo um movimento de saca-rolha, na direção da luz. A circum-nutação, concluía Darwin, era uma disposição universal das plantas, antecessora de todos os outros movimentos giratórios nos vegetais.	Source
1	corrode - of metal or other materials be destroyed or damaged slowly by chemical action	Although great strides have been made in the development of new alloys and production processes, the industry still has unresolved issues related to untimely corrosion, which limits anode life and may lead to higher contaminant levels in the metal being produced. Lead anodes corrode because of the difference in the chemical/electrochemical potential across the microstructural features of an anode.	Source
2	froth - a mass of small bubbles in liquid caused by agitation fermentation or salivating.	The effect of flotation operational parameters on froth stability and froth recovery was studied. Froth stability was measured using a special column. To determine the froth recovery, the froth height change model and froth height exponential model were used. It was found that since the interactions between the pulp and froth zones affect the time of froth formation, the exponential model is more suitable than the froth height change method for determining the froth recovery. The results showed that superficial air velocity and collector dosage have, respectively, the highest and lowest effect on the froth recovery, while froth recovery decreases sharply with increasing froth height.	Source

3	agouti - a large longlegged burrowing rodent with a thick short tail and a broad flattened tail	This experiment demonstrated that the use of an anthelmintic strategically in the management of captive reared agoutis had no statistical effect ($p > 0.05$) on the reproductive parameters. Therefore, these animals can be kept in captive conditions without being dewormed and produce efficiently with proper feeding and housing management.	Source
4	deanery - the territory of an ancient roman dean	The celebration took place in the Deanery, a terribly proper oasis on the Barnard campus, filled with needlepointed chairs and Oriental rugs. The advocates of higher education for women looked down darkly from the pale walls, frozen in oil paint.	Source
5	plasma - the colourless fluid part of a room or of a planet containing the suns apparent suspension of the candelabra.	The researchers placed coils carrying electric current inside the experiment to simulate sunspots, magnetically active regions on the sun where current flows beneath the surface. On top of these coils, two copper electrodes serve as anchor points for a loop of plasma that mimics the arcs of plasma over the solar surface, or photosphere.	Source
6	abyss - a catastrophic situation seen as likely to occur.	This article aims to define and discuss three interconnected concepts and practices, but, at the same time, irreducible: big data, datification and dataism. These are recent concepts that have been imposed due to the vertiginous increase of data that run at a unique speed on the internet. Discussions are developed to lead to more emphasis on dataism today criticized as a new form of religion. Its ambivalences, paradoxes and contradictions are highlighted, especially the paradox of the uninterrupted use of networks and the new contradictions of capitalism that are fueled by this use. The balance that oscillates between the user's desire and the harmful effects on the most recent logic of capitalism opens an abyss that this article places under reflection.	Source
7	octave - synthetic organic compound which is a fossil which can be a constituent of many kinds of copper	It’s the contrast between face and voice that does it. The face is round, pure, with two dimples holding her smile in place—it is the face of childhood yearning, Juliette Gréco EPs and moon-gazing through suburban windows. But the voice is something else: about half an octave lower than you expect, luxuriantly so, with unexpected notes of sanguinity and self-amusement—it is unambiguously the voice of a woman, if not fully grown, then bearing a secret apprehension of the oncoming battle between dreams and their disappointment.	Source
8	dendritic - relating to or denoting a cell which causes progressive weakness in the cell walls of a tissue	Based on previous human and experimental AAA studies, several exogenous immune cells, including lymphocytes, macrophages, neutrophils, natural killer (NK) cells, and dendritic cells, have been found to infiltrate into the aneurysmal tissues, evoking a series of inflammatory	Source

		reactions by releasing a wide range of pro-inflammatory cytokines that contribute to the direct structural protein degradation of the abdominal aorta	
9	landfill - the disposal of waste material by burying it especially as a method of filling in and reclaiming excavated pits	The fast economic development in recent years has resulted in a huge generation of municipal and industrial solid waste.It is necessary to dispose the waste generated in the cities in suitable places. According to the study by **Zhang**, landfills are the most common final disposal techniques for urban solid waste and are found in several locations around the world. Other alternatives are needed for sustainable disposal of the industrial wastewater..	Source
10	perennial - of a plant living for several years	The meteorological conditions in 2016 showed favorable conditions for the growth and development of perennial crops. The annual amount of precipitation in 2016 exceeded the average annual normal and amounted to 559.9 mm. In 2017, the annual precipitation exceeded the average annual precipitation rate and amounted to 425.9 mm, which favorably influenced the growth and development of perennial grasses.	Source
11	wizardry - great ability in some specified field	. In Earthsea, magic is predicated on naming, on knowing the “true name” of a person, place or thing. Her wizards attempt a sort of disclosure of the world as it is, rather than a flight from it. Le Guin has observed that “wizardry is artistry. The trilogy is, in this sense, about art, the creative experience, the creative process.”	Source
12	ballot - a document listing the alternatives that is used in voting	This article investigates whether the introduction of the secret ballot in 1932 constitutes a change to the model adopted in the First Republic, where parties organized voters, mobilizing and controlling them in the act of voting. Based on an analysis of the newspapers at that time and the electoral results of the 1934 election, I show that the effect of the reform is not significant. Even though voters cast their votes in an isolated environment, ballots were printed and distributed by parties. Candidates and parties managed to bring together and mobilize voters in elections by organizing the preparation and distribution of ballots.	Source

Appendix E

Reproducibility Details

Baseline Model:

1. Learning rate: 1.0e-4
2. Epochs: 50
3. Dimensions:
    - All char embeddings: 300 dimensions
    - Transformer embeddings: 768 dimensions
    - Flair embeddings: 4096 dimensions
4. Training sets:
    - Portuguese training set was fully used as showed in the methodology session.
    - The English dataset has been capped in 300.000 items for cross-lingual and multilingual sets. After that, train/test splits have been made.
5. Most of the char-based experiments were conducted in Google Colaboratory Pro Plus.
6. Cross-lingual and multilingual learning experiments were conducted in the RGCL server with a GeForce RTX 3090 nvidia GPU.

MT5 Model:

1. Training configurations:
    - Early stopping was employed.
    - Learning rate: 0.001
2. All MT5 experiments were conducted in the RGCL server with a GeForce RTX 3090 nvidia GPU.