

LREC-COLING 2024

**The 17th Workshop on
Building and Using Comparable Corpora (BUCC)
@ LREC-COLING-2024**

Workshop Proceedings

Editors

Pierre Zweigenbaum, Reinhard Rapp and Serge Sharoff

20 May, 2024
Torino, Italy

Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024

Copyright ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL), 2024

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-31-9

ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association and the International Committee on Computational Linguistics

Preface

The 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024

In the language engineering and linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical and neural machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties. Parallel corpora are on the one end of this spectrum, unrelated corpora on the other.

Comparable corpora have been used in various applications, including Information Retrieval, Machine Translation, Cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for statistical natural language processing applications, for example, to extract parallel corpora from comparable corpora for neural machine translation. As such, it is of great interest to bring together builders and users of such corpora. The aim of the workshop series on "Building and Using Comparable Corpora" (BUCC) is to promote progress in this field.

The previous editions of the workshop took place in Africa (LREC 2008 in Marrakech), America (ACL 2011 in Portland and ACL 2017 in Vancouver), Asia (ACL-IJCNLP 2009 in Singapore, ACL-IJCNLP 2015 in Beijing, LREC 2018 in Miyazaki, Japan), Europe (LREC 2010 in Malta, ACL 2013 in Sofia, LREC 2014 in Reykjavik, LREC 2016 in Portoroz, RANLP 2019 and RANLP 2023 in Varna, LREC 2022 in Marseille) and also on the border between Asia and Europe (LREC 2012 in Istanbul). Due to the Corona crisis, the workshop was also held online in conjunction with LREC 2020 and RANLP 2021. The materials of the past workshops and related studies have also been summarised in a recent textbook from Springer:

<https://link.springer.com/book/10.1007/978-3-031-31384-4>.

We want to thank all the people who, in one way or another, helped make this workshop once again a success, especially the LREC management chairs, workshop chairs, and publication chairs.

Our special thanks go to our invited speaker, François Yvon, and to the members of the program committee, who did a great job in reviewing the submitted papers under strict time constraints. Last but not least, we would like to thank the authors and all workshop participants.

Pierre Zweigenbaum, Reinhard Rapp, Serge Sharoff

May 2024

Workshop Chairs

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Reinhard Rapp (University of Mainz and Magdeburg-Stendal University of Applied Sciences, Germany)

Serge Sharoff (University of Leeds, United Kingdom)

Program Committee

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Shervin Malmasi (Amazon, USA)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (Language Weaver, Inc., USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Ayla Rigouts Terryn (KU Leuven, Belgium)
- Reinhard Rapp (University of Mainz and Magdeburg-Stendal University of Applied Sciences, Germany)
- Nasredine Semmar (CEA LIST, Paris, France)
- Silvia Severini (Leonardo Labs, Italy)
- Serge Sharoff (University of Leeds, UK)
- Richard Sproat (OGI School of Science & Technology, USA)
- Tim Van de Cruys (KU Leuven, Belgium)
- Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Acknowledgements

The BUCC workshop was supported by ANR-20-IADJ-0005-01 under the trilateral ANR-DFG-JST AI Research project KEEPHA, *Knowledge-enhanced information extraction across languages for pharmacovigilance*, through support to the participation of Pierre Zweigenbaum and François Yvon.

Table of Contents

<i>On a Novel Application of Wasserstein-Procrustes for Unsupervised Cross-Lingual Alignment of Embeddings</i> Guillem Ramrez, Rumen Dangovski, Preslav Nakov and Marin Soljagic	1
<i>Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal</i> Julius Steuer, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach and Dietrich Klakow	12
<i>PORTULAN ExtraGLUE Datasets and Models: Kick-starting a Benchmark for the Neural Processing of Portuguese</i> Tomas Freitas Osorio, Bernardo Leite, Henrique Lopes Cardoso, Luıs Gomes, Joao Rodrigues, Rodrigo Santos and Antonio Branco	24
<i>Invited Talk: The Way Towards Massively Multilingual Language Models</i> Franois Yvon	35
<i>Exploring the Necessity of Visual Modality in Multimodal Machine Translation using Authentic Datasets</i> Zi Long, ZhenHao Tang, Xianghua Fu, Jian Chen, Shilong Hou and Jinze Lyu	36
<i>Exploring the Potential of Large Language Models in Adaptive Machine Translation for Generic Text and Subtitles</i> Abdelhadi Soudi, Mohamed Hannani, Kristof Van Laerhoven and Eleftherios Avramidis	51
<i>INCLUDE: a Dataset and Toolkit for Inclusive French Translation</i> Paul Lerner and Cyril Grouin	59
<i>BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation</i> Sourav Saha, Zeshan Ahmed Nobin, Mufassir Ahmad Chowdhury, Md. Shakirul Hasan Khan Mobin, Mohammad Ruhul Amin and Sudipta Kar	69

Posters

<i>Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity</i> Anna Laskina, Eric Gaussier and Gaelle Calvary	85
<i>EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research</i> Marc Kupietz, Piotr Banski, Nils Diewald, Beata Trawinski and Andreas Witt	94
<i>Building Annotated Parallel Corpora Using the ATIS Dataset: Two UD-style treebanks in English and Turkish</i> Neslihan Cesur, Aslı Kuzgun, Mehmet Kose and Olcay Taner Yildız	104
<i>Bootstrapping the Annotation of UD Learner Treebanks</i> Arianna Masciolini	111

<i>SweDiagnostics: A Diagnostics Natural Language Inference Dataset for Swedish</i> Felix Morger	118
<i>Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese and Turkish</i> Deniz Zeyrek, Giedrė Valūnaitė Oleškevičienė and Amalia Mendes	125
<i>mini-CIEP+ : A Shareable Parallel Corpus of Prose</i> Annemarie Verkerk and Luigi Talamo	135

Workshop Program

Monday, 20 May, 2024

9:00–10:30 Session 1

9:00–9:30 *On a Novel Application of Wasserstein-Procrustes for Unsupervised Cross-Lingual Alignment of Embeddings*

Guillem Ram3rez, Rum3n Dangovski, Preslav Nakov and Marin Solj3acic

9:30–10:00 *Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal*

Julius Steuer, Marie-Pauline Krielke, Stefan Fischer, Stefania Degaetano-Ortlieb, Marius Mosbach and Dietrich Klakow

10:00–10:30 *PORTULAN ExtraGLUE Datasets and Models: Kick-starting a Benchmark for the Neural Processing of Portuguese*

Tom3s Freitas Os3rio, Bernardo Leite, Henrique Lopes Cardoso, Lu3s Gomes, Jo3o Rodrigues, Rodrigo Santos and Ant3nio Branco

10:30–11:00 Coffee break

11:00–12:00 Invited talk

11:00–12:00 *The Way Towards Massively Multilingual Language Models*

Fran3ois Yvon

12:00–13:00 Session 2

12:00–12:30 *Quality and Quantity of Machine Translation References for Automatic Metrics*

Vil3m Zouhar and Ondřej Bojar

12:30–13:00 *Exploring the Necessity of Visual Modality in Multimodal Machine Translation using Authentic Datasets*

Zi Long, ZhenHao Tang, Xianghua Fu, Jian Chen, Shilong Hou and Jinze Lyu

13:00–14:00 Lunch break

14:00–16:00 Session 3

14:00–14:30 *Exploring the Potential of Large Language Models in Adaptive Machine Translation for Generic Text and Subtitles*

Abdelhadi Soudi, Mohamed Hannani, Kristof Van Laerhoven and Eleftherios Avramidis

14:30–15:00 *INCLURE: a Dataset and Toolkit for Inclusive French Translation*

Paul Lerner and Cyril Grouin

15:00–15:30 *BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation*

Sourav Saha, Zeshan Ahmed Nobin, Mufassir Ahmad Chowdhury, Md. Shakirul Hasan Khan Mobin, Mohammad Ruhul Amin and Sudipta Kar

Monday, 20 May, 2024 (continued)

15:30–16:00 **Booster presentations**
poster authors

16:00–16:30 **Coffee break**

16:30–18:00 **Poster session**

Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity

Anna Laskina, Eric Gaussier and Gaelle Calvary

EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research

Marc Kupietz, Piotr Banski, Nils Diewald, Beata Trawinski and Andreas Witt

Building Annotated Parallel Corpora Using the ATIS Dataset: Two UD-style treebanks in English and Turkish

Neslihan Cesur, Asli Kuzgun, Mehmet Kose and Olcay Taner Yıldız

Bootstrapping the Annotation of UD Learner Treebanks

Arianna Masciolini

SweDiagnostics: A Diagnostics Natural Language Inference Dataset for Swedish

Felix Morger

Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese and Turkish

Deniz Zeyrek, Giedrė Valūnaitė Oleškevičienė and Amalia Mendes

mini-CIEP+ : A Shareable Parallel Corpus of Prose

Annemarie Verkerk and Luigi Talamo

On a Novel Application of Wasserstein-Procrustes for Unsupervised Cross-Lingual Alignment of Embeddings

Guillem Ram rez^{*1}, Rumen Dangovski^{*1}, Preslav Nakov², Marin Solja c¹

Massachusetts Institute of Technology (MIT)¹

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)²

gramirez@ed.ac.uk

Abstract

Unsupervised word embeddings, pre-trained on vast monolingual text corpora, have driven the neural revolution in Natural Language Processing (NLP). Initially developed for English, these embeddings soon expanded to other languages, spurring efforts to align embedding spaces for cross-lingual NLP applications. Unsupervised cross-lingual alignment of embeddings (UCAE) is particularly appealing due to its minimal data requirements and competitive performance against supervised and semi-supervised approaches. In this work, we scrutinize prevalent UCAE methods and discover their objectives inherently resemble the Wasserstein-Procrustes problem. Consequently, we propose a direct solution for Wasserstein-Procrustes, enhancing popular UCAE techniques such as iterative closest point (ICP), multilingual unsupervised and supervised embeddings (MUSE), and supervised Procrustes methods. Evaluation on benchmark datasets demonstrates significant improvements over existing approaches. Our reexamination of the Wasserstein-Procrustes problem fosters further research, paving the way for more effective algorithms to align word embeddings across languages.

Keywords: Wasserstein-Procrustes, cross-lingual embeddings, unsupervised alignment

1. Introduction

Pre-trained word embeddings, which map words to dense vectors of low dimensionality, have been the key enabler of the ongoing neural revolution, and today they serve as the basic building blocks of contemporary Natural Language Processing (NLP) models. While initially introduced for English (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017; Joulin et al., 2017), pre-trained embeddings quickly emerged for a number of other languages (Heinzerling and Strube, 2018), and the idea of cross-language embedding spaces was born. In a cross-language embedding space, two semantically similar (or dissimilar) words would be close to (or far from) each other regardless of whether they are from the same or from different languages. Using such a space is attractive, as for a number of NLP tasks, it enables the application of an NLP model trained for one language on input from another language.

Ideally, such spaces could be trained on parallel bilingual datasets, but such resources are of limited size, e.g., compared to the large-scale monolingual resources typically used to pre-train monolingual word embeddings. Thus, it has been more attractive to train monolingual word embeddings for different languages independently, and then to try to align the corresponding embedding spaces in what is commonly known as bilingual lexicon induction. This has been attempted in a supervised (Mikolov et al., 2013b; Irvine and Callison-Burch, 2013), in a semi-supervised (Artetxe et al., 2017), and in an

unsupervised setting (Lample et al. (2017); Lample and Conneau (2019); Alipour et al. (2022); Feng et al. (2022); Tian et al. (2022); Liang et al. (2023); Li et al. (2023); Liu and Piccardi (2023); Ghayoomi (2023); Ghazvininejad et al. (2023)).

Initial space alignment efforts used word translation pairs as anchors, inferring transformations between languages in a supervised setup (Mikolov et al., 2013b). The alignment employs an orthogonal transformation minimizing the Frobenius norm in the Procrustes problem, with a closed-form solution obtainable via SVD. For the translation of word embeddings, W is taken to be an orthogonal matrix due to a self-similarity argument (Smith et al., 2017). The convenience of using an orthogonal matrix has also been supported empirically (Xing et al., 2015; Zhang et al., 2016; Artetxe et al., 2016). The orthogonal Procrustes problem has a closed-form solution $W = UV^T$, where $U\Sigma V^T$ is the singular value decomposition (SVD) of $X^T Y$ as shown by Sch nemann (1966).

Procrustes Given two ordered clouds of points $X, Y \in \mathbb{R}^{N \times d}$, each with N points of dimension d , the orthogonal Procrustes problem finds the orthogonal matrix $W \in \mathbb{R}^{d \times d}$ that minimizes the following Frobenius norm:

$$\arg \min_{W \in O(d)} \|XW - Y\|_2^2 \quad (1)$$

A popular unsupervised formulation of the problem is known as the Wasserstein-Procrustes (Grave et al., 2019; Alaux et al., 2019), which is more challenging as it needs to optimize a generalization

^{*}Equal contribution.

of the Procrustes objective. One-to-one maps are encouraged through a permutation matrix P .

The convenience of one-to-one maps is justified for different reasons. First, the hubness problem (Dinu and Baroni, 2014) occurs in high-dimensional vector spaces where certain vectors are the nearest neighbor to a disproportionate number of other vectors, thus reducing the quality of the embedding space (Radovanovic et al., 2010). Second, one-to-one maps can be linked to Wasserstein distance and computational optimal transport.

Wasserstein-Procrustes Given two clouds of points $X, Y \in \mathbb{R}^{N \times d}$, each with N points of dimension d , the Wasserstein-Procrustes problem finds an orthogonal matrix $W \in \mathbb{R}^{d \times d}$ and a permutation matrix $P \in \mathbb{R}^{N \times N}$ that minimize the Frobenius norm:

$$\arg \min_{P \in \pi(N), W \in O(d)} \|XW - PY\|_2^2 \quad (2)$$

where $\pi(N)$ is the set of N -dimensional permutation matrices and $O(d)$ is the set of d -dimensional orthogonal matrices.

In practice, most approaches modify the objective yet achieve good accuracy in synthetic dictionary induction tasks. We ask: Can we find approximate Wasserstein-Procrustes solutions (Equation 2) with high accuracy in dictionary tasks? Can we enhance existing methods using refinements to optimize Equation 2? Can we identify scenarios with good solutions? We address these questions by analyzing different objective functions in the literature, adhering to Artetxe et al. (2020)’s call for fair model comparison.

2. Background: Towards a Unifying Framework

There have been attempts to compare different methods proposed for the Unsupervised Cross-Lingual Alignment of Embeddings, or UCAE (Hartmann et al., 2019), and there have been papers that have tried to generalise the different possibilities one approach could possibly have. Artetxe et al. (2018a) proposed a framework based on different steps and showed how existent methods would fit in it. Ruder et al. (2019) described the most general framework for UCAE. However, we are not aware of a unified description of the existing methods from the point of view of what is being optimized, namely the loss function. We start by analyzing methods based on optimal transport methods, as they are most relevant to our approach.

2.1. Optimal Transport Methods

There have been some approaches framing the problem of unsupervised dictionary induction as an

optimal transport problem, and this is the approach we will adopt in the following sections. Haghghi et al. (2008) proposed a self-learning method for bilingual lexicon induction, representing words with orthographic and contextual features and using the Hungarian algorithm (Tomizawa, 1971) to find an optimal one-to-one matching.

With the emergence of word embeddings (Mikolov et al., 2013a), words were interpreted as vectors in high-dimensional spaces, and concepts such as distance between words started to gain attention. Ruder et al. (2018) presented Viterbi EM, where words were mapped following a one-to-one map between subsets X' and Y' of X and Y , respectively, and the isometry was induced by an orthogonal matrix. They deviated from the Wasserstein-Procrustes objective by including a penalization term for unmatched words $Y'_\perp = Y - Y'$. They did not consider all possible matches, instead imposing a restriction on the k nearest neighbors when running the Jonker-Volgenant algorithm for optimal transport (Jonker and Volgenant, 1987).

Zhang et al. (2017) proposed two different methods: WGAN (an adversarial network that optimizes the Wasserstein distance) and EMDOT (an iterative procedure that uses Procrustes and solves a linear transport problem). Both methods are inspired by the Earth Mover’s Distance (EMD), which defines a distance between probability distributions, which they applied to frequencies of words. They found that, although EMDOT could converge to bad local minima, it improved the results when used as a refinement tool after first optimizing with WGAN. Alvarez-Melis and Jaakkola (2018) used the concept of Gromov-Wasserstein distance to provide an alternative to Wasserstein-Procrustes. This distance does not operate on points but on pairs of points, turning the problem of finding optimal matching Γ^* from a linear into a quadratic one. This new loss function can be optimized efficiently with first-order methods, whereby each iteration involves solving a traditional optimal transport problem. Artetxe et al. (2018b) achieved better results by combining this idea with a refinement method called stochastic dictionary induction, i.e., randomly dropping dimensions out of the similarity matrix when extracting a seed dictionary for the next iteration of the Procrustes analysis.

2.2. Other Methods

Wasserstein-Procrustes is one of the recurring loss functions in the literature, but there have been also deviations from the original problem. Grave et al. (2019) suggested an iterative procedure whose initial condition minimizes the convex relaxation $\|X^\top PY\|_2^2$ instead of the original problem. This relaxation is known as the Gold-Rangarajan relax-

ation and can be solved using the Frank-Wolfe algorithm (Gold and Rangarajan, 1996; Frank and Wolfe, 1956). The solution to this relaxation is then used as the initial condition for a gradient-based iterative procedure that stochastically samples different subsets of words for which there is not necessarily a direct translation.

This deviates strongly from Objective 2: not only the initial condition does not optimize the Wasserstein-Procrustes objective, but also the iterative procedure does not optimize it, as it translates words that are not necessarily the optimal matches. Alaux et al. (2019) were also inspired by Objective 2 for aligning multiple languages in a common vector space. However, they minimized a loss function based on the CSLS metric from Lample et al. (2018). In a similar fashion, the entropy regularization of the Gromov-Wasserstein problem (Mémoli, 2011) has been used for bilingual lexicon induction.

Generative Adversarial Network (GAN) optimization was first introduced for bilingual lexicon induction by Barone (2016), but its canonical implementation was given by Lample et al. (2018), who presented *multilingual unsupervised and supervised embeddings* (MUSE), an adversarial method in which the transformation matrix W is considered as a generator, and thus is trained by a generative adversarial network, so that the mapped word embeddings XW cannot be distinguished from the set Y via a discriminator (Goodfellow et al., 2014). However, a simple thought experiment can convince us that this approach does not minimize distances. We elaborate on that experiment in the Appendix.

Hoshen and Wolf (2018) were inspired by the Iterative Closest Point (ICP) method used in 3D point cloud alignment. Although their transformation matrix is not necessarily orthogonal, this property is enforced using the regularization $L(X, Y, W; \lambda) := \lambda \|XW - Y\|_2^2 + \lambda \|Y - XW\|_2^2$. Another fundamental difference to Objective 2 is that they do not use a one-to-one mapping for P .

This list is not exhaustive, as there have been successful methods that do not rely on loss functions, and such that go beyond the geometry of the trained word embeddings. For example, Artetxe et al. (2019) used both the word embeddings and the monolingual corpus used to train them.

To sum up, in Table 1, we list the relevant objectives from above using our formalism from Equation 2. In the table, Γ^* is the optimal Gromov-Wasserstein matching, X' and Y' are subsets of the corresponding X and Y , Y'_\perp is the complement of Y' in Y , and \bar{Y}'_\perp is the average of the complements.

3. Properties of the Wasserstein-Procrustes Problem

We begin by simplifying Objective 2 to arrive at some essential properties, described below.

Proposition 1 (Grave et al. (2019)) *The Wasserstein-Procrustes problem is equivalent to maximizing the trace norm on the permutation matrix $X^\top PY$ over P , described as follows:*

$$\arg \min_{P \in \pi(N), W \in O(d)} \|XW - PY\|_2^2 = \arg \max_{P \in \pi(N)} \|X^\top PY\|_* \quad (3)$$

where $\|\cdot\|_*$ denotes the nuclear norm and W is selected, so that it fulfills that $U^\top WV = \mathbb{I}_d$, where both $U(P)$ and $V(P)$ are evaluated at a matrix P^* that achieves the optimum of Equation 3.

Hungarian algorithm Given two clouds of points $X, Y \in \mathbb{R}^{N \times d}$, each with N points of d dimensions, the Hungarian algorithm finds the permutation matrix P that gives the correspondence between the different points by solving the following problem:

$$\arg \min_{P \in \pi(N)} \|X - PY\|_2^2. \quad (4)$$

Replacing W in Proposition 1 with the identity matrix \mathbb{I}_d and noting that $\langle \mathbb{I}_d, X^\top PY \rangle_2 = \text{Tr}(X^\top PY)$ holds for the Frobenius inner product, we obtain the following:

Corollary 1 *Problem 4 is equivalent to maximizing the trace of $X^\top PY$ over P :*

$$\arg \min_{P \in \pi(N)} \|X - PY\|_2^2 = \arg \max_{P \in \pi(N)} \text{Tr}(X^\top PY), \quad (5)$$

which is the maximum weight matching problem. The latter can be solved using the Hungarian algorithm, which has a complexity of $O(N^3)$ (Tomizawa, 1971).

Even though the Hungarian algorithm has cubic complexity, we could still run it feasibly for $N = 45,000$. In principle, our refinement methods work well by using a subset of the full vocabulary, which typically has $N = 200,000$ words. Speedups of the Hungarian algorithm and approximations could be pursued in future work.

Equivalent problems One useful property of the trace norm is that $\|UA\|_* = \|AV\|_* = \|A\|_*$, where U and V are orthogonal matrices. Knowing this, and writing $U_X \Sigma_X V_X^\top$ and $U_Y \Sigma_Y V_Y^\top$ as the SVD decompositions for X and Y , respectively, we obtain the following:

$$\|X^\top PY\|_* = \|V_X \Sigma_X U_X^\top P U_Y \Sigma_Y V_Y^\top\|_* \quad (6)$$

Method	Objective
Grave et al. (2019) and Ours	$\min_{W \in O(d), P \in \pi(N)} \ XW - PY\ _2^2$
Alvarez-Melis and Jaakkola (2018)	$\min_{\Gamma^* \text{ best coupling}, W \in O(N)} \ X\Gamma^* - WY\ _2^2$
Hoshen and Wolf (2018)	$\min_{W \in O(d)} \ XW - Y\ _2^2 + \ YW^\top - X\ _2^2 + L(X, Y, W; \lambda)$
Ruder et al. (2018)	$\min_{W \in O(d), P' \in \pi(N')} \ X'W - P'Y'\ _2^2 + \ Y'_\perp - \bar{Y}'_\perp\ _2^2$
Lample et al. (2017)	$\min_W \max_{\theta_D} \mathbb{P}_{\theta_D}(\text{source} WX) \mathbb{P}_{\theta_D}(\text{target} Y)$
Zhang et al. (2017)	$\min_{W \in O(d), P \in \pi(N)} \sum_{i=1, j=1}^{N, N} P_{i,j} \left((X_i W)_j - Y_i \right)^2$

Table 1: Objective functions of relevant existing methods in the language of our formalism.

which yields

$$\arg \max_{P \in \pi(N)} \left\| \Sigma_X U_X^\top P U_Y \Sigma_Y \right\|_* \quad (7)$$

Let us define $\tilde{X} = U_X \Sigma_X$ and $\tilde{Y} = U_Y \Sigma_Y$. Then, the optimal solution P would be the same for translations involving all of the following pairs of word embeddings: (X, Y) , (\tilde{X}, Y) , (X, \tilde{Y}) and (\tilde{X}, \tilde{Y}) . However, the optimal transformation matrix W^* will be different for each of these problems. There is a different, yet interesting way of looking at this: if we follow the iterative procedure that starts from an initial transformation matrix $X_0 = XW_0$ (where W_0 is our initial approximation to the transformation matrix), and then we want to solve Problem (5), the equivalent problems will induce a set of *natural initializations* of the transformation W , which we formalize below:

Given the iterative procedure that tries to minimize the Wasserstein-Procrustes objective by first obtaining the permutation matrix $P_n = \arg \min_{P \in \pi(N)} \text{Tr}(X_n^\top P Y_n)$ and then the transformation matrix $W_n = \arg \min_{W \in \mathbb{R}^{N \times N}} \|X_n W - P_n Y_n\|_2^2$, the procedure aims for the same solution P as the problems with initial conditions $X_0 = XW_0$, $X_0 = X V_X W_0$, $X_0 = X W_0 V_Y^\top$, $X_0 = X V_X W_0 V_Y^\top$.

The significance of the different natural initialization is that it gives us a starting point for different problems that have the same solution P . It must be noted, however, that these transformations of X_0 are not the unique ones that will have the same original solution, as the trace norm is invariant to any orthogonal transformation; however, they help to avoid bad local minima as we will show in Section 5 below. Another way of looking at these initialization is that we are performing PCA to the embedding matrices without a dimensionality reduction. Hoshen and Wolf (2018) proposed using PCA in a similar context.

4. Approach

Below, we present a general iterative algorithm to solve the Wasserstein-Procrustes problem.

Joint optimization on W and P . For the Wasserstein-Procrustes problem from Equation 2, a joint iterative procedure involving the Procrustes problem and the Hungarian algorithm (see Algorithm 1) has been dismissed due to its computational cost and convergence to bad local minima (Zhang et al., 2017). However, as we will show below, there are a number of situations where such an approach can be extremely beneficial if we apply some improvements based on the discussion in the previous section.

Algorithm 1 *Cut Iterative Hungarian (CIH) Algorithm*

1. We initialize as follows: $X \leftarrow XW_0$.
2. We find $P \leftarrow \text{Hungarian}(X, Y)$ and $W \leftarrow \text{Procrustes}(X, PY)$.
3. If the trace norm has increased, update $X_{NEW} \leftarrow XW$ and $Y_{NEW} \leftarrow PY$, repeat Step 2.

Variants of the natural initializations. The first improvement is to consider the different equivalent problems or the natural initialization transformations, mentioned in the previous section. We observe empirically that apart from the four problems that share the same optimal P , it is possible to improve the results by considering the opposite optimization problem: instead of maximizing the costs for the two clouds of points (X, Y) , sometimes *minimizing* the costs yields a solution with a higher trace norm, and thus the algorithm eventually converges to a better solution. The matrix $X^\top PY$ is generally not symmetric with non-negative eigenvalues, and thus the trace norm and the trace are not the same. The minimization is achieved by simply considering the cloud $-X$ instead of X . Algorithm 2 is the most general iterative procedure that we consider here, and it serves as the backbone for our experiments below:

Algorithm 2 *Iterative Hungarian (IH) Algorithm. It is the same as Algorithm 1, but in Step 2 we also consider the solutions for four natural initializations: $X_0 = XW_0$, $X_0 = X V_X W_0$, $X_0 = X W_0 V_Y^\top$, $X_0 =$*

$XV_XW_0V_Y^\top$, also considering the cloud $-X$ for the four different initializations.

Supervised translation. Although the scope of this paper is the unsupervised cross-lingual alignment of embeddings, we also decided to run some experiments that involve minimal supervision. There are different ways of doing this, but the procedure that converges the fastest is to fix n pairs of words when calculating the Hungarian map, where typically $n \ll N$. We also consider similar approaches, e.g., deciding how to update Algorithm 2, taking into account the accuracy of the maps on a small subset of the data. Choosing among these methods could be motivated by how trustworthy the initial dictionary is. By *trustworthy* here we mean how many of the corresponding cloud points are correctly matched.

We use a fast implementation of the Hungarian algorithm¹ for dense matrices based on shortest path augmentation (Edmonds and Karp, 1972). Relaxations of the original problem can achieve higher speed ups. Cuturi (2013) showed how smoothing the classical optimal transport problem with an entropic regularization term results in a problem that can be solved using the Sinkhorn-Knopp’s matrix scaling algorithm (Sinkhorn and Knopp, 1967) at a speed that is orders of magnitude faster than that of transportation solvers.

Mapping. Although our method finds a permutation matrix P , this is not necessarily the best possible mapping as the set of word-to-word translations does not have to represent a one-to-one mapping. Nearest neighbor approaches can be used, but they suffer from the so-called hubness problem: in high-dimensional vector spaces, certain vectors are universal nearest neighbors (Radovanovic et al., 2010), and this is a common problem for word-embedding-based bilingual lexicon induction (Dinu and Baroni, 2014). Lample et al. (2018) presented *cross-domain similarity local scaling* (CSLS), which is a method intended to reduce the influence of hubs by expanding high-density areas and condensing low-density ones.

Given a source vector x_s , the mean similarity of its transformation Wx_s to its k target nearest neighbors $\mathcal{N}_T^k(Wx_s)$ is defined as

$$\mu_T^k(Wx_s) = \frac{1}{k} \sum_{y_t \in \mathcal{N}_T^k(Wx_s)} \cos(Wx_s, y_t).$$

Likewise is defined $\mu_S^k(y_t)$, i.e., the mean similarity of a target word y_t to its neighborhood of source mapped vectors. Then, the CSLS similarity between a mapped source vector x_s and a target vector y_t is calculated as follows: $\text{CSLS}(Wx_s, y_t) =$

$2 \cos(Wx_s, y_t) - \mu_T^k(Wx_s) - \mu_S^k(y_t)$. Intuitively, this mapping increases the similarity associated with isolated word vectors, and it decreases the one for vectors lying in dense areas. In the following experiments, we use the mapping induced by CSLS with $k = 10$.

5. Experiments

Below, we describe our experiments. In our first set of experiments, we deploy our method on top of well-known methods for cross-lingual alignment of embeddings and we show that it improves their accuracy, meaning that it can be used as a refinement tool. In the second set of experiments, we recreate the benchmarks from (Grave et al., 2019), and we show that our method can align word embedding spaces without a good initialization matrix.

5.1. The Iterative Hungarian Algorithm as a Refinement Tool

The experiments in this section use the Iterative Hungarian (IH) algorithm starting with the initial condition W_0 produced from the following methods:

- The adversarial approach by Lample et al. (2017). This combines the adversarial training described in Section 2 with a refinement step, which consists of creating a dictionary from the best matches and then running the supervised Procrustes algorithm using that dictionary.
- The supervised Procrustes approach.
- The Iterative Closest Point (ICP) method by Hoshen and Wolf (2018).

We used the word embeddings, the dictionaries and the evaluation methods from Lample et al. (2018). We trained the transformation matrix obtained from MUSE (Lample et al., 2018) on 200,000 words. Then we ran the Iterative Hungarian algorithm on a subsample of 45,000 words. Finally, we refined the new transformation matrix following the procedure in Lample et al. (2018). Also, inspired by their work, we induced mappings using CSLS with $k = 10$ nearest neighbors.

We ran the Iterative Hungarian algorithm after normalizing the word embeddings (divide them by their Euclidean norm), which we found to converge faster. It must be noted that, since the adversarial part does not normalize the word embeddings, the W_0 matrices do not match exactly and thus not normalizing them should yield better results at a higher computational cost. Hartmann et al. (2019) showed that unit-length normalization makes GAN-based methods more unstable and also deteriorates their performance, but supervised alignments or Procrustes refinement are not affected by this.

¹<http://github.com/cheind/py-lapsolver>

Method	en-es	es-en	en-fr	fr-en	en-it	it-en	en-de	de-en	en-ru	ru-en	mean
MUSE (1)	82.6	83.7	82.5	82.0	76.8	77.6	75.1	72.5	42.5	60.1	73.5
MUSE (1) + IH	82.5	84.1	82.7	82.4	78.3	77.9	74.9	73.3	44.5	60.7	74.1
MUSE (2)	81.9	83.2	82.1	82.4	77.5	77.5	74.7	72.9	37.0	61.9	73.1
MUSE (2) + IH	82.5	84.1	82.7	82.4	77.3	78.1	74.7	73.3	42.3	62.5	74.0
MUSE (3)	82.1	84.0	82.1	82.3	77.9	77.7	74.8	69.9	37.1	60.1	72.8
MUSE (3) + IH	82.3	83.9	82.6	82.4	77.8	77.8	75.1	72.9	38.9	62.1	73.6
Procrustes	81.7	83.3	82.1	81.9	77.3	77.0	73.7	72.7	49.9	60.8	74.0
Procrustes + IH	82.5	84.2	82.2	82.6	78.1	78.0	75.0	73.5	47.9	63.9	74.8
ICP (1)	81.9	82.7	81.9	81.5	76.0	75.5	72.3	72.3	46.4	56.6	72.7
ICP (1) + IH	82.5	84.1	82.1	82.7	78.1	78.0	76.6	72.7	46.2	63.2	74.6
ICP (2)	80.8	82.5	81.3	80.4	76.3	76.3	72.3	72.4	46.5	57.5	72.6
ICP (2) + IH	82.2	84.1	82.4	82.3	78.2	77.9	76.4	73.3	46.6	63.1	74.7
ICP (3)	82.0	82.6	82.0	81.8	75.7	76.6	73.1	72.6	45.1	56.2	72.8
ICP (3) + IH	82.5	84.2	82.0	82.4	77.7	77.7	76.9	73.5	45.2	63.1	74.5

Table 2: The Iterative Hungarian (IH) Algorithm starts with a transformation matrix W from MUSE, Procrustes or ICP and then refines it. The numbers 1, 2 and 3 represent runs over different seeds for non-deterministic methods (MUSE and ICP).

The results can be seen in Table 2. We can see that our Iterative Hungarian algorithm improves the accuracy when used as a refinement tool. We believe that this is because the other methods do not try to optimize the Wasserstein-Procrustes objective directly, even though they achieve very good translations without relying on it. In the Appendix we report the performance of our algorithm on more language pairs.

We also tried Zhang et al. (2019)’s Iterative Normalization: before applying IH, we subtracted the mean of the word embeddings, and we normalized them. We repeated this process three times, and then we applied IH. The results appear in Table 3: although this method improved the initialization produced by MUSE, better results were obtained by simply normalizing the word embeddings (as shown in Table 2).

5.2. Aligning Word Embeddings from the Same Data

The second set of experiments justify that the simple iterative procedure displayed in Algorithm 2 works and we explain under what circumstances it can be relaxed or needs some help in the form of either supervision or a natural initialization matrix W_0 . For the following controlled experiments, we set the initialization matrix to be the identity. We experiment with the following four approaches:

- *Hungarian*. Run the Hungarian algorithm for only one iteration, and then taking the permutation matrix P as the map.
- *Cut Iterative Hungarian (CIH)*. Run the Hun-

garian algorithm to update $Y \leftarrow PY$ and $X \leftarrow XW$ (see Algorithm 1).

- *Iterative Hungarian (IH)*. Run the previous iterative procedure but considering the different natural initializations (see Algorithm 2).
- *Supervised Iterative Hungarian (SIH)*. Learn the correct mapping from a random 5% subsample of the words, and then we run the IH algorithm for the remaining words.

The experiments from this subsection recreate those by Grave et al. (2019); the idea is that English word embeddings are trained after changing some parameters, and the different spaces of word embeddings are rotated in order to match. We use fastText (Bojanowski et al., 2017; Joulin et al., 2017) to train word embeddings on 100M English tokens from the 2007 News Crawl corpus.²

The different experiments in this section consist of changing the different training conditions and correctly mapping the results. We train the models using Skipgram (Mikolov et al., 2013c) unless stated otherwise, using the standard parameters of fastText.³ We perform four experiments:

- **Seed**. We only change the seed used to generate the word embeddings in our fastText runs. The source and the target are word embeddings trained using the same parameters.

²<http://statmt.org/wmt14/translation-task.html>

³<https://github.com/facebookresearch/fastText>

Method	en-es	es-en	en-fr	fr-en	en-it	it-en	en-ru	ru-en	mean
MUSE	81.7	83.5	82.5	81.9	77.5	77.7	45.3	61.0	73.9
MUSE + IH	82.3	84.0	82.3	82.5	77.9	77.9	44.9	61.9	74.2

Table 3: The Iterative Hungarian (IH) Algorithm starts with a transformation matrix W from MUSE, applies the iterative normalization from (Zhang et al., 2019) and then it refines the mapping.

Method	Seed	Window	Algorithm	Data
Hungar.	99%	7%	7%	1%
CIH	100%	100%	100%	0%
IH	100%	100%	100%	0%
SIH	100%	100%	100%	100%

Table 4: Our method correctly aligns the word embeddings. *Hungar.* is short for *Hungarian*.

- **Window.** We use window sizes of 2 and 10, respectively. The source and the target correspond to word embeddings trained on the same data but with different window sizes.
- **Algorithm.** We train the first algorithm with Skipgram and the second one with CBOW (Mikolov et al., 2013c). The source and the target correspond to word embeddings trained on the same data but using a different method.
- **Data.** We separate the dataset in two different parts of the same length. We train corresponding word embeddings from the two separate parts. The source and the target correspond to word embeddings trained with the same parameters but on different data.

We run the above algorithms on the 10,000 most frequent words. Table 4 shows the results for the different algorithms. We perform the final mapping using the nearest neighbor for CSLS with $k = 10$, and the reported score is the percentage of words correctly mapped. Notice, that since we are *translating English to English*, the correct map is trivial. Some observations follow:

- The supervised approach works well with very little supervision, but all other attempts failed when facing the problem of mapping data from different datasets. This is probably because, by adding some supervision, we improve the initial W_0 . This effect may be similar (although with less impact) to the help introduced in the IH algorithm with the equivalent problems or the natural initial transformations.
- The first three experiments converged in three iterations or less. The SIH algorithm took around twenty iterations to converge for the *Data* experiment.

Method	Seed	Window	Algorithm	Data
\mathbb{I}	9.49	12.59	12.45	14.11
V_X	14.13	14.14	14.18	14.19
V_Y^T	14.15	14.18	14.18	14.14
$V_X V_Y^T$	13.95	14.10	14.09	14.16

Table 5: Distance between the natural initialization and the optimal solution for the four experiments.

- The Hungarian algorithm, which was not designed for the Wasserstein-Procrustes method, correctly finds the mapping for the seed experiment, whereas some other reported iterative experiments failed to achieve good results in this experiment (Grave et al., 2019).

The proposed iterative procedures do converge, but they usually need good initial conditions or the help of supervision to converge to a good minimum. This suggests that Algorithm 1 could work well as long as we start from an initial transformation matrix W_0 close enough to the true solution. The importance of the initial condition can be shown by the natural initial conditions. The solution of the four different equivalent problems induce different optimal transformation matrices W^* . In the first iteration of the IH algorithm, a branch among these four is chosen. Table 5 shows the Euclidean distance between each of the four natural initializations (assuming $W_0 = \mathbb{I}$) and their respective optimal solution W^* for the four experiments. These distances are different for the four branches, and to choose the best one (the one minimizing this distance) is key for convergence.

The distances that are too big do not converge to a good solution. For the *Seed* experiment, such a small distance explains why a single iteration of the Hungarian algorithm was enough for a strong result. The Window and the Algorithm do not converge when running on a branch different from the first one—also the one that has the smallest distance—and when they run on the first branch, they converge in a few iterations. Hence, being able to provide a good initial transformation matrix W_0 and to correctly discriminate what the best branches are is essential for this approach.

In the Appendix we present further experiments on English to Spanish that test whether our method can be used without a good initialization, but with

little supervision. We found that our method works well when little supervision is given.

6. Conclusion and Future Work

We have underlined some mathematical properties of the Wasserstein-Procrustes problem and hence used the concept of the different natural initialization transformations in an iterative algorithm to achieve improved results for mapping word embeddings between different languages. In particular, we have shown that it is possible to use our algorithm as a refinement tool for UCAE and we have demonstrated improved results after using the transformation of [Lample et al. \(2018\)](#) as the initialization matrix W_0 . We hope that our rethinking of the Wasserstein-Procrustes problem would enable further research and would eventually help develop better algorithms for aligning word embeddings across languages, especially taking into account that most unsupervised approaches try to minimize loss functions different from Objective 2.

In future work, we plan to study other loss functions. We are further interested to see how well the objectives in Table 1 correlate with CSLS. Finally, we plan combinations with other existing methods.

7. Limitations

While our work provides valuable insights and improvements for unsupervised cross-lingual alignment of embeddings, there are some limitations to consider:

- Our analysis primarily focuses on non-contextual unsupervised word embeddings. In future work, it is essential to extend this analysis to contextualized word embeddings, which are prevalent in modern NLP applications and offer additional challenges and opportunities for alignment.
- Our study is more theoretical in nature, and the Wasserstein-Procrustes problem may not always hold true in practice due to factors such as noisy datasets or significant differences among languages. Despite these potential discrepancies, we believe our unified framework can inspire future research for improving word embeddings and contribute to more effective algorithms in aligning them across languages.

Overall, these limitations highlight potential avenues for further research and emphasize the importance of continued exploration in the field of unsupervised cross-lingual alignment of embeddings.

8. Ethics Statement

As researchers in the field of natural language processing, we recognize the importance of addressing ethical considerations in our work. In this study, we focused on unsupervised cross-lingual alignment of embeddings, with the aim of improving alignment techniques and fostering further research in this area. Below, we outline some of the ethical aspects that we have considered in this research:

- **Fairness and Bias:** We are aware that word embeddings can unintentionally capture and propagate biases present in the training data. By improving alignment techniques across languages, our work could potentially contribute to the mitigation of biases and the promotion of fairness in multilingual applications. However, we also acknowledge that our methods could inadvertently introduce or amplify biases. Future work should include thorough assessments of potential biases in the embeddings and the development of strategies to address them.
- **Accessibility:** Our research aims to advance unsupervised cross-lingual alignment methods, which can contribute to the democratization of NLP technologies by enabling their application in low-resource languages with minimal data requirements.
- **Privacy:** As our work is based on unsupervised word embeddings pretrained on large text corpora, it is crucial to ensure that the underlying data does not contain sensitive or personally identifiable information. We have made efforts to use publicly available and well-vetted datasets for our experiments and evaluations, minimizing potential privacy concerns.
- **Impact:** The advancements in unsupervised cross-lingual alignment could lead to improved performance in various multilingual NLP tasks, such as machine translation, cross-lingual information retrieval, and sentiment analysis. While these improvements can have positive effects, it is essential to consider potential misuse of such technologies and remain vigilant against unintended consequences.

Acknowledgements

This research was sponsored in part by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. This material is based upon work supported in part by the U.S. Army Research

Office through the Institute for Soldier Nanotechnologies at MIT, under Collaborative Agreement Number W911NF-18-2-0048. The work was also supported by the Technical University of Catalonia (UPC), the CFIS program and Fundació Cellex. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force, the U.S. Government, or the U.S. Army. The U.S. Government is authorized to reproduce and to distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

- Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2019. Unsupervised hyperalignment for multilingual word embeddings. In *Proceedings of the International Conference on Learning Representations*, ICLR '19, New Orleans, LA, USA.
- Ghafour Alipour, Jamshid Bagherzadeh Mohasefi, and Mohammad-Reza Feizi-Derakhshi. 2022. Learning bilingual word embedding mappings with similar words in related languages using gan. *Applied Artificial Intelligence*, 36(1):2019885.
- David Alvarez-Melis and Tommi Jaakkola. 2018. [Gromov-Wasserstein alignment of word embedding spaces](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.
- M. Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. [Learning principled bilingual mappings of word embeddings while preserving monolingual invariance](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. [Learning bilingual word embeddings with \(almost\) no bilingual data](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. [A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 5002–5007, Florence, Italy.
- Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7375–7388, Seattle, WA, USA.
- Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, RepL4NLP '16, pages 121–126, Berlin, Germany.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Marco Cuturi. 2013. Sinkhorn distances: Light-speed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, NIPS '13, pages 2292–2300.
- Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the 3rd International Conference on Learning Representations (Workshop Track)*, ICLR '14, San Diego, CA, USA.
- Jack Edmonds and Richard M. Karp. 1972. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19(2):248–264.
- Zihao Feng, Hailong Cao, Tiejun Zhao, Weixuan Wang, and Wei Peng. 2022. Cross-lingual feature extraction from monolingual corpora for low-resource unsupervised bilingual lexicon induction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5278–5287.
- Marguerite Frank and Philip Wolfe. 1956. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

- Masood Ghayoomi. 2023. Training vs post-training cross-lingual word embedding approaches: A comparative study. *International Journal of Information Science and Management (IJISM)*, 21(1):163–182.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.
- Steven Gold and Anand Rangarajan. 1996. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS '14*, pages 2672–2680, Montreal, Canada.
- Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS '2019*, pages 1880–1890, Naha, Japan.
- Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. [Learning bilingual lexicons from monolingual corpora](#). In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.
- Mareike Hartmann, Yova Kementchedjieva, and Anders Søgaard. 2019. Comparing unsupervised word translation methods step by step. In *Advances in Neural Information Processing Systems 32, NeurIPS '19*, pages 6033–6043, Vancouver, BC, CA.
- Benjamin Heizerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC '18*, Miyazaki, Japan.
- Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP '18*, pages 469–478, Brussels, Belgium.
- Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '13*, pages 518–523, Atlanta, GA, USA.
- Roy Jonker and A. Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL '17*, pages 427–431, Valencia, Spain.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32, NeurIPS '19*, pages 7059–7069.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations, ICLR '18*, Vancouver, BC, Canada.
- Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations, ICLR '18*, Vancouver, BC, Canada.
- Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. *arXiv preprint arXiv:2305.09148*.
- Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*.
- Yuzhi Liu and Massimo Piccardi. 2023. Topic-based unsupervised and supervised dictionary induction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–21.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. [Efficient estimation of word representations in vector space](#). In *Proceedings of the 1st International Conference on Learning Representations (Workshop Track), ICLR '13*, Scottsdale, AZ, USA.

- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS '13*, page 3111–3119, Red Hook, NY, USA.
- Facundo Mémoli. 2011. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP '14*, pages 1532–1543, Doha, Qatar.
- Milo Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531.
- Sebastian Ruder, Ryan Cotterell, Yova Kementchedjieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium. Association for Computational Linguistics.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.
- Peter H. Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.
- Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348.
- Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of the 2017 International Conference on Learning Representations, ICLR '17*, Toulon, France.
- Zhoujin Tian, Chaozhuo Li, Shuo Ren, Zhiqiang Zuo, Zengxuan Wen, Xinyue Hu, Xiao Han, Haizhen Huang, Denvy Deng, Qi Zhang, et al. 2022. Rapo: An adaptive ranking paradigm for bilingual lexicon induction. *arXiv preprint arXiv:2210.09926*.
- Nobuaki Tomizawa. 1971. On some techniques useful for solution of transportation network problems. *Networks*, 1:173–194.
- Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.
- Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover’s distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.
- Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan L. Boyd-Graber. 2019. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. *CoRR*, abs/1906.01622.
- Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.

Modeling Diachronic Change in English Scientific Writing over 300+ Years with Transformer-based Language Model Surprisal

Julius Steuer, Marie-Pauline Krielke, Stefan Fischer
Stefania Degaetano-Ortlieb, Marius Mosbach and Dietrich Klakow

Department of Language Science and Technology, Saarland University, Saarbrücken, Germany

{jsteuer,mmosback,dietrich.klakow}@lsv.uni-saarland.de

s.degaetano@mx.uni-saarland.de

{mariepauline.krielke,stefan.fischer}@uni-saarland.de

Abstract

This study presents an analysis of diachronic linguistic changes in English scientific writing, utilizing surprisal from transformer-based language models. Unlike traditional n-gram models, transformer-based models are potentially better at capturing nuanced linguistic changes such as long-range dependencies by considering variable context sizes. However, to create diachronically comparable language models there are several challenges with historical data, notably an exponential increase in no. of texts, tokens per text and vocabulary size over time. We address these by using a shared vocabulary and employing a robust training strategy that includes initial uniform sampling from the corpus and continuing pre-training on specific temporal segments. Our empirical analysis highlights the predictive power of surprisal from transformer-based models, particularly in analyzing complex linguistic structures like relative clauses. The models' broader contextual awareness and the inclusion of dependency length annotations contribute to a more intricate understanding of communicative efficiency. While our focus is on scientific English, our approach can be applied to other low-resource scenarios.

Keywords: Scientific English, Digital Humanities, Language Change, Evaluation, Language Modeling, Transformer

1. Introduction

Language models, particularly those rooted in machine learning and neural networks, have revolutionized the way we analyze and understand the intricacies of linguistic change (Kim et al., 2014; Hamilton et al., 2016; Dubossarsky et al., 2019). Different models, such as n-gram, LSTM or transformer models, offer diverse possibilities for analyzing language variation and change due to their underlying architectures. N-gram models are usually based on rather small and fixed-size context windows, excelling in capturing local patterns of variation. Transformer models, instead, employ attention mechanisms and deep neural networks capturing long-range dependencies and global context in language data. While they are less efficient in terms of training compared to n-gram models, they excel at capturing complex syntactic and semantic relationships, making them well-suited for analyzing possibly broader and more complex linguistic trends.

Various studies, especially concerned with lexical semantic change, already employ transformer-based models successfully (Giulianelli et al., 2020). However, comparability of the models over time is not a trivial task as the data sets often vary greatly in terms of corpus and vocabulary size, especially for historical material where the data cannot be extended.

In this paper, we apply transformer-based mod-

els to explore diachronic linguistic change in 300 years of English scientific writing. In particular, we create models of surprisal (the predictability of a word given its previous context, Shannon, 1948), which are comparable over time. Surprisal models allow us to investigate how change in language use is possibly driven by optimization effects, given that surprisal is proportional to cognitive effort (Hale, 2001; Levy, 2008). A major assumption for the evolution of scientific writing is that it becomes more informationally dense over time (Biber and Gray, 2016) due to the increasing specialization and diversification of scientific disciplines. On the other hand, conventionalization effects are at play which modulate the informational load. This balance between highly informative content and conventionalized ways of expression allows for an optimal code for expert-to-expert communication (Degaetano-Ortlieb and Teich, 2019). Our overarching aim is to create robust diachronic language models capable of capturing and quantifying optimization effects in language use. We begin by outlining previous research on changes in English scientific writing, emphasizing the use of information-theoretic notions (specifically surprisal) to capture changes related to efficiency in communication. We continue by elaborating on the challenges in using models with restricted window sizes (n-gram models) and the motivation to apply transformer-based models as well as the challenges associated with

the implementation of these models to diachronic data. Next, we describe the dataset used in our study and discuss the methods we adopt for n-modeling changes over time using transformer-based models. Working with historical linguistic data presents unique challenges, and we address some of these in detail (vocabulary shifts and train set bias). In our analysis section, we assess our transformer-based surprisal models, comparing surprisal trends with those found in earlier studies. We supplement this with a focused study on relative clauses, which require understanding long-range dependencies. These dependencies can be effectively captured by transformer-based models as they consider larger context windows.

The contributions of this study lie in addressing key modeling challenges inherent in historical data analysis, however, our approach has broader applications, extending to other areas where resources are limited.

2. Previous Work and Rationale

Diachronic change in the English scientific register has received ample attention in previous work. Earlier, descriptive (Halliday, 1988; Halliday and Martin, 1993) as well as corpus-based studies (e.g. Biber et al., 1999; Biber and Gray, 2011, 2016) report on a central mechanism in scientific language which shifts grammatical complexity from the clausal level (subordination and coordination) to the phrasal level (see also Hundt et al.) leading to an increasingly nominal instead of verbal style. Another central development in the scientific register is the conventionalization of lexico-grammatical features, which has been detected to be a necessary condition for innovation on the one hand and grammaticalization on the other (Schmid, 1994). Innovation is probably the most obvious mechanism, as a natural reaction to the need to create new vocabulary for newly arising concepts. Furthermore, diversification of certain features in increasingly distinct contexts has been observed to be at play in the course of the creation of new scientific disciplines and the formation of their respective sublanguages (Halliday, 1988; Harris Sabbetai, 1991).

While the mentioned studies are either qualitative in nature or at most frequency-based, more recent studies have employed information-theoretic measures such as n-gram-based surprisal to detect diachronic changes in the register (Degaetano-Ortlieb and Teich, 2016, 2018; Teich et al., 2021). Surprisal is formalized as the negative log probability of a unit in context $Surprisal(unit_i) = -\log_2 P(unit_i | Context)$, which results in bits of information (Shannon, 1948). The motivation to abandon a mere

frequency-based approach in favour of n-gram-based surprisal is the assumption that linguistic change underlies the rational strive for communicative efficiency. Since surprisal is a widely-used measure of information, which has been shown to be correlated with cognitive effort in online language processing (e.g. Levy, 2008; Demberg and Keller, 2008) it is well suited to giving a communicative explanation for changes in the lexical as well as grammatical level. Degaetano-Ortlieb and Teich (2019), for instance, show that in scientific writing certain grammatical patterns become less surprising over time, i.e. increasingly conventionalized in their contexts, while specific lexical items show a trend toward “innovation and increase in expressivity” (Degaetano-Ortlieb and Teich, 2019, p. 26) indicated by phases of high surprisal when new concepts enter the language and phases of stability/consolidation when items of lexical usage become conventionalized in their contexts. An example of the interplay between lexical innovation and grammatical consolidation is the noun–preposition–noun pattern (e.g. *oxide of iron*) becoming extremely predictable as a grammatical pattern while serving as a “habitual host for lexical innovation and terminology formation” (Degaetano-Ortlieb and Teich (2019, p. 26). The mentioned studies give a plausible explanation for the underlying motives of language change, however, their underlying 4-gram language models (i.e. surprisal based on a word’s predictability given its previous three words as context) are fairly restricted in terms of context size. While the narrow context of only three preceding words is well suited for detecting optimization of shorter linguistic units such as the above-mentioned noun-preposition-noun pattern, it is less well suited for drawing conclusions about the diachronic development of linguistic structures exceeding this window size. A possible solution to this is to replace the n-gram with a model covering a larger context window such as an LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Radford et al., 2018), which is the approach we present in the present paper. However, applying such models, especially transformer-based ones, poses several challenges (e.g., varying corpus and vocabulary sizes or selection of data for modeling). In this paper, we work towards addressing some of these challenges (cf. Section 3.2). While there is a growing body of research on which language model architecture to choose if restricted to a limited compute budget (e.g. Scaboro et al., 2021) or a specific dataset size (e.g. Hoffmann et al., 2022), it is less clear what approach one should take if either one’s compute budget or dataset size is very small. This becomes especially pressing in the case of historical corpora as there are only limited

ways of extending the data. While it is possible to train large language models on historical English data (e.g. Hosseini et al., 2021), doing so might be undesirable for a number of reasons. When the reason for training a language model is to create a computational model which serves as an approximation – ideally a cognitively plausible one – of a speaker of a specific time period, the option of training the language model on large-scale text data is not available, since the training data should, for example, be restricted to a time period *preceding* any text from that time period, with the basic assumption that a speaker did not have access to future text productions. This is of course a simplified assumption; in the case of written text, it is plausible to assume that every reader has access to some amount of data that lie in the future from the perspective of any given text. However, there are domains in which this amount can plausibly be assumed to be small, such as scientific English writing.

3. Data and Methods

3.1. The Royal Society Corpus

We use the Royal Society Corpus (RSC) (Fischer et al., 2020; Kermes et al., 2016) as a data set. The RSC is based on the *Philosophical Transactions* and the *Proceedings* of the Royal Society of London. In total, it comprises 295 895 749 tokens and 47 837 texts, which were published between 1665 and 1996. The RSC incorporates a comprehensive set of metadata such as text categories (e.g., articles, abstracts), authorship, title, publication date, and historical periods (ranging from decades to half-centuries), along with linguistic annotations at multiple layers including tokens (featuring to some extent both normalized and original forms), lemmas, and parts of speech, utilizing TreeTagger (Schmid, 1994), and Universal Dependency parsing achieved by Stanza (Qi et al., 2020) with combined models. Given that the texts underwent OCR, several preliminary procedures were employed to counteract OCR inaccuracies to the greatest extent feasible (for an in-depth explanation, refer to Kermes et al., 2016; Menzel et al., 2021).

Even though the RSC is large enough for language modeling, the distribution of texts and tokens poses a challenge. Figures 1 and 2 show that texts and tokens are not equally balanced across time. This can be attributed to an increase in publication activity as well as significantly longer texts in recent time periods (see Figure 3).

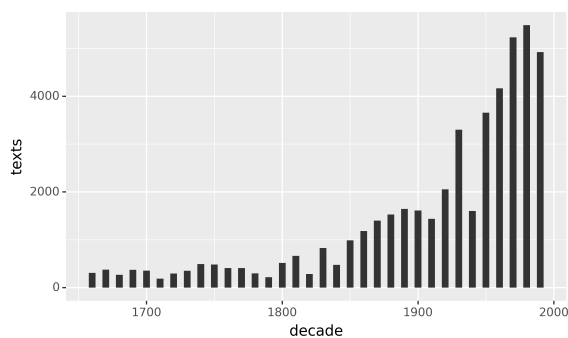


Figure 1: Distribution of texts in the RSC.

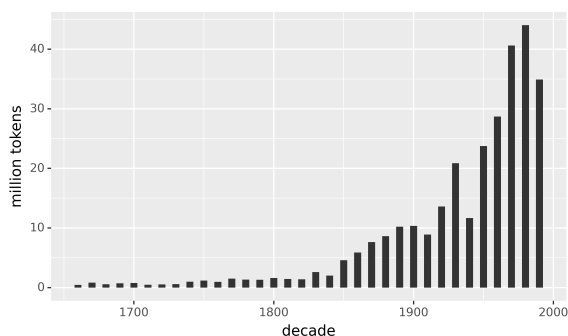


Figure 2: Distribution of tokens in the RSC.

3.2. Modeling Diachronic Change with Transformer-Based Models

Problem Statement Previous research on the diachronic linguistic development of English scientific writing (Degaetano-Ortlieb et al., 2018; Degaetano-Ortlieb and Teich, 2018, 2019) has employed surprisal derived from n-gram language models as a proxy of a model’s linguistic knowledge. The hypothesis is that as syntactic structures are conventionalized (i.e. become more predictable) over time, surprisal from n-gram language models fitted to texts from successive time slices of the RSC decreases. While previous work indeed found such an effect (Krielke, 2021), n-gram language models are only a good approximation for local effects and ideally, a more cognitively plausible model should have access to the full sentence-level context.

A possible solution is to replace the n-gram by a large language model with a larger context window (LSTM or Transformer). However, for historical data such as the RSC, this is far from trivial as the number of texts and the number of tokens per year decrease exponentially as we go back in time (see Figure 2). In such a setting, it becomes increasingly difficult to train and compare language models for two reasons:

1. **Vocabulary Shift.** When sampling from peri-



Figure 3: Average number of tokens per document in the RSC.

ods t and $t+1$, the sets of word types or vocabularies V^t, V^{t+1} will be partially disjoint, i.e. $V_t \cap V^{t+1} \neq \emptyset$. When training language models M^t, M^{t+1} on t and $t+1$, the probability distributions $P_{M^t}, P_{M^{t+1}}$ cannot be directly compared since they are defined over different sets of events.

2. **Train Set Bias.** Let C^t be the set of texts from period t . Since $|C^t| \ll |C^{t+1}|$, M^{t+1} will see much more training data than M^t when naively sampling from the corpus. The probability estimates derived from M^{t+1} will be tighter than those derived from M^t as a function of the train set size, if the vocabulary stays constant i.e. for an identical prefix $w_{0\dots i-1}$ we expect that $P_{M^{t+1}}(w_i|w_{0\dots i-1}) \geq P_{M^t}(w_i|w_{0\dots i-1})$.

Approach

Continuous Pre-training While vocabulary shifts can be addressed by sharing a unified vocabulary over all models, train set bias requires sampling the train set such that M^t and M^{t+1} are trained on a similar number of tokens, which is problematic because for earlier periods we may have only very little data. In order to alleviate the effects of train set bias, we make use of the default NLP pipeline of pretraining a transformer model on a more general dataset D_{PT}^0 sampled uniformly from the each time period C^t , and then continue pre-training on the documents of a specific year C^t . In our experiments we use the smallest version decoder-only OPT architecture (Zhang et al., 2022), with randomly initialized weights.

Pretraining Dataset We sample a pretraining dataset D_{PT}^0 from all documents in the corpus such that an equal number of tokens is sampled from the documents C^t . Sampling 10^5 tokens

yields $|D_{PT}^0| \approx 2 \times 10^6$. We derive a unified vocabulary by training a BPE tokenizer with $|V| = 5 \times 10^4$ on D_{PT} . We then pre-train on D_{PT}^0 , obtaining a set of pre-trained weights θ_{PT} . Surprisal for words that are split into subwords by the tokenizer is calculated by summing their respective log probabilities.

Pre-training on Individual Years We sample datasets D_{PT}^t for each year t in the corpus. Each D_{PT}^t consists of the documents from a period of k years prior to t such that starting from $t_0 = t - k$:

$$D_{PT}^t = \bigcup_{t'=t_0}^{t-1} C^{t'} \quad (1)$$

We then initialize the model with θ_{PT} and fine-tune on D_{PT}^t until validation loss converges. Hyperparameters for continuous pre-training can be found in Table 1. This results in a similar number of training steps (300-400) on each D_{PT}^t , independent of $|D_{PT}^t|$.

Hyperparam	Value
Batch Size	128
Learning Rate	1^{-3}
Warmup	Linear, 10%
Optimizer	AdamW

Table 1: Pre-training hyperparameters

Implementation We used HuggingFace Transformers (Wolf et al., 2020) to train the BPE tokenizer and to pretrain and fine-tune the OPT model. The source code will be made available on GitHub alongside instructions to replicate the result upon publication.

4. Analyzing Linguistic Change

One main assumption regarding the development of scientific English is a balance between informationally dense scientific content and conventionalized scientific style (cf. Degaetano-Ortlieb and Teich, 2019). In fact, it has been shown that scientific English changes from a verbal involved style with embedded clausal structure (see Example (1)) toward a heavy nominal style with long nominal phrases (see Example (2)) and predictable grammatical structures (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2019).

- (1) *And if the greatest part of these Vessels are Arteries, or other Vessels, **that immediately receive liquors from them; I may prove, I think, from another Experiment, made by Injection into a part of the Arteria praeparans, before I began***

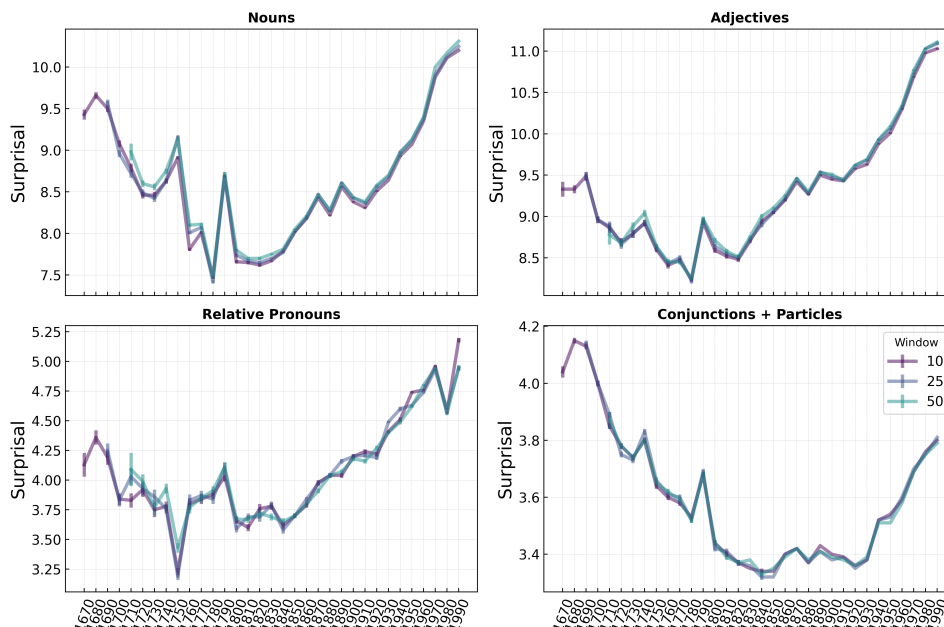


Figure 4: Surprisal on nouns, adjectives, relative pronouns ('which'+ 'that' with XPOS tag 'WDT') and conjunctions/particles as defined by the UPOS annotations of the RSC. Error bars show standard error.

to expand the Body of the Testis; **whereupon** opening the part, **which I saw** discoloured, **I found**, that many of these Tubes **had received** some of the fine particles of that matter, **which I tinged** my injected Spirit with. (King and de Grieff, 1669)

- (2) *On the other hand, a clear red-green stripe pattern of predominantly positive or negative response emerges in the vertical motion signal.* (Zanker, 1996)

We start by testing whether the results obtained by transformer-based surprisal models are in line with previous findings. We employ the Mann-Kendall trend test (in the Python implementation by Hussain and Mahmud (2019)) to confirm visually salient increases or decreases of surprisal (either on specific words or averaged over POS tags). In particular, we report the direction of the trend (increasing, decreasing, or no trend), its slope s and the p-value p associated with it. Figure 4 shows surprisal of the lexical word classes nouns and adjectives as well as the grammatical function words relative pronouns (e.g., *which*, *that*) and conjunctions/ particles (e.g., *and*, *to*). The surprisal values are calculated as the mean surprisal of all words belonging to a class per decade, determined by the word's POS tag in the CoNLLU. From the 1800s onward we can see an increase in surprisal for word classes associated with nominal style, nouns ($s = 0.0181$, $p < 0.001$) and adjectives ($s = 0.174$, $p < 0.001$). Function words, instead, show a steady decrease during the 17th up to the

1840s ($s = -0.01$, $p < 0.001$) followed by a slight but not significant increase from the 1930s onward ($s = 0.0093$, $p = 0.2097$). Thus, while nouns and adjectives in general carry more information (higher surprisal, between 7.5 - 12 bits) on average, function words carry much less information (between 3-5 bits). While these findings are in line with the general trend observed in previous work (cf. Kermes and Teich, 2017), we should state that considering average lexical surprisal of words belonging to specific word classes only shows a very aggregated picture as a result of the confluence of different factors, e.g. word frequency, vocabulary diversity, collocational behaviours of words per decade.

4.1. Modeling Convention and Innovation with Surprisal

A more thorough inspection at the lexical level (nouns and adjectives in Figure 4) seems to indicate a wave-like tendency with periods of alternating peaks (e.g., 1680, 1750, 1790, 1890, 1990) and troughs (e.g., 1730, 1780, 1820, 1910) in surprisal. Considering that peaks in surprisal indicate an increase in the use of unpredictable words given their previous context and troughs stand for more predictable usages, the observed changes could be related to discoveries triggering new vocabulary in the corpus at hand, especially at points with abrupt changes such as in the 1790s for nouns. In fact, this peak coincides with the chemical revolution where the 100-year-old phlogiston theory was replaced by the ev-

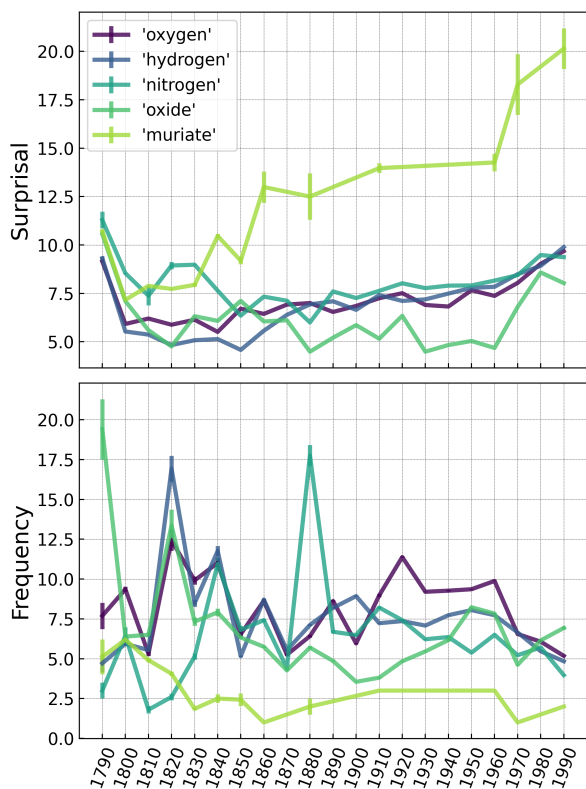


Figure 5: Average surprisal and frequency of nouns contributing to the surprisal increase and related to the chemical revolution. Error bars show standard error over documents.

idence around the discovery of oxygen and hydrogen. Troughs, on the other hand, reflect periods of consolidation, where new vocabulary is integrated into language use possibly becoming established terminology (cf. [Degaetano-Ortlieb and Teich, 2019](#)). To test this assumption, we further inspect the nouns contributing most to the surprisal increase in the 1790s (see Figure 5 showing surprisal and frequency). At their first mentioning in the 1790s, these nouns are relatively high in surprisal, but strongly decrease in the decade 1800, when their frequency increases, stabilizing at a mid-surprisal range in the coming decades (1800-1840). This is clearly an indication of a point in time (1790s) of innovation in terms of the use of new lexemes followed by a period of conventionalization, where new terminology was established in the new chemistry field. Thereafter, the chemical elements oxygen ($s = 0.031$, $p < 0.001$), hydrogen ($s = 0.0424$, $p < 0.001$), and nitrogen ($s = 0.0188$, $p < 0.05$) show a continuous increase in surprisal, with a clear peak from the 1970s to the 1990s. This tendency seems to indicate two distinct mechanisms that might have an impact on the nouns' surprisal: (1) given that the frequency is not decreasing until the 1960s, the nouns might be used

in more diverse contexts which would explain their increase in surprisal (e.g., *thought that/permeable to/the ketonic oxygen* with high surprisal of oxygen >10 bits), (2) in the period of the 1970s to the 1990s, their frequency decreases, which might explain their even stronger increase in surprisal in that later period.

4.2. Modeling Surprisal for Long-Range Dependencies

In this second analysis, we focus on relative clauses (RCs), which inherently involve long dependencies (see Example (3)), necessitating models that can appropriately handle such complexities. In this regard, transformer-based models are arguably more effective than n-gram or LSTM models as they can make use of very long contexts, offering deeper, context-sensitive analysis that is crucial for accurately capturing the nuanced aspects of these linguistic structures.

- (3) *The **protein**, for which the detailed biochemical pathway analysis conducted by the researchers identified several novel interaction partners, **exhibits** properties consistent with increased metabolic resistance.*

4.2.1. Surprisal of Relativizer

We start by considering the surprisal of the relative pronouns in Figure 4, which show a clear turn in 1920, while conjunctions and particles seem to stabilize in surprisal for more than 100 years between 1800 and 1920. Interestingly for relativizers, the development until the 19th century is not exactly in line with previous research on the diachronic development of relativizer informativity (cf. [Krielke, 2021](#)), which reports on an overall slightly increasing surprisal of relativizers (*which* and *that*). [Krielke \(2021\)](#) explains the upward trend with the frequency decrease of relativizers in scientific writing. However, they report on an increasing conventionalization of *which* as the preferred relativizer in scientific writing occurring in increasingly uniform contexts accounting for some very low surprisal values. An explanation for the differences in our results might be the different modeling of surprisal since [Krielke \(2021\)](#) uses a 4-gram surprisal model based on probability estimates of relativizers within 50-year periods. As our estimates are based on a dynamic context size and models are more precise in terms of surprisal prediction in different time slices due to the balanced vocabulary size of each slice, we can assume that our results reflect changes in relativizer informativity more reliably. Moreover, the increase in surprisal in the 20th century in our data may have two mutually non-exclusive explanations: (a) rela-

tivizers become less frequent in the 20th century, and (b) they occur in increasingly diverse contexts as a reaction to specialization, e.g., they might follow a higher diversity of nouns, which are also less predictable (cf. Figure 4).

4.2.2. Relativizers in Context

We furthermore inspect more thoroughly the reasons for the increasing surprisal values of relativizers in the 20th century by examining a) the frequency distributions of relativizers over time (Figure 6) as well as the immediate lexical contexts of relativizers (*that* and *which*, Table 2).

The frequencies show that the overall strongest decrease in relativizers over time happens roughly in the 19th century. This shows that the sudden increase in surprisal of relativizers in the 20th century does not seem to be motivated by a major frequency drop but rather by a specialization of contexts that relativizers tend to occur in. The most frequent part-of-speech 3-grams preceding the relativizers *that* and *which* reflect this: *that* in the decade 1990 mostly occurs after complex noun phrases (Table 2), which can be assumed to decrease the predictability of the relativizer. The preceding contexts of *which* are more predictive since they introduce either prepositional RCs (e.g., *the way in which*) or restrictive RCs separated from the matrix clause with a comma (e.g. *the experiment, which*). What is interesting here is the fact that the more predictable *which* steadily drops in frequency while *that* steadily increases from 1900 onward. This could explain the overall increase in surprisal since the strongly conventionalized *which* becomes less influential in the average surprisal while *that* becomes less predictable in context and more frequent.

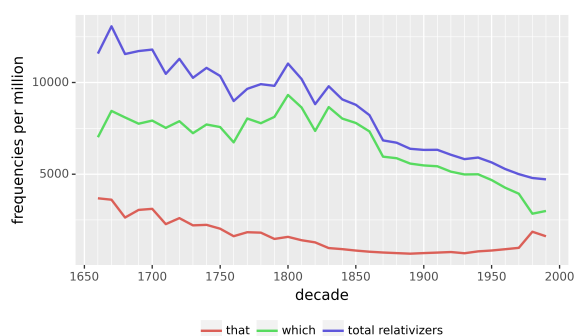


Figure 6: Distribution of relativizers in the RSC.

While this kind of pattern-based context analysis would also be possible using 4-grams, our surprisal model should also account for larger contexts and better surprisal estimates compared to the 50-year-based 4-gram surprisal model used in

previous studies allowing more reliable interpretations of diachronic trends.

freq.	trigram	example
3034	IN DT NN	of the acid <i>that</i>
2893	DT JJ NN	the muriatic acid <i>that</i>
1627	NN IN NNS	number of experiments <i>that</i>
1473	IN JJ NNS	in various instances <i>that</i>
1415	DT NN NN	the iron particles <i>that</i>
6392	DT NN IN	the way in <i>which</i>
3123	JJ NN ,	unique solution, <i>which</i>
2866	JJ NN IN	special case in <i>which</i>
2818	NN NN ,	length scale <i>which</i>
2722	DT JJ NN	the complex plane <i>which</i>
1767	(CD)	(3.1) <i>which</i>

Table 2: POS trigrams preceding *that* and *which*

4.2.3. Surprisal in Long-Range Dependencies

Here we ask whether syntactic context has an influence on the predictability of syntactic elements relying on longer distances to their syntactic heads. As a plausible measure to define syntactic context, we make use of the well-established metric *dependency length* (DL) describing the distance from an element X to its syntactic head (Hinger et al., 1980; Hudson, 1995). We apply this metric to measure the distance between the head noun of an RC and its embedded verb to find out whether the distance correlates with the predictability of the embedded verb. One assumption could be that the closer the relevant information (head noun) is located to the upcoming dependent (embedded verb) the lower the surprisal of the upcoming word (see Example (4)).

- (4) a. *The woman who ate the sandwich was hungry.* (DL = 2)
 b. *The woman whom the manager of operations wanted to talk to was upset.* (DL = 10)

We start by inspecting the diachronic development of DL and the surprisal of embedded verbs in RCs by decades (see Figure 7). DL and surprisal are negatively correlated, i.e. the opposite of our assumption is the case: in decades where the embedded verb on average is further away from its syntactic head the verb's average surprisal is lower, and in decades where the verb is closer to its syntactic head the verb is more surprising (compare Examples (5-a) and (5-b)).

- (5) a. *The questions that we might ask are not easy to answer.* (DL = 4)
 b. *The questions that lie on the table are not easy to answer.* (DL = 2)

One explanation is that at the point of encountering

the relativizer the entropy (i.e. uncertainty about the rest of the sentence, Hale, 2006) is higher than at a later point in the relative clause.

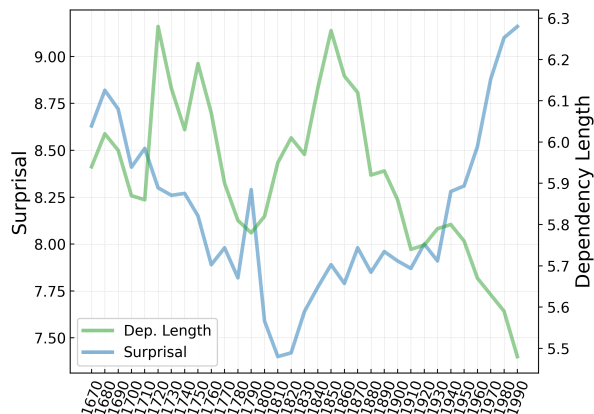


Figure 7: Average surprisal on the verb in the RC and dependency length to its head noun in the RSC. Negative correlation for surprisal and dependency length (Spearman’s $\rho = -0.35$, $p < 0.05$).

To get a better intuition about the reasons for the lower predictability in shorter syntactic contexts and the higher predictability in longer syntactic contexts, we extract the most surprising contexts of embedded verbs in RCs in 1990 plus the least surprising contexts in 1820. Example (6) reflects two aspects we have mentioned so far. First, the conventionalized pattern of prepositional RCs (i.e. *determiner noun preposition*) contexts not only seem to increase the predictability of the upcoming relativizer but also that of the embedded verb. Second, surprisal at the main verb (participle) in the RC might be reduced due to being in a highly predictable passive construction (*has been + participle*).

- (6) *the first **case** in which a quantitative attempt has been **made*** (*Surprisal = 0.0142*) Conversely, the high-surprisal contexts are extremely short where the head noun is directly followed by the relativizer and the embedded verb (Example (7)).
- (7) *recordings in the region of the pontine **nuclei** that **VERB*** (*Surprisal = 9.9998*) The latter, high-surprisal cases syntactically belong to the subject RC type, while the first, low surprisal cases belong to the oblique relative clause type. The negative correlation between surprisal and DL can thus also be explained with Hale’s Entropy Reduction Hypothesis (ERH, Hale, 2006) – uncertainty about the rest of the sentence tends to decrease as new words are introduced, and the degree of this reduction aligns with the information that the word

conveys within the context of the current sentence (cf. Frank, 2013, p. 476). Thus, a high surprisal value at the verb of a subject RC directly following the relativizer is equivalent to a strong reduction of uncertainty about the rest of the sentence since at this point the relativized grammatical relation can be resolved. The low surprisal at the embedded verbs of RCs extracted from other positions (e.g., oblique) implies that entropy reduction here is much lower since a lot of information for disambiguation about the rest of the sentence has been given before.

For comparison, we consider the 1820s where DL is comparatively long, while surprisal is fairly low. Example (8-a) shows a particularly long dependency relation between the head noun (*bundles*) and the embedded verb (*composed*) with extremely low surprisal.

- (8) a. *denote the **bundles** of fibres of which the brain is **composed*** (*Surprisal = 0.1132*)

Since the immediate left context (i.e. *the brain is*) of *composed* is not particularly predictive, while the head noun *bundles* is much more so, our surprisal estimates seem to rely on a context beyond the immediately preceding 3-gram. Thus, our modeling approach allows us to capture long-range syntactic relations. Together with the fact that our results align with observations from psycho-linguistic studies (e.g. Frank, 2013), we conclude that our methodology for generating diachronically comparable surprisal estimates provides plausible metrics to investigate the interplay between information content and syntactic context in diachronic language change.

5. Conclusion

We demonstrated the efficacy of transformer-based surprisal models in analyzing diachronic linguistic change, highlighting their capacity to accommodate long-range dependencies and global context. The application of these models to historical data is not without challenges given the exponential increase in texts and tokens over time, leading to partially disjoint vocabularies and non-comparable probability distributions across different periods. Significant disparities in training set sizes between time periods further complicate the modeling process. To address this, we implement two key strategies: (1) sharing a unified vocabulary over all models, (2) pre-training on a more general dataset sampled uniformly from the whole corpus and then continue pre-training on documents of a specific year. Our models also uniquely incorpo-

rate a temporal aspect, restricting training data to texts published before the target period (which can be adapted for other research questions).

Our empirical analysis, compared against prior studies using n-gram models on the Royal Society Corpus, revealed both corroborative and novel insights. Specifically, the examination of linguistic phenomena, such as relative clauses, underscored the superiority of transformer-based models in predicting changes not solely based on changes related to frequency distributions. These models, with their broader contextual awareness, facilitated a more nuanced exploration of communicative efficiency, aligning with theoretical frameworks like Hale's Entropy Reduction Hypothesis. The inclusion of DL annotations further enriched our analysis, allowing for a more granular examination of syntactic structures. This provided deeper insights into the adaptive mechanisms of language, reflecting shifts in complexity and efficiency within scientific discourse.

While our research focused on diachronic scientific English, the methodologies employed are universally applicable, especially in low-resource environments facing similar challenges with vocabulary consistency and corpus size disparities. This universality significantly broadens the potential impact of our findings, suggesting that our approaches could be instrumental in diverse linguistic and historical analyses.

6. Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074 – SFB 1102.

7. Bibliographical References

- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- Rie Kubota Ando and Tong Zhang. 2005. [A framework for learning predictive structures from multiple tasks and unlabeled data](#). *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. [Scalable training of \$L_1\$ -regularized log-linear models](#). In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Douglas Biber and Bethany Gray. 2011. [Grammatical change in the noun phrase: the influence of written language use](#). 15(2):223–250.
- Douglas Biber and Bethany Gray. 2016. [Grammatical Complexity in Academic English: Linguistic Change in Writing](#). Studies in English Language. Cambridge University Press.
- Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Longman.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](#). *Journal of the Association for Computing Machinery*, 28(1):114–133.
- James W. Cooley and John W. Tukey. 1965. [An algorithm for the machine calculation of complex Fourier series](#). *Mathematics of Computation*, 19(90):297–301.
- Stefania Degaetano-Ortlieb. 2021. [The rise of compounds as informationally dense structures in 20th-century scientific English: Chapter 11. measuring informativity](#). In Elena Seoane and Douglas Biber, editors, *Corpus-based Approaches to Register Variation*, Studies in Corpus Linguistics, pages 291–312. John Benjamins Publishing Company.
- Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2018. [An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English](#). Brill. Pages: 258-281 Section: From Data to Evidence in English Language Research.
- Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based modeling of diachronic linguistic change: from typicality to productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173.
- Stefania Degaetano-Ortlieb and Elke Teich. 2018. [Using relative entropy for detection and analysis of periods of diachronic linguistic change](#). In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.
- Stefania Degaetano-Ortlieb and Elke Teich. 2019. [Toward an optimal code for communication: The case of scientific English](#). *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.
- Vera Demberg and Frank Keller. 2008. [Data from eye-tracking corpora as evidence for theories of syntactic processing complexity](#). 109(2):193–210.

- Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. [Time-out: Temporal referencing for robust modeling of lexical semantic change](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.
- Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. [The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802, Marseille, France. European Language Resources Association.
- Stefan L. Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. [Large-scale evidence of dependency length minimization in 37 languages](#). 112(33):10336–10341. Publisher: National Academy of Sciences.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing lexical semantic change with contextualised word representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 446–457. Association for Computational Linguistics.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL '01*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- John Hale. 2006. [Uncertainty about the rest of the sentence](#). *Cognitive Science*, 30(4):643–672.
- M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter. Tex.date-added: 2010-03-09 11:19:11 +0100 tex.date-modified: 2010-03-30 15:16:26 +0200.
- M.A.K. Halliday and James R. Martin. 1993. *Writing science: Literacy and discursive power*. Falmer Press.
- William L. Hamilton, J. Leskovec, and Dan Jurafsky. 2016. [Cultural shift or linguistic drift? comparing two computational measures of semantic change](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Zellig Harris Sabbettai. 1991. *A Theory of Language and Information: A Mathematical Approach*. Oxford University Press, Oxford, New York.
- Hans Jürgen Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. Fink.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Computation*, 9(8):1735–1780.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. [An empirical analysis of compute-optimal large language model training](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.
- Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. [Neural Language Models for Nineteenth-Century English](#). ArXiv:2105.11321 [cs].
- Richard Hudson. 1995. Measuring syntactic difficulty. *Manuscript, University College, London*.
- Marianne Hundt, David Denison, and Gerold Schneider. [Relative complexity in scientific discourse](#). 16(2):209–240.
- Md. Hussain and Ishtiak Mahmud. 2019. [pymannkendall: a python package for non-parametric Mann Kendall family of trend tests](#). *Journal of Open Source Software*, 4(39):1556.
- Tom S Juzek, Marie-Pauline Krielke, and Elke Teich. 2020. [Exploring diachronic syntactic shifts with dependency length: the case of scientific English](#). In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, Barcelona, Spain (Online). Association for Computational Linguistics.
- Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich.

2016. [The Royal Society Corpus: From uncharted data to corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931, Portorož, Slovenia. European Language Resources Association (ELRA).
- Hannah Kermes and Elke Teich. 2017. [Average Surprisal of Parts-of-\(s\)peech](#). Birmingham, UK. Corpus Linguistics 2017.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [Temporal analysis of language through neural language models](#). In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.
- Reinhard Kneser and Hermann Ney. 1995. [Improved backing-off for M-gram language modeling](#). In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA. IEEE.
- Marie-Pauline Krielke. 2021. [Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German](#). *Bergen Language and Linguistics Studies*, 11(1):91–120.
- Marie-Pauline Krielke, Luigi Talamo, Mahmoud Fawzi, and Jörg Knappen. 2022. [Tracing syntactic change in the scientific genre: Two Universal Dependency-parsed diachronic corpora of scientific English and German](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4808–4816, Marseille, France. European Language Resources Association.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. [Generating Linguistically Relevant Metadata for the Royal Society Corpus](#).
- Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. [An Analysis of Neural Language Modeling at Multiple Scales](#). Publisher: arXiv Version Number: 1.
- Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukas Burget, and Jan Honza Cernocky. 2011. [RNNLM - Recurrent Neural Network Language Modeling Toolkit](#). In *IEEE Automatic Speech Recognition and Understanding Workshop*. Edition: IEEE Automatic Speech Recognition and Understanding Workshop.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#).
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. [Yara parser: A fast and accurate dependency parser](#). *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Simone Scabro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2021. [NADE: A benchmark for robust adverse drug events extraction in face of negations](#). In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 230–237, Online. Association for Computational Linguistics.
- Hans-Jörg Schmid. 2015. [A blueprint of the Entrenchment-and-Conventionalization model](#). 3(1):3–26. Publisher: De Gruyter Mouton.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.
- Claude E. Shannon. 1948. [A Mathematical Theory of Communication](#). *Bell System Technical Journal*, 27(3):379–423.
- Andreas Stolcke. 2002. [SRILM - an extensible language modeling toolkit](#). In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904. ISCA.
- Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).
- Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. [Less is more/more diverse: On the communicative utility of linguistic conventionalization](#). 5:142.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick

von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Huggingface's transformers: State-of-the-art natural language processing](#).

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#).

8. Language Resource References

Fischer, Stefan and Knappen, Jörg and Menzel, Katrin and Teich, Elke. 2020. *The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study*. European Language Resources Association. [\[link\]](#).

Kermes, Hannah and Degaetano-Ortlieb, Stefania and Khamis, Ashraf and Knappen, Jörg and Teich, Elke. 2016. *The Royal Society Corpus: From Uncharted Data to Corpus*. European Language Resources Association (ELRA). [\[link\]](#).

PORTULAN ExtraGLUE Datasets and Models: Kick-starting a Benchmark for the Neural Processing of Portuguese

Tomás Freitas Osório¹, Bernardo Leite¹, Henrique Lopes Cardoso¹,
Luís Gomes², João Rodrigues², Rodrigo Santos², António Branco²

¹Artificial Intelligence and Computer Science Laboratory (LIACC),
Faculdade de Engenharia da Universidade do Porto
Rua Doutor Roberto Frias, s/n, 4200-465 Porto, Portugal
tomas.s.osorio@gmail.com, {bernardo.leite, hlc}@fe.up.pt

²University of Lisbon
NLX—Natural Language and Speech Group, Dept. Informatics
Faculdade de Ciências (FCUL), Campo Grande, 1749-016 Lisboa, Portugal
{lmdgomes, jarodrigues, rdsantos, antonio.branco}@fc.ul.pt

Abstract

Leveraging research on the neural modelling of Portuguese, we contribute a collection of datasets for an array of language processing tasks and a corresponding collection of fine-tuned neural language models on these downstream tasks. To align with mainstream benchmarks in the literature, originally developed in English, and to kick start their Portuguese counterparts, the datasets were machine-translated from English with a state-of-the-art translation engine. The resulting PORTULAN ExtraGLUE benchmark is a basis for research on Portuguese whose improvement can be pursued in future work. Similarly, the respective fine-tuned neural language models, developed with a low-rank adaptation approach, are made available as baselines that can stimulate future work on the neural processing of Portuguese. All datasets and models have been developed and are made available for two variants of Portuguese: European and Brazilian.

Keywords: Machine translation, Portuguese, Benchmark, LoRA

1. Introduction

Neural language models are pervasive in Natural Language Processing (NLP) applications and have radically changed the state-of-the-art since the Transformer architecture (Vaswani et al., 2017) was proposed. This has given rise to encoder (Devlin et al., 2019), decoder (Radford et al., 2018), and encoder-decoder architectures (Raffel et al., 2020). To support the development of such models, several benchmarks have been created to assess their performance in several downstream tasks (Wang et al., 2018, 2019). However, most research in NLP has focused on the English language (Bender, 2011), and as a consequence, many other languages lack sufficient resources – in particular, benchmarks for neural language models.

Developing benchmark datasets is hard, usually demanding labeling by experts, especially for complex semantic-level tasks. An alternative path that has been resorted to in the literature is to rely on state-of-the-art Machine Translation (MT) to produce dependable datasets, namely those that support the evaluation of neural models in downstream tasks (Conneau et al., 2018; Eger et al., 2018; Yang et al., 2019; Carrino et al., 2020; d’Hoffschmidt et al., 2020; Shavrina et al., 2020; Carvalho et al., 2021; Sousa et al., 2021; Žagar

and Robnik-Šikonja, 2022). Though possibly imperfect, such datasets can fit the purpose of greatly leveraging research in less-resourced languages, possibly complemented with human-curated test sets.

In this paper, we contribute to enriching the set of benchmarks publicly available for Portuguese by relying on MT applied to tasks from the well-known GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, which were originally developed for English. We discuss the issues encountered with our approach and provide versions of several tasks for European (pt-PT) and Brazilian (pt-BR) Portuguese, which altogether we named PORTULAN ExtraGLUE.

As a way of their practical validation, for most tasks, we include experimental evaluation of different Portuguese language models fine-tuned with the respective datasets. Hence, for many of them, these will be the first models to address that task in Portuguese, and we thus contribute the first baselines for them. To that end, we resort to the encoder Albertina language model (Rodrigues et al., 2023) and the low-rank adaptation approach (Hu et al., 2022). The resulting fine-tuned language models for these tasks are openly distributed as open source under an open license.

2. Related Work

Producing benchmarks to evaluate language models in downstream tasks is a daunting endeavor. The more complex the task, the more difficult it is to produce quality data that can be used to train models in a fine-tuning approach and test their capabilities. While highly resourced languages, such as English, include quite elaborate benchmarks (Wang et al., 2018, 2019), few evaluation datasets are available for other, less-resourced languages.¹ The particular case of Portuguese is a paradigmatic example, with only a few tasks being available for this purpose (Fonseca et al., 2016; Real et al., 2020; Santos et al., 2006; Freitas et al., 2010).

A few examples of manually produced multilingual parallel corpora exist (Yang et al., 2019; Artetxe et al., 2020b; Ponti et al., 2020; Sen et al., 2022), as well as collections of tasks in multiple languages (Srivastava et al., 2023). At the same time, machine translation has come to a point in which it can be useful to create corpora that, while lacking human curation, can, up to a certain extent, be used to evaluate language models in the target languages (Conneau et al., 2018; Eger et al., 2018; Yang et al., 2019; Carrino et al., 2020; d’Hoffschmidt et al., 2020). Some have been created to allow cross-lingual evaluation of pre-trained encoders (Hu et al., 2020; Liang et al., 2020).

State-of-the-art MT systems still struggle to produce accurate translations in several situations. Short texts, for instance, often lack enough context to obtain proper translations (Wan et al., 2022). Because of this, translation at the sentence level often falls short of translating longer texts, which provide more context (Jin et al., 2023). Translating from mostly gender-poor to gender-rich languages is also often a source of translation errors (Savoldi et al., 2021). Idioms are among the most intricate artifacts for MT systems, which tend to over-generate compositional and literal translations (Dankers et al., 2022). Additionally, translation-based data can arguably be seen as a dialect of the target language (Volansky et al., 2013; Artetxe et al., 2020a), with the possible effect of over-estimating the performance in the target language of models trained on such data. Still, MT has progressed notably over the last few years; it can, we believe, be used to produce datasets that are useful as a proxy in assessing the comparative merits of different (monolingual) language models.

Following this trend, some works have leveraged MT to produce corpora in Portuguese (Carvalho et al., 2021; Sousa et al., 2021). We leverage state-of-the-art MT in producing Portuguese variants of

¹For instance, treebank annotations (Nivre et al., 2020) are available, but do not comprise benchmarks *per se*.

several GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) tasks. Similar efforts have been made for other languages (Shavrina et al., 2020; Žagar and Robnik-Šikonja, 2022).

In tandem with developing and making these datasets available, and as a way of their practical validation, we also release low-ranked adaptations (Hu et al., 2022) of Albertina-based models (Rodrigues et al., 2023), arguably the best open encoder models for both European and Brazilian Portuguese available at the time of this writing.

Low-ranked adaptations (LoRA) reduce the number of training parameters, alleviating storage requirements for language models adapted to specific tasks while outperforming other fine-tuning techniques. For that, pre-trained model weights are frozen, and two additional weight matrices are used to adapt the model to the downstream task. After training, such weights can be merged with the frozen weights so that no latency is added at inference time, which is a main advantage compared to other low-rank adapters (Houlsby et al., 2019; Mahabadi et al., 2021; He et al., 2022). Concerning LoRA, more recent proposals (Valipour et al., 2023; Audibert et al., 2023) rely on the GLUE benchmark (Wang et al., 2018) to report improvements.

3. General Language Understanding Evaluation Benchmarks

The General Language Understanding Evaluation (GLUE) tasks are meant to measure the progress toward general-purpose language understanding technologies for English. Both GLUE and SuperGLUE are aggregations of existing public datasets accompanied by a single-number performance metric and an analysis toolkit. The tasks included in these benchmarks can be grouped as follows².

3.1. Single sentence tasks

The Corpus of Linguistic Acceptability (CoLA)^G (Warstadt et al., 2019) is a task including sentences annotated for grammatical acceptability by experts in linguistics. The Stanford Sentiment Treebank (SST-2)^G (Socher et al., 2013), in turn, is a task for predicting the sentiment polarity of movie reviews.

3.2. Similarity tasks

The Microsoft Research Paraphrase Corpus (MRPC)^G (Dolan and Brockett, 2005) is a task for determining whether a pair of sentences are mutual paraphrases. Quora Question Pairs (QQP)^{G,3} is

²We superscript each task regarding its inclusion in (G)LUE, (S)uperGLUE, or both.

³<https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

a task for determining whether a pair of questions are semantically equivalent. The Semantic Textual Similarity Benchmark (**STS-B**)^G (Cer et al., 2017) is a task for predicting a similarity score (from 1 to 5) for each sentence pair. Word-in-Context (**WIC**)^S (Pilehvar and Camacho-Collados, 2019) comprises a word sense disambiguation task, where given two sentences containing a polysemous target word, the aim is to determine whether the word is used in the same sense in both sentences.

3.3. Inference tasks

The Multi-Genre Natural Language Inference Corpus (**MNLI**)^G (Williams et al., 2018) is a task to determine if a given premise sentence entails, contradicts, or is neutral to a hypothesis sentence; the task includes matched (in-domain) and mismatched (cross-domain) validation and test sets. Question NLI (**QNLI**)^G (Rajpurkar et al., 2016) is a question-answering task converted to determine whether the context sentence contains the answer to the question. Recognizing Textual Entailment (**RTE**)^{GS} is a task for determining whether a premise sentence entails a hypothesis sentence. Winograd Natural Language Inference (**WNLI**)^G (Levesque et al., 2012) is a pronoun resolution task formulated as sentence pair entailment classification where, in the second sentence, the pronoun is replaced by a possible referent. Similarly, the Winograd Schema Challenge (**WSC**)^S is a co-reference resolution task also formulated as sentence pair entailment classification, where each example comprises a sentence and a pair pronoun-noun, the objective being to determine if they are co-referent. CommitmentBank (**CB**)^S (de Marneffe et al., 2019) comprises short texts with embedded clauses; one such clause is extracted as a hypothesis and should be classified as neutral, entailment or contradiction.

GLUE and SuperGLUE also include expert-constructed diagnostic datasets covering diverse linguistic phenomena. Broadcoverage Diagnostics (**AX_b**)^{GS} (Wang et al., 2018) is a Natural Language Inference (NLI) task designed to test models across a wide spectrum of linguistic, commonsense, and world knowledge; each instance contains a sentence pair labeled with entailment or not entailment. Winogender Schema Diagnostics (**AX_g**)^S (Rudinger et al., 2018) is a similar task, designed to measure gender bias, where each premise sentence includes a male or female pronoun and a hypothesis includes a possible referent for the pronoun.

3.4. Question-answering tasks

Boolean Questions (**BoolQ**)^S (Clark et al., 2019) is a question-answering task where yes/no questions are given for short text passages. In the Multi-Sentence Reading Comprehension (**MultiRC**)^S

task (Khashabi et al., 2018), given a context paragraph, a question, and an answer, the goal is to determine whether the answer is true; for the same context and question, more than one answer may be correct. In the Reading Comprehension with Commonsense Reasoning Dataset (**ReCoRD**)^S, each sample is a multiple-choice question including a news article passage and a Cloze-style question with one entity masked out; the aim is to predict the masked entity from a list of alternatives.

3.5. Reasoning tasks

Choice of Plausible Alternatives (**COPA**)^S (Gordon et al., 2012) is a casual reasoning task: given a premise, two choices, and a cause/effect prompt, the system must choose one of the choices.

4. PORTULAN ExtraGLUE

Creating a Portuguese version of the tasks introduced in the previous section via machine translation (MT) requires a thoughtful understanding of the nature of each task, together with the limitations of the selected MT engine. While we are aware that, for a small subset of these tasks, Portuguese-translated versions have already been created (Rodrigues et al., 2023), such considerations have not been taken into account. In fact, the inner workings of MT and the differences between languages (in our case, English and Portuguese) may impact the validity of the gold labels in supervised tasks. This is something we analyze in this section before providing details on the PORTULAN ExtraGLUE datasets we obtained.

For MT, we use DeepL⁴, a commercial MT tool that tailors translation to two Portuguese variants, European (pt-PT) and Brazilian (pt-BR).

4.1. More than translation

Both statistical and neural sequence-to-sequence MT models are affected by language model probabilities. As a side effect, ill-formed or ungrammatical source sentences are affected in the translation process, hindering the faithfulness of the output in the target language as a direct counterpart of the input in the source language. In fact, MT has been used in grammatical error correction (Rozovskaya and Roth, 2016; Kementchedjheva and Søgaard, 2023). For this reason, we abstain from machine-translating the CoLA dataset, as the obtained translation may easily corrupt the target labels. As an example, the sentence “They drank the pub” (linguistically *ungrammatical*) is translated to pt-BR

⁴All the examples in this section are obtained via DeepL’s web interface (<https://www.deepl.com/translator>) at the time of writing.

as “Eles beberam *no bar*” (“They drank *in the pub*”, *grammatical*). As another example, the sentence “The professor talked us” (*ungrammatical*) is translated to pt-PT as “O professor falou-*nos*” (“The professor talked *to us*”, *grammatical*).

4.2. Gendered nouns and pronoun resolution

English common nouns do not express grammatical gender. On the other hand, Portuguese common nouns do and are used with corresponding gendered determiners (as opposed to English gender-neutral *the* or *a*). This exacerbates the difficulty of properly addressing pronoun reference resolution, given that third-person singular pronouns (and also plural in Portuguese) are gendered. Tasks specifically dealing with pronoun resolution or evaluating the gender robustness of language models are thus prone to corruption via MT. These include WNLI, WSC, and AX_g . While we provide translated versions of WNLI and AX_g , we conduct error analysis to diagnose the quality level of their Portuguese versions.

An example of a translation issue in WNLI is as follows: “Tom said “Check” to Ralph as he took his bishop” / “Tom said “Check” to Ralph as he took Ralph’s bishop” is translated to pt-PT as “O Tomás disse “Xeque” ao Rafa quando *este* lhe tirou o bispo” / “O Tomás disse “Xeque” ao Rafa quando tirou o bispo ao Rafa”. The first sentence in the pair is wrongly translated (*este* means *the latter*), and even though it does not make sense, the target label should change from *entailment* to *not_entailment*.

For WSC, the situation is more critical, as parts of the input are isolated words (usually nouns and pronouns). Thus, obtaining a proper Portuguese equivalent requires more than MT. An example is as follows: from “The *mothers* of Arthur and Celeste have come to the town to fetch them. They are very happy to have them back, but they scold *them* just the same because they ran away”, we want to determine whether the italicised words are co-referent. In this example, there is no separated word matching *them* (which should translate to *elas*) in the translation “As mães do Artur e da Celeste vêm buscá-los à cidade. Estão muito contentes por os terem de volta, mas repreendem-*nos* na mesma por terem fugido”.

AX_g focuses on gender bias, explicitly combining both concerns expressed above. For instance, the hypothesis “The investigator tried to get in contact” is translated into Portuguese as “O investigador tentou entrar em contacto”; its possibly accompanying premises “The investigator wanted to interview the witness in person, but [he | she] was unable to get in contact” are translated into Portuguese as “O

investigador queria entrevistar a testemunha pessoalmente, mas não conseguiu entrar em contacto com *ela*” (for he), or to “O investigador queria entrevistar a testemunha pessoalmente, mas *ela* não conseguiu entrar em contacto” (for she). In the latter case, limiting the possible referents of pronoun *ela* (she) – the only feminine noun is *testemunha* (witness), since *investigador* (investigator) is masculine in Portuguese – renders the *entailment* label wrong, as it should be changed to *not_entailment*.

4.3. Named entities

Another issue we have encountered when using DeepL is the non-deterministic translation of common or proper names, which might make fine-tuning models in these datasets harder or even impact label quality. Consider the following example, taken from WNLI: “Jane gave Joan candy because she wasn’t hungry” / “Jane wasn’t hungry” is translated to pt-PT as “A Joana deu doces à Joana porque ela não tinha fome” / “A Joana não tinha fome”; in this example, one of the distinct proper names is lost. The reverse can also happen: “Bill passed the half-empty plate to John because he was full” / “John was full” is translated to pt-PT as “O Bill passou o prato meio vazio ao John porque estava cheio” / “O João estava cheio”; in this case, a single entity, *John*, is either kept or translated to *João* in the same short text.

As another example from the same dataset, now concerning the same common noun being translated differently, “I couldn’t put the *pot* on the shelf because it was too tall” / “The *pot* was too tall”. is translated to pt-PT as “Não podia colocar a *panela* na prateleira porque era demasiado alta” / “O *pote* era demasiado alto”.

These issues may be prevalent in every dataset, particularly in pt-PT variants.

4.4. Machine-translated tasks

The set of datasets that have been translated and are part of PORTULAN ExtraGLUE⁵ are included in Table 1. As mentioned in Sections 4.1 and 4.2, we leave out the CoLA and WSC datasets.

For MNLI, we provide translations only for the matched and mismatched validation and test sets due to the excessive size of the training set⁶. Likewise, we do not translate the QQP dataset⁷.

Given the nature of the WiC task (based on word sense disambiguation), we posit that a (human or machine) translated version of this dataset is not viable and thus leave it out. Finally, given the focus

⁵Made available at <https://huggingface.co/datasets/PORTULAN/extraglue>.

⁶The training set for MNLI contains 393k rows.

⁷QQP includes a total of 795k rows.

Task	Train	Val	Test	Tokens (en)	Version	Tokens (pt)	mt _e	lab _e	low _q
SST-2	67.3k	872	1.82k	686.1k	pt-PT	725.3k	4%	0%	0%
					pt-BR	724.9k	4%	0%	0%
MRPC	3.67k	408	1.73k	254.3k	pt-PT	287.2k	4%	0%	2%
					pt-BR	284.7k	6%	0%	2%
STS-B	5.75k	1.5k	1.38k	197.5k	pt-PT	220.6k	2%	0%	0%
					pt-BR	217.8k	2%	0%	0%
MNLI _matched	–	9.82k	9.8k	649.4k	pt-PT	660.6k	0%	0%	0%
					pt-BR	661.4k	4%	0%	0%
MNLI _mismatched	–	9.83k	9.85k	680.6k	pt-PT	710.3k	6%	0%	0%
					pt-BR	705.3k	4%	0%	0%
QNLI	105k	5.46k	5.46k	4.82M	pt-PT	5.22M	2%	2%	2%
					pt-BR	5.14M	0%	0%	0%
RTE	2.49k	277	3k	333.8k	pt-PT	364.4k	2%	0%	0%
					pt-BR	360.8k	2%	0%	0%
WNLI	635	71	146	29.7k	pt-PT	30.2k	6%	4%	4%
					pt-BR	29.5k	8%	6%	6%
CB	250	56	250	43.3k	pt-PT	40.4k	6%	2%	2%
					pt-BR	40.5k	8%	2%	4%
AX _b	–	–	1.1k	40.2k	pt-PT	43.3k	20%	4%	14%
					pt-BR	42.7k	20%	4%	12%
AX _g	–	–	356	8.7k	pt-PT	8.9k	22%	10%	10%
					pt-BR	8.8k	20%	6%	8%
BoolQ	9.43k	3.27k	3.25k	1.93M	pt-PT	2.07M	22%	2%	12%
					pt-BR	2.06M	18%	2%	8%
MultiRC	27.2k	4.85k	9.69k	12.99M	pt-PT	13.69M	10%	2%	2%
					pt-BR	13.65M	10%	4%	4%
CoPA	400	100	500	19.5k	pt-PT	18.6k	2%	2%	2%
					pt-BR	19.3k	2%	2%	2%

Table 1: PORTULAN ExtraGLUE datasets. For each task, we include the size of each partition, the number of tokens in each Portuguese variant, and results from the sample analysis in percentages (mt_e = machine translation errors, lab_e = corrupted labels, and low_q = low-quality translated samples).

of the ReCoRD task on named entities and the issues encountered and described in Section 4.3, we abstain from translating this dataset as well.

To improve translation quality, we concatenate each dataset entry’s textual columns with a line break. This ensures that the MT model can access as much context as is available (which may be critical for datasets with very short text spans) and is in line with previous findings (Artetxe et al., 2020a).

As it can be seen in Table 1, the number of tokens varies among the Portuguese language variants. To better assess how different these are in the resulting machine-translated datasets, we calculate the BLEU score (Papineni et al., 2002) between both variants. For that, we rely on 4-grams; BLEU is calculated independently for each feature (text column in a dataset) and then averaged for the whole dataset. The BLEU score averaged over both directions (pt-PT → pt-BR and pt-BR → pt-PT) and for all datasets is 57.3, with the lowest value of 46.7 on the CoPA dataset and the highest of 64.5 on RTE. These values demonstrate that there are significant

differences between the translations obtained for each variant via DeepL.

To assess the quality of each machine-translated dataset, we resort to sampling 50 randomly selected examples, which were manually checked by three of the authors⁸ for translation correctness and target label consistency. The rightmost columns in Table 1 show the results of this analysis: obvious translation errors, label corruption, and low-quality entries that should be removed from the dataset, given its nature.

The main translation problems we have observed concern pronoun resolution or gender issues (as already emphasized in Section 4.2), idiomatic expressions, inconsistent translations in pairs of sentences, and a few cases of ‘hallucinations,’ among other problematic mistranslations. In some cases, these problems have an impact on the correctness of the labels (mainly in WNLI and AX_g); in other cases, they mostly imply a dataset of lesser quality (such as in AX_b and BoolQ). In the specific case of

⁸Portuguese native speakers and fluent in English.

Hyper-parameter	Value
r	8
alpha	32
dropout	0.05
batch size	8
learning rate	2×10^{-5}
weight decay	0.05

Table 2: LoRA hyper-parameters.

AX_g , even when the translation is correct, it does not do justice to the nature of the task, which loses its purpose (e.g., *his/her* translate the same way to Portuguese).

Despite these problems, machine translation errors amount to only an average of 8%, with a mode as low as 2%. Label errors are even lower, with an average of 2% and a zero mode. We did not observe relevant differences between Portuguese language variants.

5. Albertina LoRA Models

We train and make available a set of fine-tuned low-rank adaptations of Albertina-based language models. For several PORTULAN ExtraGLUE datasets, we fine-tune a 1.5B Albertina language model for two Portuguese variants, European (pt-PT) and Brazilian (pt-BR). The resulting models are a practical validation for the created datasets.

5.1. Set up

First, we adapt each task example for tokenization regarding their input components. For this, we concatenate the input features with a special token separator. On the MRPC and STS-B similarity tasks, we concatenate the first and second sentences. On the CB and RTE inference tasks, the hypothesis and premise; on QNLI, the sentence and question. For the BoolQ Question-answering task, we concatenate the passage and question; for MultiRC, the paragraph, question, and answer, truncating the paragraph if needed. For the CoPA reasoning task, we concatenate the premise and question and then join with each choice, resulting in two inputs. During tokenization, we truncate the examples with a maximum context length of 128 tokens, except in MultiRC, which uses 256 tokens.

After tokenization, we apply a low-rank adapter (Hu et al., 2022) with the hyper-parameters shown in Table 2. Due to hardware limitations, it was unfeasible to perform a grid search on these hyper-parameters. We chose the current hyper-parameters by resorting to small-scale exploratory experiments. Because several datasets lack test labels, we fine-tuned models on the training split and evaluated them on the validation split.

5.2. Results

The fine-tuning results are presented in Table 3. All these models are the first baselines for the tasks regarding these new datasets.

Comparing the empirical results between the two variants (pt-PT and pt-BR), we observe that the pt-BR variant achieves better scores than the pt-PT variant in seven tasks (SST-2, MRPC, STS-B, RTE, WNLI, CB, and BoolQ), while the pt-PT variant has better scores in three tasks (QNLI, MultiRC, and CoPA). It is worth noting, however, that the differences are marginal in most cases. The larger discrepancies are observed for the WNLI, BoolQ and CoPA tasks. The first two tasks yield better results with the pt-BR variant, whereas the CoPA task achieves a better outcome in the pt-PT variant.

We can also compare the results with those available for a subset of tasks and the current state-of-the-art Albertina models, as reported in Rodrigues et al. (2023). For the pt-PT variant: in MRPC we obtain 0.8969 accuracy compared to 0.9171 in the original 900M Albertina model; in STS-B we obtain a Pearson correlation of 0.8905 compared to Albertina’s 0.8801; in RTE we obtain 0.7870 accuracy against .8339; and in WNLI we obtain 0.6197 accuracy against 0.4225. For the pt-BR variant: in MRPC we obtain 0.9184 accuracy compared to 0.9071 in the original 900M Albertina model; in STS-B we obtain a Pearson correlation of 0.8940 compared to Albertina’s 0.8910; in RTE we obtain 0.7978 accuracy against 0.7545; and in WNLI we obtain 0.6901 accuracy against 0.4601. We note, however, that the translations of these tasks in PORTULAN ExtraGLUE may differ from the translations used by the authors of the Albertina model for their evaluations. This is certainly true for the pt-BR variant, as the MT model used differed.

Table 3 also includes the results obtained by fine-tuning the multilingual XLM-RoBERTa-XL⁹ model (Conneau et al., 2020) following the same LoRA approach. XLM-RoBERTa-XL is significantly larger (3.5B parameters) than Albertina 1.5B. Even so, we note the benefits of using monolingual models when comparing such results with our Albertina 1.5B LoRA models. In fact, we observe improvements in Albertina 1.5B LoRA models for all tasks and in both Portuguese variants. In some cases, improvements are significant.

When comparing with the DeBERTa¹⁰ (He et al., 2021) model (the foundation model for Albertina) applied to the original English datasets, the results of our low-rank adapters on the PORTULAN ExtraGLUE datasets fall behind in most cases. This

⁹<https://huggingface.co/facebook/xlm-roberta-xl>

¹⁰<https://huggingface.co/microsoft/deberta-v2-xxlarge>

Task	Albertina 1.5B		XLM-RoBERTa-XL		DeBERTa-V2-XXLarge en
	pt-PT	pt-BR	pt-PT	pt-BR	
Single sentence					
SST-2	0.9392	0.9450	0.9323	0.9392	0.9633
Similarity					
MRPC	0.8969	0.9184	0.8696	0.8651	0.9266
STS-B	0.8905	0.8940	0.8743	0.8734	0.9170
Inference					
QNLI	0.9398	0.9361	0.9237	0.9237	0.9608
RTE	0.7870	0.7978	0.6571	0.6606	0.8917
WNLI	0.6197	0.6901	0.5634	0.5634	0.7887
CB	0.8385	0.8554	0.6280	0.6160	0.8936
QA					
BoolQ	0.7456	0.7807	0.6538	0.6587	0.8900
MultiRC	0.7257	0.7169	0.6926	0.6925	0.8243
Reasoning					
CoPA	0.8500	0.8200	0.5000	0.5600	0.9200

Table 3: Evaluation scores on validation sets for both variants regarding the different categories of datasets (Single Sentence, Similarity, Inference, Question-Answering, and Reasoning). Performance on SST-2, QNLI, RTE, WNLI, BoolQ, and CoPA is measured with accuracy; on MRPC, CB, and MultiRC with F1; and on STS-B with Pearson. For comparison, we include results for the multilingual XLM-RoBERTa-XL 3.5B model, fine-tuned using the same LoRA approach. For reference, we also include results for English by applying LoRA to the DeBERTa-V2-XXLarge 1.5B model (based on which Albertina has been developed).

is expected for at least two reasons: first, Albertina was pre-trained with far fewer data than DeBERTa; second, we rely on machine translation to obtain the datasets for the tasks, which, as discussed before, isn't without issues. Tasks exhibiting significant differences in performance include WNLI, which, as explained in Section 4.2, has issues related to pronoun resolution.

prove this benchmark with manual curation of the datasets (in particular, the test sets) and expand it with new ones. Additionally, developing new datasets from scratch may better reflect the language and the cultures latent within language variants (which go well beyond European and Brazilian ones). Evolving these in a leaderboard would help foster research in the Portuguese language.

6. Conclusion

We contribute an open benchmark suite to support the development of the neural processing of Portuguese. In this initial version, this suite comprises 14 datasets for downstream tasks of various types, including single sentence tasks, similarity tasks, inference tasks, and reasoning tasks. To kick-start benchmarking for this language, these datasets were machine-translated from mainstream benchmarks in the literature and designated as PORTULAN ExtraGLUE. We also make available baseline models for 10 of these tasks, developed with the low-rank adaptation approach over a state-of-the-art and open language model for Portuguese.

Even though MT datasets have their limitations and pitfalls, our manual analysis has found a relatively reduced amount of (translation and label) errors. We believe this renders our obtained datasets highly useful for assessing the comparative performance of neural language models for Portuguese.

In future work, it would be important to im-

Acknowledgements

This research was partially supported by: PORTULAN CLARIN – Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT (PIN-FRA/22117/2016); ACCELERAT.AI – Multilingual Intelligent Contact Centers, funded by IAPMEI (C625734525-00462629); ALBERTINA – Foundation Encoder Model for Portuguese and AI, funded by FCT (CPCA-IAC/AV/478394/2022); and Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC).

Bibliographical References

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. [Translation artifacts in cross-lingual transfer learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Alexandre Audibert, Massih R Amini, Konstantin Usevich, and Marianne Clausel. 2023. [Low-rank updates of pre-trained weights for multi-task learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7544–7554, Toronto, Canada. Association for Computational Linguistics.
- E. M. Bender. 2011. [On achieving and evaluating language-independence in NLP](#). *Linguistic Issues in Language Technology*, 6.
- Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. [Automatic Spanish translation of SQuAD dataset for multilingual question answering](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.
- Nuno Ramos Carvalho, Alberto Simões, and José João Almeida. 2021. Bootstrapping a data-set and model for question-answering in Portuguese (short paper). In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). 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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can transformer be too compositional? analysing idiom processing in neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. [The CommitmentBank: Investigating projection in naturally occurring discourse](#). *Proceedings of Sinn und Bedeutung*, 23(2):107–124.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Martin d’Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. [FQuAD: French question answering dataset](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.
- William B. Dolan and Chris Brockett. 2005. [Automatically constructing a corpus of sentential](#)

- paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. [Cross-lingual argumentation mining: Machine translation \(and a bit of projection\) is all you need!](#) In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. ASSIN: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.
- Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. 2010. [Second HAREM: Advancing the state of the art of named entity recognition in Portuguese](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. [SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. [Towards a unified view of parameter-efficient transfer learning](#). In *International Conference on Learning Representations*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, and et al. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. [Challenges in context-aware neural machine translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15246–15263, Singapore. Association for Computational Linguistics.
- Yova Kementchedjheva and Anders Søgaard. 2023. [Grammatical error correction through round-trip machine translation](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2208–2215, Dubrovnik, Croatia. Association for Computational Linguistics.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. [Looking beyond the surface: A challenge set for reading comprehension over multiple sentences](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12*, page 552–561. AAAI Press.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. [XGLUE: A new benchmark dataset](#)

- for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.
- Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. **Compacter: Efficient low-rank hypercomplex adapter layers**. In *Advances in Neural Information Processing Systems*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. **Universal Dependencies v2: An evergrowing multilingual treebank collection**. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. **WiC: the word-in-context dataset for evaluating context-sensitive meaning representations**. 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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. **XCOPA: A multilingual dataset for causal commonsense reasoning**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. **Exploring the limits of transfer learning with a unified text-to-text transformer**. *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. **SQuAD: 100,000+** questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.
- João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Freitas Osório. 2023. **Advancing neural encoding of portuguese with transformer Albertina PT-***. In *Progress in Artificial Intelligence - 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part I*, volume 14115 of *Lecture Notes in Computer Science*, pages 441–453. Springer.
- Alla Rozovskaya and Dan Roth. 2016. **Grammatical error correction: Machine translation and classifiers**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2205–2215, Berlin, Germany. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. **Gender bias in coreference resolution**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. **HAREM: An advanced NER evaluation contest for Portuguese**. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. **Gender bias in machine translation**. *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. **Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

- Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. [RussianSuperGLUE: A Russian language understanding evaluation benchmark](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Afonso Sousa, Bernardo Leite, Gil Rocha, and Henrique Lopes Cardoso. 2021. Cross-lingual annotation projection for argument mining in portuguese. In *Progress in Artificial Intelligence*, pages 752–765, Cham. Springer International Publishing.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and et al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2023. [DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3274–3287, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. [On the features of translationese](#). *Digital Scholarship in the Humanities*, 30(1):98–118.
- Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. [Challenges of neural machine translation for short texts](#). *Computational Linguistics*, 48(2):321–342.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, and et al. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Procs. 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. [PAWS-X: A cross-lingual adversarial dataset for paraphrase identification](#). 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)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.
- Aleš Žagar and Marko Robnik-Šikonja. 2022. [Slovene SuperGLUE benchmark: Translation and evaluation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France. European Language Resources Association.
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. [Evaluating commonsense in pre-trained language models](#). In *Procs. 34th AAAI, New York, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.

Invited Talk: The Way Towards Massively Multilingual Language Models

François Yvon

Sorbonne Université, CNRS, ISIR
yvon@isir.upmc.fr

Abstract

In this talk, I will discuss the training and evaluation of massively multilingual language models, that can handle dozens or even hundreds of languages. After motivating the development of such models, I will draw some lessons learned in the course of developing Glot500, a language model covering 500 languages, and some associated resources such as language identification softwares. I will also focus on the challenges raised by “low resourced” languages, i.e. languages for which (a) the available training data is often incomplete, highly specialised and also possibly very noisy; (b) the evaluation data are non existent, requiring to use innovative evaluation strategy, e.g. based on various cross-lingual alignment tasks.

Keywords: multilingual language model, low-resource languages, language identification

Bio

François Yvon is a senior CNRS researcher at the ISIR laboratory of Sorbonne-Université in Paris, France, working on Machine Translation and Multilingual Language Models. Before this, F. Yvon has been leading activities in Machine Translation at LISN / LIMSI in Orsay for about 15 years, resulting in more than one hundred scientific publications on all aspects related to the development and evaluation of multilingual language processing technologies, from word and sentence alignment to translation modelling and evaluation, including recent work on multi-domain adaptation in Machine Translation and on cross-lingual transfert learning issues. He has acted as coordinator or Principal Investigator in multiple past national and international projects in Machine Translation and has supervised more than 20 PhDs on related topics. Between 2013 and 2020, Dr. Yvon has also been the general director of the LIMSI laboratory in Orsay. He is a board member of the European chapter of the Association for Computational Linguistics, of the MetaNet network, and has recently contributed as an expert on linguistic technologies for the French language to several European projects (European Language Resource Collection, ELE – European Language Equality, ELG – European Language Grid).

Exploring the Necessity of Visual Modality in Multimodal Machine Translation using Authentic Datasets

Zi Long¹, Zhenhao Tang², Xianghua Fu¹, Jian Chen², Shilong Hou², Jinze Lyu²

¹College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China

²College of Application and Technology, Shenzhen University, Shenzhen, China

Abstract

Recent research in the field of multimodal machine translation (MMT) has indicated that the visual modality is either dispensable or offers only marginal advantages. However, most of these conclusions are drawn from the analysis of experimental results based on a limited set of bilingual sentence-image pairs, such as Multi30k. In these kinds of datasets, the content of one bilingual parallel sentence pair must be well represented by a manually annotated image, which is different from the real-world translation scenario. In this work, we adhere to the universal multimodal machine translation framework proposed by Tang et al. (2022). This approach allows us to delve into the impact of the visual modality on translation efficacy by leveraging real-world translation datasets. Through a comprehensive exploration via probing tasks, we find that the visual modality proves advantageous for the majority of authentic translation datasets. Notably, the translation performance primarily hinges on the alignment and coherence between textual and visual contents. Furthermore, our results suggest that visual information serves a supplementary role in multimodal translation and can be substituted.

Keywords: MMT, image retrieval, visual noise filtering, supplementary text retrieval

1. Introduction

With the development of neural machine translation (NMT), the role of visual information in machine translation has attracted researchers' attention (Specia et al., 2016; Elliott et al., 2017; Barraud et al., 2018). Different from those text-only NMT (Bahdanau et al., 2014a; Gehring et al., 2016), a bilingual parallel corpora with manual image annotations are used to train an MMT model by an end-to-end framework, and therefore visual information can assist NMT model to achieve better translation performance (Calixto and Liu, 2017; Calixto et al., 2017; Su et al., 2021).

Concurrently, researchers have also undertaken a diverse range of experiments in an effort to validate the specific role of visual information in NMT. For example, Grönroos et al. (2018a) and Lala et al. (2018) observed that the robustness of MMT systems remains unaffected when the input image lacks direct relevance to the accompanying text. Notably, the absence of visual features, as highlighted by Elliott (2018), also does not yield detrimental effects. Wu et al. (2021) underscores that the utilization of the visual modality serves as a regularization mechanism during training rather than serving as a true complement to the textual modality. Oppositely, Caglayan et al. (2019) delve into the correlation between visual features and text. Their investigation reveals that incorporating the input image aids translation, particularly when certain input words are masked. Li et al. (2022) design more detailed probing tasks and found that stronger vision features strengthen MMT systems.

Note that most of the previous conclusions are drawn from the analysis of experimental results based on a restricted selection of manually annotated bilingual sentence-image pairs, known as the Multi30k dataset (Elliott et al., 2016). Within the Multi30k dataset, as depicted in Table 1, the sentences primarily comprise common and straightforward vocabulary, with each bilingual parallel sentence pair being effectively depicted by a single image. Table 1 also presents an illustration of a bilingual sentence-image pair extracted from a genuine news report from the United Nations News¹, alongside examples of sentence pairs derived from other authentic translation datasets. Evidently, a substantial disparity exists between the Multi30k dataset and the authentic translation data. Hence, the evidence and findings derived from Multi30k may potentially exhibit inadequate generalizability and offer limited utility when attempting to analyze the role of the visual modality in MMT within real-world translation scenarios. In these scenarios, sentences often incorporate rare and uncommon words and are only partially depicted by accompanying images.

In a recent study, Tang et al. (2022) introduced a universal multimodal neural machine translation model that integrates open-vocabulary image retrieval techniques. In this work, inspired by Tang et al. (2022), we formulate a set of comprehensive probing tasks aimed at assessing the extent to which the visual modality enhances MMT within real-world translation scenarios. In addition to com-

¹<https://news.un.org/en/>



Data source	Sentences	Image
Multi30k	EN: A dog is running in the snow. DE: Ein Hund rennt im Schnee.	
UN News	EN: Rescue workers look for survivors in a building in Samada, Syria destroyed by the February 6 earthquake. DE: Rettungskräfte suchen nach Überlebenden in einem Gebäude in Samada, Syrien, das durch das Erdbeben vom 6. Februar zerstört wurde.	
Bible	EN: I saw, and behold, there was no man, and all the birds of the sky had fled. DE: Ich sah, und siehe, da war kein Mensch, und alle Vögel unter dem Himmel waren weggeflogen.	no image
MultiUN	EN: Development assistance cannot by itself prevent or end conflict. DE: Entwicklungshilfe allein kann Konflikte weder verhüten noch beenden.	no image

Table 1: Comparison between Multi30k Dataset and Authentic Datasets

only used Multi30k, we conduct an extensive set of experiments across four authentic text-only translation datasets. We further evaluated two visual noise filtering approaches based on the correlation between textual and visual content. Furthermore, we investigate the necessity of visual modality in the current multimodal translation process by substituting visual data with closely equivalent textual content. To summarize, our findings are:

- (1) Visual modality is mostly beneficial for translation, but its effectiveness wanes as text vocabulary becomes less image-friendly.
- (2) The MMT performance depends on the consistency between textual and visual contents, and utilizing filters based on the textual-visual correlation can enhance the performance.
- (3) Visual information plays a supplementary role in the multimodal translation process and can be substituted by the incorporation of additional textual information.

2. Related Work

The integration of extra knowledge to build fine-grained representations is a crucial aspect in language modeling (Li et al., 2020a,b; Zhang et al., 2020). Incorporating the visual modality into language modeling has the potential to enhance the machine’s understanding of the real world from

a more comprehensive perspective. Inspired by the studies on the image description generation task (Elliott et al., 2015; Venugopalan et al., 2015; Xu et al., 2015), MMT models have gradually become a hot topic in machine translation research. In some cases, visual features are directly used as supplementary information to the text presentation. For example, Huang et al. (2016) take global visual features and local visual features as additional information for sentences. Calixto and Liu (2017) initialize the encoder hidden states or decoder hidden states through global visual features. Calixto et al. (2017) use an independent attention mechanism to capture visual representations. Caglayan et al. (2016) incorporate spatial visual features into the MMT model via an independent attention mechanism. On this basis, Delbrouck and Dupont (2017b) employs compact bilinear pooling to fuse two modalities. Lin et al. (2020) attempt to introduce the capsule network into MMT, they use the timestep-specific source-side context vector to guide the routing procedure. Su et al. (2021) introduce image-text mutual interactions to refine their semantic representations.

Researchers have also come to recognize the potential redundancy of the visual modality. Inconsequential images exhibit minimal impact on translation quality, and the absence of images does not yield a significant drop in BLEU scores, as noted by Elliott (2018). Encouraging findings emerged in the study by Caglayan et al. (2019). They high-

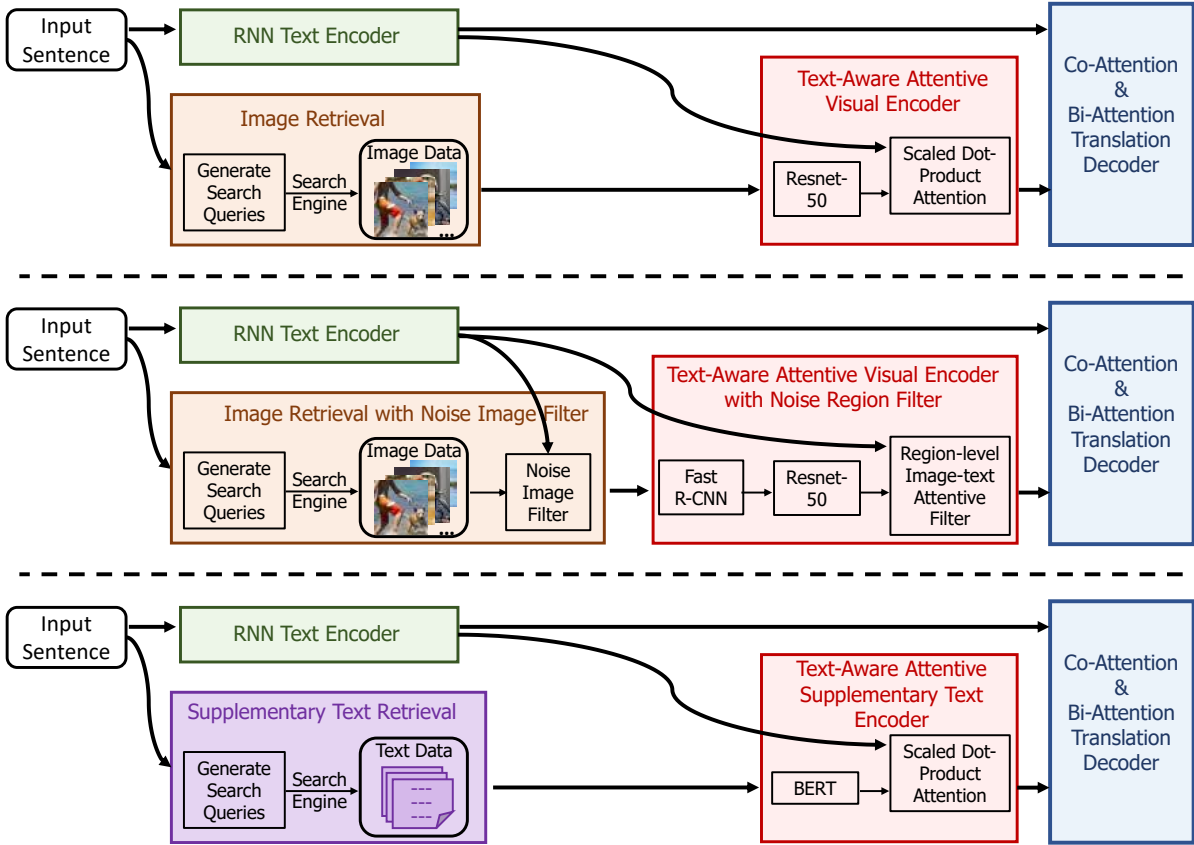


Figure 1: Frameworks of three probing methods

lighted the continuing utility of the visual modality in scenarios where linguistic context is limited but noted its diminished sensitivity when exposed to complete sentences. In a more recent investigation, Wu et al. (2021) attributed the observed BLEU improvement in MMT tasks to training regularization. They underscored the importance of constructing appropriate probing tasks with inadequate textual input. It’s important to highlight that the proposed probing task represents an enhanced iteration building upon prior research (Caglayan et al., 2019; Wu et al., 2021). Li et al. (2022) made a systematic study on whether stronger vision features are helpful. All the preceding research has been conducted exclusively on the Multi30k dataset, which has limitations in scale and considerably differs from real-world translation scenarios. In this study, we employ the framework introduced by Tang et al. (2022) to systematically examine the influence of visual information across various authentic translation datasets, extending our analysis beyond the limitations of the small and specialized Multi30k dataset.

3. Preliminary

We start with a description of three probing methods employed in this work, which encompass the approach introduced by Tang et al. (2022) and two additional methods derived from it. Figure 1 shows frameworks of these three methods.

3.1. MMT with Search Engine Based Image Retrieval

As depicted in the top section of Figure 1, Tang et al. (2022) introduced a search engine-based image retrieval technique and a text-aware attention image encoder. This innovation enables the handling of authentic text-only translation data within MMT systems. We implement this approach across multiple authentic translation datasets to examine the influence of visual information across datasets with varying styles. To ensure the comprehensiveness of this paper, this section will provide a brief overview of the approach proposed by Tang et al. (2022).

Text Encoder In this work, we employ a commonly utilized bi-directional LSTM as the RNN text encoder. For a given sentence denoted as X ,

the output of the text encoder is represented as $C = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N)$, where N denotes the length of the sentence X .

Image Retrieval To emphasize the core components of the sentence and mitigate the impact of noise, including stopwords and infrequent words, Tang et al. (2022) utilized the TF-IDF method (Witten et al., 2005) to generate search queries for image search engines. Subsequently, the generated search queries are utilized in image search engines to retrieve the first available image associated with each query. For each given sentence X , M search queries denoted as (q_1, q_2, \dots, q_M) are generated, and subsequently M images represented as (A_1, A_2, \dots, A_M) are retrieved from search engines.

Text-Aware Attentive Visual Encoder Each image A_m ($m = 1, \dots, M$) is transformed into a 196×1024 dimensional feature vector using ResNet-50 (He et al., 2016). A simple but effective scaled dot-product attention in visual encoder is subsequently employed in the visual encoder to derive a resultant visual representation. Here, we utilize the average pooling C of the text representation $C = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N)$ as the query, while the visual feature vectors A_1, A_2, \dots, A_M serve as the keys and values in this attention mechanism. The resultant visual representation A is also expressed as a 196×1024 dimensional feature vector, which can be regarded as a matrix $A = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L)$, where $L = 196$ and each $\mathbf{a}_l \in R^{1024}$ ($l = 1, \dots, L$). Visual representation $A = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L)$ and text representation $C = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N)$ are then used as the inputs of translation decoder.

Translation Decoder For the decoder, we adopt the approach introduced by Su et al. (2021), implementing both a bidirectional attention network and a co-attention network to effectively capture the underlying semantic interactions between textual and visual elements. Based on the results of the preliminary experiment, it was evident that transformer-based models did not confer a performance advantage on datasets like Global Voices and other smaller ones. Consequently, we followed the approach of Tang et al. (2022) and selected LSTM as our foundational model. The bidirectional attention network enhances the representations of both text and image. These enhanced representations are subsequently input into the co-attention network to obtain the time-dependent context vector c_t and the visual vector v_t . Within the co-attention network, we calculate the probability distribution for the next target word y_t using the previous hidden state s_{t-1} , the previously generated target word

y_{t-1} , the time-dependent context vector c_t , and the time-dependent visual vector v_t .

3.2. MMT with Visual Noise Filtering

Considering that the noise images obtained from search engines could have a substantial impact on the performance of the MMT system, we further evaluated two visual noise filtering approaches based on the correlation between textual and visual content, as depicted in the central part of Figure 1. One approach utilizes the pretrained CLIP model to filter out noise images, while the other employs a region-level image-text attentive filter module to filter out noisy image regions.

Noise Image Filter In the CLIP-based noise image filtering approach, we begin by retrieving M' ($M' > M$) images from search engines for each input sentence. Following this, we calculate the correlation between the input text and the retrieved images using a pretrained CLIP model (Radford et al., 2021). Subsequently, we select only the top- M images with the highest correlation to the input source text as the output of the image retrieval process.

Noise Region Filter In the noise image region filtering approach, we begin by extracting convolutional feature maps from the top- O most confident regions denoted as (r_1, \dots, r_O) in each collected image. This is achieved using a pretrained Faster R-CNN model (Ren et al., 2015), aiding in the initial filtration of visual information that may be challenging to distinguish as distinct regions in the images. The image region of each collected image is then represented as a 1024 dimensional feature vector using ResNet-50. For all the retrieved M images, we extract a total of $M \times O$ regions $(r_1, \dots, r_{M \times O})$, resulting in $M \times O$ feature vectors $(\mathbf{a}_1, \dots, \mathbf{a}_{M \times O}, \mathbf{a}_o \in R^{1024})$. Subsequently, we compute the correlation score between each image region and the input text using the following equation:

$$S(\mathbf{a}_o, C') = V_a \tanh(W_a \mathbf{a}_o + U_a C')$$

Here, C' represents the average pooling of the text representation $C = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N)$. We retain only the visual information from the top- O most relevant regions out of the initially extracted $M \times O$ regions. This preserved visual information serves as the visual representation for the given input sentence, denoted as $A = \{\mathbf{a}_o | S(\mathbf{a}_o, C') \text{ ranks in the top-}O, 1 \leq o \leq M \times O\}$, and it is subsequently fed into the translation decoder module. Less relevant regions are discarded during this process.

3.3. Supplementary Text Enhanced NMT

As discussed by Caglayan et al. (2019), multimodal translation models typically view visual information as a complementary component to textual information. However, we raise the question of whether this complementary role can also be achieved by incorporating additional textual information, potentially obviating the need for images in the process. Hence, our investigation aims to assess the necessity of visual information in the existing multimodal translation process by substituting visual data with nearly equivalent textual information. As illustrated in the lower section of Figure 1, we replace the image retrieval module with a supplementary text retrieval module and substitute the text-aware attentive visual encoder with a similar text-aware attentive supplementary text encoder.

Supplementary Text Retrieval Similar to the process of retrieving images from search engines, we collected supplementary textual data from search engines. For every input source sentence X , we follow the same approach as outlined in Section 3.1 to generate M search queries (q_1, \dots, q_M) . Subsequently, we collect M sentences (T_1, \dots, T_M) that contains all the terms present in the respective search queries $(q_i \subseteq T_i, 1 \leq i \leq M)$.

Text-Aware Attentive Supplementary Text Encoder Each gathered supplementary text T_m ($m = 1, \dots, M$) is transformed into a $N \times 1024$ dimensional feature vector using BERT (Devlin et al., 2018), where N denotes the length of the gathered text data. To ensure consistency, these textual feature vectors are subsequently padded to match the dimensions of $L \times 1024$ ($L = 196$), aligning them with the visual feature vectors. These feature vectors are then integrated into the scaled dot-product attention module as keys and values, with the average pooling C' representing the input text serving as the query. The resultant supplementary text representation is then passed to the translation decoder.

4. Experiment Setup

4.1. Dataset

We conducted experiments on five commonly used machine translation datasets, including multimodal machine translation dataset Multi30k (Elliott et al., 2016) English-to-German, Global Voices (Tiedemann, 2012) English-to-German, and WMT'16 (100k) English-to-German (Newstest2016 as the test set)², Bible (Christodouloupoulos and

²To ensure a focused evaluation of the retrieved visual information's effectiveness, we intentionally sought

dataset	training set	dev set	test set
Multi30k	29,000	1,014	1,000
Global Voices	69,227	2,000	2,000
WMT'16 (100k)	100,000	2,000	3,000
Bible	56,734	1,953	1,821
MultiUN	56,235	4,000	4,000

Table 2: Statistics of datasets

Steedman, 2015) English-to-German, and MultiUN (Eisele and Chen, 2010) English-to-German. The statistics for each dataset are presented in Table 2.

4.2. Model Implementation

For image retrieval, we used the Microsoft Bing³ as the image search engine. In contrast, for supplementary text retrieval, we gathered sample sentences that included all the terms found in the respective search queries by referencing the Microsoft Bing Dictionary⁴. As described in Section 3.1 and Section 3.3, we set M to 5. This choice signifies that we formulated 5 search queries and procured 5 images or supplementary text instances⁵ for every source language sentence.

Regarding the text encoder, we used a bi-directional RNN with GRU to extract text features. Specifically, we used a 256 dimensional single-layer forward RNN and a 256 dimensional single-layer backward RNN. For the translation decoder, we adhered to the approach proposed by Su et al. (2021) and utilized a modified cGRU with hidden states of 256 dimensions. Furthermore, we configured the embedding sizes for both source and target words to be 128.

As described in Section 3.1, the visual encoder we employed leveraged the res4f layer of a pre-trained ResNet-50(He et al., 2016) model to extract visual features of dimensions 196×1024 . Furthermore, as described in Section 3.3, the supplementary text encoder utilized a BERT model pretrained on the BooksCorpus(Zhu et al., 2015) and English Wikipedia⁶. This model was employed to extract

to minimize the impact of data size. Consequently, we opted to construct our training set by randomly sampling 100,000 sentence pairs from the total pool of 4.5 million sentence pairs. This sampling approach aligns our dataset size more closely with that of other datasets for a fairer assessment.

³<https://global.bing.com/images>

⁴<https://www.bing.com/dict>

⁵When an insufficient number of sample sentences can be collected, we resort to large pretrained models like ChatGPT to generate sentences that meet the search query.

⁶https://en.wikipedia.org/wiki/English_Wikipedia

Method		BLEU Score
Text-only NMT	Bi-LSTM (Calixto et al., 2017)	33.70
	Transformer (Zhang et al., 2019)	36.86
MMT with Original Images	Zhang et al. (2019)	36.86
	Zhao et al. (2021)	38.40
	Su et al. (2021)	39.20
	Tang et al. (2022) (Section 3.1)	38.14
MMT with Retrieved Images	Zhang et al. (2019)	36.94
	Tang et al. (2022) (Section 3.1)	38.43
	MMT with Visual Noise Filtering (Section 3.2)	38.51
NMT with Retrieved Supplementary Text (Section 3.3)		39.13

Table 3: Results on Multi30K

Method	Dataset				
	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
Text-only NMT	33.70	9.22	7.99	35.23	39.49
MMT with Random Images	37.65	9.29	8.11	35.31	39.48
MMT with Blank Images	37.79	9.46	8.31	35.39	39.52
MMT with Retrieved Images	38.43	9.81	8.41	35.42	39.53

Table 4: Translation performance across diverse datasets under varied image conditions (BLEU score)

textual features of dimensions $N \times 1024$, where N represents the length of the retrieved supplementary text.

Regarding the noise image filter, we set $M' = 10$ and used a CLIP model (Radford et al., 2021) pretrained on the YFCC100M dataset (Thomee et al., 2016) to filter out noisy images. For the noise region filter, we configured it with $O = 128$. Here, we utilized a pretrained Faster R-CNN model (Ren et al., 2015) that had been trained on the Open Images dataset (Kuznetsova et al., 2020). This model was employed to identify and filter noisy regions in images effectively.

4.3. Training Parameters

We initiated the word embeddings and other associated model parameters randomly, following a uniform distribution with a range of -0.1 to 0.1 . During training, we employed the Adam optimizer with a mini-batch size of 32 and set the learning rate to 0.001. Additionally, a dropout strategy with a rate of 0.3 was applied to further enhance the models. The training process continued for up to 15 epochs, with early stopping activated if the BLEU (Papineni et al., 2002) score on the development set did not exhibit improvement for 3 consecutive epochs. The model with the highest BLEU score on the dev set was selected for evaluation on the test set. To minimize the impact of random seeds on experimental results and ensure result stability, we conducted the experiment 5 times with fixed random seeds and reported the macro-average of BLEU scores as the final result.

4.4. Baselines

In the case of the Multi30k dataset, we conducted a quantitative comparison of the probing methods with several recent MMT models (Zhang et al., 2019; Zhao et al., 2021; Su et al., 2021; Tang et al., 2022). However, the main focus of this research is to evaluate the necessity of visual information within real-world translation scenarios. Four out of the five datasets utilized in our evaluation experiments are authentic text-only translation datasets without any visual annotation. Consequently, for each dataset, we exclusively employed the text-only Bi-LSTM (Calixto et al., 2017) as a baseline.

The baseline model and the models detailed in Section 3 were all trained using the same training set and identical training parameters. For all these models, we present the 4-gram BLEU score (Papineni et al., 2002) as the primary evaluation metric.

5. Results and Analysis

Table 3 presents the experimental results of the Multi30k dataset. Compared to various baseline models, all three probing methods mentioned in Section 3 have achieved promising results. Notably, the MMT model with visual noise filtering (Section 3.2) achieved a BLEU score of 38.51, while the NMT model with retrieved supplementary text (Section 3.3) achieved an impressive BLEU score of 39.13. In comparison to text-only NMT models (Calixto et al., 2017; Vaswani et al., 2017), the NMT model with retrieved supplementary text significantly outperforms them, showcasing a sub-

stantial increase in BLEU score. When compared to existing MMT methods that utilize original images (Zhang et al., 2019; Zhao et al., 2021; Su et al., 2021), the NMT model with retrieved supplementary text obtains a comparable BLEU score. Furthermore, in contrast to the MMT methods with retrieved images (Zhang et al., 2019; Tang et al., 2022), the NMT model with retrieved supplementary text demonstrates performance gains of approximately 2.2 and 0.7 BLEU points, respectively.

Further experimental results and analysis will be presented in the following sections.

5.1. Translation Performances across Varied Datasets

We firstly quantitatively compared text-only NMT (Calixto et al., 2017) with MMT utilizing retrieved images (Section 3.1) across five diverse datasets mentioned in Section 4.1. As demonstrated in Table 4, MMT achieved significantly higher BLEU scores on Multi30k, higher BLEU scores on Global Voices and WMT'16 (100k), and slightly higher BLEU scores on Bible and MultiUN. It is intriguing to note that the improvement in translation performance is substantial on Multi30k, with an increase of approximately 4.7, whereas the gain on MultiUN is relatively modest, at approximately 0.04.

We speculate that the variations in results among the aforementioned translation datasets, such as Multi30k and other datasets, may be attributed to the differing qualities of images collected through the search engine. To evaluate the influence of the quality of collected images, we train the MMT model with randomly retrieved unrelated images, blank images, and retrieved images from image search engines, respectively.

The evaluation results are shown in table 4. It is obvious that MMT models with retrieved images achieves the highest BLEU score on all Multi30k and other four datasets, demonstrating the effectiveness of visual information from retrieved images. Compared with the model with random images and blank images, the performance gain of collected images is approximately 0.7 & 0.6 BLEU score on Multi30k, and 0.5 & 0.3 BLUE score on Global Voices. However, on WMT'16 (100k), Bible, and MultiUN datasets, models with retrieved images achieve almost the same BLEU score as the model with blank images.

One of the possible reason is that sentences from those three datasets contains fewer entity words that can be represented by images, and therefore, the search engine based image retrieval method collects numbers of noise images. Sentences from WMT'16 (100k), Bible, and MultiUN datasets describe abstract concepts and complex events, while

sentences from Multi30k and Global Voices describe real objects and people, which is more reliable for image retrieval.⁷

To validate the hypotheses, we manually analyzed the image retrieval outcomes of each dataset. In detail, we initially conducted a random sampling of 1,000 sentences and employed the image retrieval methods outlined in Section 3.1 to gather keywords and images for each sentence. Regarding the extracted keywords, we conducted manual assessments to identify whether each keyword qualifies as an entity word. Regarding the collected images, we carried out manual evaluations to determine if an image could offer pertinent visual information for the search query, and those that could not. Images in the latter category were categorized as noise images. Lastly, we tallied the quantity of sentences containing at least half of non-entity keywords and the quantity of sentences harboring at least half of noise images among the collected images.

As presented in Table 5, for the Multi30k dataset, out of 1000 sentences, only 27 sentences contained half or more non-entity keywords, and 61 sentences gathered half or more noise images from search engines. However, in the WMT'16 (100k) dataset, there are 796 sentences with half or more non-entity keywords and 685 sentences with half or more noise images, accounting for more than half of the sampled sentences. Consequently, our method shows poor performance on the WMT'16 (100k) dataset. The Bible dataset and MultiUN dataset exhibit a similar situation. For the Global Voices dataset, there are 94 sentences with half or more non-entity keywords and 228 sentences with half or more noise images. This falls between the Multi30K and WMT'16 (100k) datasets. It is interesting to note that the Multi30k dataset, which has the smallest proportions of non-entity keywords and noise images, achieves the most significant gain in translation performance. On the other hand, datasets with the largest proportions of non-entity keywords and noise images show the smallest gain in translation performance.

5.2. Influence of the Correlation between Text and Images

Table 6 shows the evaluation results of applying two filtering approaches described in Section 3.2 in MMT. It is obvious that MMT models with both noise image filter and noise region filter achieves the highest BLEU score across all datasets, including Multi30k and the other four, underscoring the

⁷Examples of retrieved images from various datasets are presented in Table 8.

	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
Number of sentences with half or more non-entity keywords	27	94	796	398	818
Number of sentences with half of more noise images	61	228	685	761	663

Table 5: Summary of manual analysis of image retrieval outcomes for each dataset

Method	Dataset				
	Multi30k	Global Voices	WMT'16 (100k)	Bible	MultiUN
MMT with retrieved images (Tang et al., 2022)	38.43	9.81	8.41	35.42	39.53
+ noise image filter	38.50	10.12	8.89	36.12	39.91
+ noise region filter	38.46	9.95	8.78	35.84	39.72
+ noise image & region filter	38.51	10.23	8.93	36.38	39.95

Table 6: Results of image and region filtering method across diverse datasets (BLEU score)

effectiveness of these two filtering approaches.⁸

Notably, it is intriguing to note that the noise filtering techniques exhibited more substantial enhancements in translation performance for the WMT'16 (100k), Bible, and MultiUN datasets, in contrast to the improvements observed in the Multi30k and Global Voices datasets. This further underscores the significant impact of the correspondence between image and text content on the translation performance the alignment and coherence between image and text content on the translation performance of the MMT system. It also elucidates why noise filtering methods yield marginal improvements on the Multi30K dataset.

In conclusion, the translation performance of the multimodal model primarily hinges on the consistency between textual and visual content. In other words, the more alignment exists between textual and visual content, the greater enhancement in translation performance with multimodal translation compared to text-only translation. Hence, we arrive at a conclusion that aligns closely with (Caglayan et al., 2019), which suggest that multimodal translation models predominantly treat visual information as a complement to textual information.

5.3. Exploring the Necessity of Visual Modality

We conducted a quantitative comparison between MMT with retrieved images (Section 3.1) and NMT with retrieved supplementary texts on the Multi30k dataset. Table 7 shows the experimental results. In comparison to MMT model employing images for translation enhancement, the approach integrating

Method	BLEU score
text-only NMT	33.70
+ visual information (MMT with retrieved images) (Tang et al., 2022)	38.43
+ textual information	39.13
+ visual & textual information	38.55

Table 7: Results on Multi30k using visual information or textual information enhanced NMT

supplementary textual data for translation enhancement demonstrated a significantly higher BLEU score of 39.13. Remarkably, the combined utilization of both images and supplementary texts for translation enhancement yielded a BLEU score of 38.55, positioning itself between image-enhanced NMT and text-enhanced NMT.

This demonstrates that both additional visual and supplementary textual information play an entirely equivalent supplementary role in the translation process. Moreover, in most cases, the utilization of supplementary textual information assists the translation process more effectively.⁹

Therefore, we speculate that multimodal translation models trained on a large volume of data might face challenges in outperforming text-only translation models trained on comparable data volumes. This is because as the volume of data used in multimodal model training increases, the potential impact of visual information could diminish. We will verify this in future work.

⁸A correct example generated by MMT with visual noise filtering is presented in Table 9.

⁹A correct example comparing NMT with retrieved supplementary texts to MMT with retrieved images is presented in Table 10.

6. Conclusions

In this paper, we conduct an in-depth exploration into the role of visual information within the multimodal translation process on Multi30k and other four authentic translation datasets. Our findings emphasize that the substantial correlation between visual and textual content significantly impacts the efficacy of multimodal translation, while employing filtering mechanisms based on the textual-visual correlation can enhance translation performance. Additionally, experimental results reveal that visual information plays a supplementary role in the multimodal translation process. This supplementary function of visual information can be substituted by the incorporation of supplementary textual information. As one of our future work, we plan to assess the impact of the visual modality on more extensive translation datasets, including the complete WMT'16 dataset. We speculate that as multimodal translation models are trained using larger datasets, the impact of visual information is likely to diminish.

Acknowledgements

This research is supported by the Research Promotion Project of Key Construction Discipline in Guangdong Province (2022ZDJS112).

References

- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.
- Rie Kubota Ando and Tong Zhang. 2005. [A framework for learning predictive structures from multiple tasks and unlabeled data](#). *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. [Scalable training of \$L_1\$ -regularized log-linear models](#). In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014a. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014b. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chirag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, volume 2, pages 308–327.
- Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North*, pages 4159–4170. Association for Computational Linguistics.
- Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](#). *Journal of the Association for Computing Machinery*, 28(1):114–133.
- Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.
- Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.
- Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181.
- James W. Cooley and John W. Tukey. 1965. [An algorithm for the machine calculation of complex Fourier series](#). *Mathematics of Computation*, 19(90):297–301.
- Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. An empirical study on the effectiveness of

- images in multimodal neural machine translation. *arXiv preprint arXiv:1707.00995*.
- Jean-Benoit Delbrouck and Stephane Dupont. 2017b. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.
- Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.
- Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.
- Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.
- Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.
- Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018a. [The MeMAD submission to the WMT18 multimodal translation task](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.
- Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018b. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.
- Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.
- Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. [Sheffield submissions for WMT18 multimodal translation shared task](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 624–631, Belgium, Brussels. Association for Computational Linguistics.
- Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. [On vision features in multimodal machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.
- Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai

- Zhao. 2020a. Explicit sentence compression for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8311–8318.
- Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020b. Data-dependent gaussian prior objective for language generation. In *Eighth International Conference on Learning Representations*.
- Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. [Yara parser: A fast and accurate dependency parser](#). *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.
- Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60.
- ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98.
- Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- Peter D Turney. 2000. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.
- Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. 2005. Kea: Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, pages 129–152. IGI global.
- Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. [Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with

visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. Nict-naist system for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 477–482.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9628–9635.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A. Qualitative Examples

In this appendix, we provide examples of retrieved images (Table 8), as well as translation examples for MMT with visual noise filtering (Table 9) and NMT with retrieved supplementary texts (Table 10).

Dataset	English Sentence	One of five retrieved images
Multi30k	The person in the striped shirt is mountain climbing.	
Global Voices	Now the city is under a siege from the security forces.	
WMT'16 (100k)	In the future, integration will be a topic for the whole of society even more than it is today.	
Bible	You are Yahweh, even you alone. You have made heaven. the heaven of heavens, with all their army, the earth and all things that are on it, the seas and all that is in them and you preserve them all.	
MultiUN	Development assistance cannot by itself prevent or end conflict.	

Table 8: Examples of retrieved image from different datasets. For the sentence from Multi30k dataset, our method efficiently retrieves an image that accurately represents the sentence’s content “A man is rock climbing”. For the sentence from Global Voice dataset, the retrieved image exhibits a degree of alignment with the source sentences, encompassing elements like “city”, “siege” and “forces”. However, for the sentence from WMT’16 (100k), Bible and MultiUN datasets, it becomes evident that the retrieved images offer limited relevant visual information and thus provide little assistance for translation.


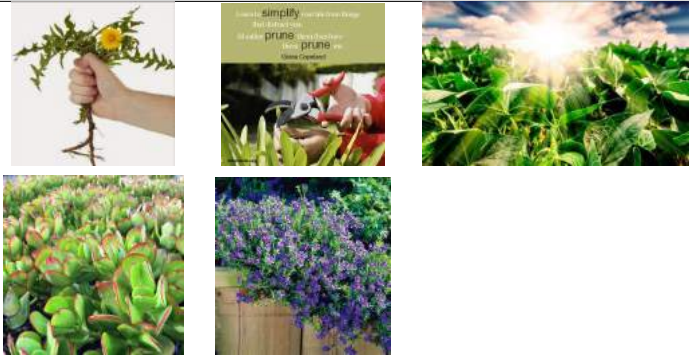
Source (En)	But he answered and said, "Every plant which my heavenly Father didn't plant will be uprooted.
Target (De)	Aber er antwortete und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte, die werden ausgereutet.
Retrieved images	
MMT with retrieved images	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht verderbte Quelle.
Retrieved images with noise image filter	
MMT with noise image filter	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte.
MMT with both noise image and region filter	Er antwortete aber und sprach: Alle Pflanzen, die mein himmlischer Vater nicht pflanzte, wird entwurzelt werden.

Table 9: A correct example generated by MMT with visual noise filtering. Due to its unique characteristics, the Bible dataset contains numerous entity words but is challenging to obtain images that effectively represent the textual content. However, visual noise filtering based on visual-text correlation can partially alleviate this situation. In this example, the filtered visual information has enabled the translation of “uprooted” to be correct.


Source (En)	Group of Asian boys wait for meat to cook over barbecue.
Target (De)	Eine Gruppe asiatischer Jungen wartet am Grill darauf, dass Fleisch gar wird.
Text-only NMT	Eine asiatische Jungen warten auf dem Fleisch, um den Grill zu kochen.
Retrieved images	
MMT with retrieved images	Eine Gruppe von asiatischen Jungen wartet darauf, um Fleisch zu grillen.
Retrieved supplementary texts	<ol style="list-style-type: none"> (1) Delivery is hardly limited to pizza at this point; everything from sushi to barbecue seems available as a to-go order. (2) While the savory aroma of barbecue filled the air, friends and family gathered around the grill, eagerly sharing stories and laughter as they waited for the delicious meal to be ready. (3) As the sun dipped below the horizon, our group of friends decided to have a barbecue in the backyard, lighting up the grill and eagerly waiting for the charcoal to heat up so that we could start cooking our favorite dishes. (4) At the lively outdoor barbecue gathering, a diverse group of friends, including a talented Asian chef, couldn't wait to cook up a mouthwatering feast. (5) While the enthusiastic Asian group gathered around the barbecue, they took turns to cook their favorite dishes, making everyone else eagerly wait in anticipation of the delicious meal.
MMT with retrieved supplementary texts	Eine Gruppe von asiatischen Jungen wartet darauf, dass Fleisch über Grill zukochen.

Table 10: A correct example generated by NMT with retrieved supplementary texts. In this example, in contrast to text-only NMT without any supplementary information, visual information and supplementary text information play an equivalent role, correctly translating “Group” to “Gruppe”. Benefiting from the rich information in the supplementary text, the NMT with retrieved supplementary text achieves more accurate translations compared to MMT with retrieved images.

Exploring the Potential of Large Language Models in Adaptive Machine Translation for Generic Text and Subtitles

Abdelhadi Soudi¹, Mohamed Hannani², Kristof Van Laerhoven²,
Eleftherios Avramidis³

¹ Ecole Nationale Supérieure des Mines de Rabat, Morocco
asoudi@enim.ac.ma

²University of Siegen, Germany

mohamed_hannani@yahoo.com, kvl@eti.uni-siegen.de

³ German Research Center for Artificial Intelligence, Germany
eleftherios.avramidis@dfki.de

Abstract

This paper investigates the potential of contextual learning for adaptive real-time machine translation (MT) using Large Language Models (LLMs) in the context of subtitles and generic text with fuzzy matches. By using a strategy based on prompt composition and dynamic retrieval of fuzzy matches, we achieved improvements in the translation quality compared to previous work. Unlike static selection, which may not adequately meet all request sentences, our enhanced methodology allows for dynamic adaptation based on user input. It was also shown that LLMs and Encoder-Decoder models achieve better results with generic texts than with subtitles for the language pairs English-to-Arabic (En→Ar) and English-to-French (En→Fr). Experiments on datasets with different sizes for En→Ar subtitles indicate that the bigger is not really the better. Our experiments on subtitles support results from previous work on generic text that LLMs are capable of adapting to In-Context learning with few-shot, outperforming Encoder-Decoder MT models and that the combination of LLMs and Encoder-Decoder models improves the quality of the translation.

Keywords: Large Language Models, Adaptive MT, Prompt Composition, LangChain, Generic Text, Subtitles.

1. Introduction

While Large Language Models (LLMs), such as GPT, Llama 2, and Falcon (Penedo et al., 2023) have made progress in tackling a variety of language-related tasks, MT is not a simple sequence-to-sequence task. It involves the complicated task of preserving the subtleties, idiomatic expressions, and distinctive stylistic features that characterize human languages.

LLMs, including but not limited to GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), Falcon (Penedo et al., 2023), and LLaMA (Touvron et al., 2023), have been designed to predict the subsequent word in a sequence based on the context. Brown et al. (2020) and Ouyang et al. (2022) introduced the concept of “In-Context learning” to describe a scenario where a pre-trained language model, during inference, assimilates specific input-output text generation patterns without the need for further fine-tuning. Their research highlighted that autoregressive LLMs, such as GPT-3, exhibit strong performance across diverse tasks, including zero-shot, one-shot, and few-shot In-Context learning without requiring updates to their weights. Instead of directly instructing the model to perform a particular task, input data can be enriched with relevant examples to facilitate the model’s adaptation. The core principle of In-Context learning revolves around learning from analogies embedded within

demonstrations (Dong et al., 2022).

A key advantage of adaptive MT, a paradigm aimed at enhancing translation by tailoring it to specific domains, genres, or styles, is its ability to achieve domain-specific translation goals without the resource-intensive processes of model training and fine-tuning.

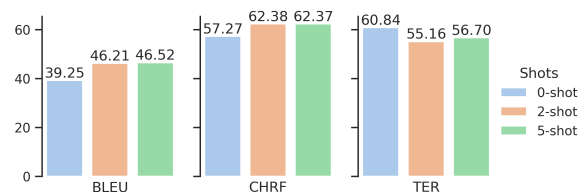


Figure 1: Evaluation results of ChatGPT 3.5 Turbo on TICO 19 for En→Ar language pair, with zero-shot, 2-shot and 5-shot fuzzy matches.

The results in Figure 1 show the performance of GPT-3.5 Turbo with zero-shot, 2-shot, and 5-shot fuzzy matches translation. When employing fuzzy matches, translation quality metrics such as BLEU and TER show substantial improvements, underlining the effectiveness of this technique in enhancing translation accuracy and fluency.

In this work, our particular emphasis lies in harnessing the capabilities of GPT-3.5 Turbo by OpenAI with In-Context examples. We examine the subtleties of adapting machine translation to domain-

specific requirements, using the TICO-19 dataset (Generic Text) and TED Talks 2013 dataset (Subtitles). By using our strategy based on prompt composition and dynamic retrieval of fuzzy matches, we report on experimental results for the language pairs English-to-Arabic (En→Ar) and English-to-French (En→Fr). To evaluate the effect of the dataset size on the translation quality of En→Ar generic text and En→Ar subtitles, we conduct experiments on different sizes for the same dataset type. An evaluation of the performance of LLMs and DeepL (Encoder-Decoder model) is also provided.

In the following sections, we provide an overview of the related work (section 2), the methodology (section 3), the experimental setup (section 4), and the results (section 5).

2. Related Work

Prior studies have focused on the application of neural language models in MT, encompassing zero-shot (Wang et al., 2021) and few-shot (Vilar et al., 2022) In-Context learning. Other researchers have proposed leveraging LLMs to generate synthetic domain-specific data to facilitate MT domain adaptation (Moslem et al., 2022). Recent research by Agrawal et al. (2022) and Zhang et al. (2023) have demonstrated the critical role of In-Context example selection in enhancing the quality of MT when employing LLMs.

One way to improve MT quality is the incorporation of fuzzy matches (Knowles and Koehn (2018), Bulte and Tezcan (2019b) and Xu et al. (2020)). Fuzzy matches comprise similar segments of previously approved translations stored within parallel datasets collected with computer-assisted translation tools, commonly referred to as translation memories (TMs). Knowles et al. (2018) showed that the utilization of fuzzy matches could enhance the quality of neural MT (NMT) systems by up to 2 BLEU points. Likewise, Bulte and Tezcan (2019b) demonstrated that fuzzy matches could enhance the consistency of MT systems, even in cases where these matches were not entirely precise (Bulte and Tezcan, 2019a). In the same vein, Moslem et al. (2022) focused on the prospect of compelling the translation of new sentence pairs to conform to the fuzzy matches found within the context dataset. They demonstrated that this approach yielded improvements in MT quality, particularly for challenging sentences.

To select fuzzy matches, Moslem et al. (2022) employed an embedding similarity-based retrieval method. This technique is initiated by generating embeddings for each sentence within the TM. These embeddings represent sentences in dense numerical forms, encapsulating their seman-

tic essence. Subsequently, the system retrieves fuzzy matches for a new sentence by identifying TM sentences with the most analogous embeddings. Previous research has established the superiority of embedding similarity-based retrieval over alternative methods, such as Edit Distance (Hosseini et al. (2020)).

Within the few-shot setting, the MT system is provided with a limited number of translated examples (e.g., 2 or 5 fuzzy matches) to assist in generating a translation for a new sentence. This stands in contrast to the zero-shot where the MT system is solely equipped with the source sentence. Moslem et al. (2022) pointed out that incorporating fuzzy matches through few-shot translation prompts could further improve the MT quality. This is attributed to fuzzy matches equipping the MT system with additional insights into the desired translation’s style and tone. In the same context, Wang et al. (2021) proposed an embedding similarity-based retrieval algorithm that improved the selection of fuzzy matches, hence the quality of the translation. Knowles and Littell (2022) investigated the role of fuzzy matches in improving low-resource language translation. Their findings underscored the potential for leveraging fuzzy matches to significantly enhance the translation of low-resource language pairs.

3. Methodology

Before the inference phase, we leverage the Sentence-Transformer model to compute embeddings for the segments of the source language (English) streamlining the retrieval of similar sentences using the Facebook AI Similarity Search (FAISS) index system (Douze et al., 2024). This technique enables us to construct contextually rich prompts, allowing the GPT-3.5 Turbo model to follow the style present in domain-specific examples. The performance of LLMs is compared with that of DeepL for the En→Fr language pair. We also evaluate the combination of both LLMs and Encoder-Decoder systems on the En→Fr language pair for the translation of subtitles.

Our particular areas of interest revolve around assessing the efficiency of LLMs in performing the following tasks without requiring additional training:

1. Adapting newly generated translations to seamlessly match the terminology and style in the context,
2. Using translations generated by Encoder-Decoder MT systems as fuzzy matches to further enhance the performance of LLMs,
3. Emphasizing the significance of prompt engineering in improving the capabilities of LLMs by using relevant translation examples for the given sentence request.

3.1. Retrieval of Fuzzy Matches

To efficiently retrieve fuzzy matches for a given input sentence, we use the FAISS system. The latter provides a variety of data structures and algorithms for efficient similarity search, and we have chosen to use the IndexFlatL2 index, which performs an exhaustive search of the index to find the nearest neighbors.

To generate the FAISS index, we first use the Sentence-Transformer model to generate embeddings for each sentence in our preprocessed dataset. Sentence embeddings are dense numerical representations of sentences that capture their semantic meaning and contextual nuances. Once the sentence embeddings are generated for all of the sentences in our dataset, the FAISS index can be created. This process involves the following steps:

1. Loading the sentence embeddings into FAISS,
2. Configuring the FAISS index with the desired parameters, such as the choice of index type and the dimensionality of the embeddings,
3. Building the FAISS index for the whole corpus.

Once the FAISS index is built, it can be used to retrieve fuzzy matches for a given input sentence. To do so, we simply compute the cosine similarity between the input sentence embedding and all of the embeddings in the index. The sentences with the highest cosine similarities are the fuzzy matches for the input sentence. The fuzzy matches are then used to compose context-aware prompts for the GPT-3.5 Turbo model. These prompts provide GPT-3.5 Turbo with additional information about the desired translation, which can help it generate more accurate translations.

3.2. Prompt Composition

For each translation request, our approach leveraged the FAISS index to retrieve the top-k closest sentence embeddings from the domain-specific dataset. The retrieved sentences serve as the foundation for constructing contextually rich prompts for the LLM model.

To facilitate prompt composition and enhance translation quality, we integrated LangChain¹ into our system. LangChain serves as a framework designed for the development of applications leveraging large language models. Its primary objective is to empower developers with the seamless integration of diverse data sources and the facilitation of interactions with other applications. To achieve this goal, LangChain offers modular components, serving as abstractions, and customizable

¹<https://www.langchain.com/>

```
Prompt: EN-AR zero-shot translation

<SystemMessage>
English: HumanMessage<source_segment>
Arabic: → AIMessage<predicted_segment>
```

Figure 2: Zero-shot translation prompt

```
Prompt: EN-AR 2-shot translation

<SystemMessage>
English: HumanMessage<source_fuzzy_match_1>
Arabic: AIMessage<g_truth_fuzzy_match_1>

English: HumanMessage<source_fuzzy_match_1>
Arabic: AIMessage<g_truth_fuzzy_match_1>

English: HumanMessage<source_segment>
Arabic: → AIMessage<predicted_segment>
```

Figure 3: 2-shot translation prompt

use case-specific pipelines, referred to as chains. We utilized the following Langchain's components settings:

- **SystemMessage**: A Message for priming AI behavior, usually passed in as the first of a sequence of input messages. This component plays a pivotal role in guiding the LLM model to follow the desired style and context for subtitle translation tasks. It acts as a foundational prompt template, providing a structured starting point for generating high-quality translations. We set the component to: "Act like a good translator from English to <target_language>. Translate the following English sentence into <target_language>".

- **HumanMessage** and **AIMessage** are Built upon the SystemMessage. We employed a combination of stacked HumanMessage and AIMessage. These messages were carefully crafted to maintain a conversational flow and ensure that the GPT model understands the user's request.

- The last **HumanMessage** in the sequence is the user's sentence request, serving as the input for the translation task.

Figures 2 and 3 show the distinction between zero-shot and few-shot translation prompts. In the zero-shot scenario, only the source sentence and language specifications are provided, prompting the model to autonomously generate the translation guided by the SystemMessage only. Conversely, the few-shot prompt incorporates translation examples, guiding the style of the generated output.

In the evaluation phase of the translation system, we leveraged the above chat message format to interact with the GPT-3.5 Turbo model effectively. Each translation request is encapsulated within a chat message, providing a structured way to communicate with the model. The chat message typically consists of a series of messages, including

a SystemMessage, AIMessages, and a final UserMessage. The SystemMessage sets the context and instructs the model to perform as a skilled translator. AIMessages provide additional guidance, context, or clarifications as needed. The UserMessage encapsulates the user’s specific translation request, serving as the input for the model. By crafting messages in this manner, we ensure that the GPT model receives a clear context.

4. Experimental Setup

In the course of our experimentation, we employed the GPT-3.5 Turbo model through its official OpenAI API ², setting parameters to top-p 1 with a temperature of 0.3 for our translation tasks (Table 1). The choice of these parameters was made deliberately to optimize model performance on the translation task.

Parameters	temperature	top_p
Values	0.3	1

Table 1: GPT-3.5 Turbo parameters with OpenAI API

To simulate a document-level scenario emulating real-world generic text translation tasks, we leveraged the TICO-19 dataset (Anastasopoulos et al., 2020), which contains 3,070 distinct segments for the language pairs under study. English is used as the source language, while Arabic and French as target languages.

With respect to the subtitle translation task, our dataset is taken from TED Talks 2013, commonly known as the Web Inventory (Cettolo et al., 2012), is composed of roughly 150,000 distinct segments for each language pair. The translations are available in more than 109 languages. For the purposes of our study, we chose portions that are relatively in the same TICO-19 domain. We strategically selected three portion sizes (3,200, 6,200, and 9,200 segments) for our experiments to be able to compare the performance with regard to the type of text being translated (generic text or subtitles) as well as to the dataset sizes.

In the following section, we evaluate our method on generic text and subtitles datasets in different portion sizes and compare our results with related work.

5. Experiments and Results

5.1. Generic Text

Previous work by Moslem et al. (2023) has shown the importance of LLMs in adaptive machine trans-

²<https://openai.com/>

lation for In-Context learning using the TICO-19 dataset. In their work, they ran extensive experiments on various language pairs and different types of models (LLMs and Encoder-Decoder models). Table 2 shows the results they obtained for English to Arabic language pair with GPT-3.5 Turbo.

Context	spBLEU [↑]	CHRF [↑]	TER [↓]
Our Results on 1500 Segments			
Zero-shot	37.42	55.48	62.8
Fuzzy 2-shot	45.52	61.7	56.26
Fuzzy 5-shot	46.43	62.41	55.98
Our Results on Full dataset			
Zero-shot	39.25	57.27	60.84
Fuzzy 2-shot	46.21	62.38	55.16
Fuzzy 5-shot	46.52	62.37	56.7
Moslem et al. (2023)’s results on Full dataset			
Zero-shot	38.06	56.35	61.34
Fuzzy 2-shot	46.04	62.18	55.03

Table 2: Our GPT-3.5 Turbo model evaluation results on TICO-19 English-to-Arabic dataset compared to those of Moslem et al. (2023).

With the same settings and parameters for the model and dataset (size and language pair), but with improvement in the prompt composition and selection of the fuzzy match (as explained in sections 3.1 and 3.2), we achieved a significant improvement in the BLEU score as is shown in Table 2 above. For instance, an improvement of 1.19 for zero-shot and 0.17 for 2-shot.

Even in the case of zero-shot translation, notable improvement in BLEU score values is achieved, which is attributed to the effective utilization of prompt composition techniques, using LangChain which helps improve the results.

With the incorporation of fuzzy matches as context for the translation task (with 2 or 5 shots), we can also see an improvement, thanks to the fuzzy matches selection as explained in the previous sections. This technique selects the most contextually relevant and representative shots to the user request on the fly instead of using static fuzzy matches for all sentences as is the case in the work of Moslem et al. (2023). In their work, when composing the prompt, the fuzzy matches were retrieved out of 10 fuzzy matches which were statistically stored as the 10-closest sentences for the overall dataset³.

To further illustrate our strategy based on prompt composition and dynamic retrieval of fuzzy matches, we conducted experiments on English-to-French language pair. As can be seen in Table 3 below, the resulting translation performance was

³<https://github.com/yomoslem/Adaptive-MT-LLM/blob/main/MT/ChatGPT-BatchTranslation.ipynb>

shown to improve in all shot settings for this language pair. We find an improvement of 0.9 for 0-shot setting over [Moslem et al. \(2023\)](#)'s results.

Context	spBLEU \uparrow	CHRF \uparrow	TER \downarrow
Our Results			
Zero-shot	47.75	67.41	47.86
Fuzzy 2-shot	50.59	69.28	45.41
Fuzzy 5-shot	53.68	71.3	42.56
Moslem et al. (2023)'s results			
Zero-shot	46.85	66.75	48.31
Fuzzy 2-shot	49.88	68.33	46.27

Table 3: Our GPT-3.5 Turbo model evaluation results on TICO-19 English-to-French dataset compared to those of [Moslem et al. \(2023\)](#).

It is worth noting that in both [Moslem et al. \(2023\)](#)'s work and ours, the results for the language pair English-Arabic are lower than those of the language pair English-French (Tables 2 and 3).

Our results show the effectiveness of both prompt composition and fuzzy match selection techniques as well as the FAISS index for efficient and fast translation quality.

5.2. Subtitles

Subtitles are short text lines usually at the bottom of the screen that allows the viewer of a film or TV program to follow the dialogue(s) without understanding the audio. We distinguish between same-language subtitles and cross-language subtitles. Same-language subtitles are usually targeted at hearing-impaired viewers or added for educational purposes, while cross-language subtitles enable viewers to enjoy a film in a language different from the audio. Same-language subtitles for hearing-impaired viewers need to include a written or a graphical representation of sounds (e.g. approaching footsteps) which hearing viewers do not need even if they do not understand the original language. Subtitles are typically limited to two rows of text with up to 37 characters on each row. They are displayed on the screen between 3 and 7 seconds. More details about the characteristics of subtitles can be found in [Jorge and Remael \(2007\)](#).

In this section, we report on experiments conducted on the TED Talks 2013 dataset. These experiments encompass various dataset sizes and are run on English-to-Arabic and English-to-French language pairs.

We conducted experiments on 3200 segments of the English-to-Arabic language pair. We found significant improvements in the BLEU score across three experiments as shown in Table 4. The experiments' settings are zero, 2, and 5 fuzzy matches. We notice that the translation performance was

shown to improve appreciably with the 5 fuzzy matches setting.

Context	spBLEU \uparrow	CHRF \uparrow	TER \downarrow
3200 Segments			
Zero-shot	21.31	44.03	78.3
Fuzzy 2-shot	22.75	45.21	76.69
Fuzzy 5-shot	24.26	46.23	75.89
6200 Segments			
Zero-shot	22.74	44.85	76.99
Fuzzy 2-shot	22.85	44.9	76.79
Fuzzy 5-shot	24.87	44.93	76.79
9200 Segments			
Zero-shot	22.97	45.14	76.4
Fuzzy 2-shot	22.97	45.14	76.35
Fuzzy 5-shot	24.98	45.12	76.27

Table 4: GPT-3.5 Turbo model evaluation results on English-to-Arabic Ted Talks 2013 dataset with 3200, 6200 and 9200 segments.

Interestingly, we noticed that there is a significant difference in the experimental results for the English-to-Arabic (generic text) and English-to-Arabic (subtitles). For the zero-shot setting and with approximately the same dataset sizes, the BLEU score of the TED Talks 2013 dataset on English-to-Arabic translation is 21.31 (Table 4), whereas the TICO-19 on English-to-Arabic translation has a BLEU score of 39.25 (Table 2). The difference in the results can be attributed to the dataset translation quality and type.

With the same previous experimental settings, we conducted experiments on 6200 subtitle segments. The results show a very slight improvement with increased data size (Table 4). For example, with the 3200 dataset, the BLEU score for the two-shot setting is 22.75, whereas with the 6200 dataset, it is 22.85. This means that the bigger the size is is not necessarily the better.

With the same settings, we tripled our dataset to 9200 segments and noticed a very minor improvement again as shown in Table 4 above. The small increase in the BLEU score even when doubling or tripling the dataset size may be due to the quality difference between the three dataset portions based on manual checks of samples of the dataset. We noticed that the translation quality of the first 3200 segments are better than the additional portions, which explains the slight improvement.

In order to verify the effect of the dataset size on performance, we also conducted experiments on generic text. Results on different size datasets for generic text show a significant improvement when doubling the dataset. Experiments on the full 3071 sentence pairs of the TICO-19 dataset presented in Table 2 show a significantly higher BLEU score than those obtained with roughly half the TICO-19 dataset. By way of example, we noticed an

additional gain of 1.83 in the BLEU score in the zero-shot setting (37.42 on the 1500 sub-dataset and 39.25 on the 3071 full dataset). This means that performance increases with more data in the case of generic text.

We also conducted experiments on English-to-French subtitles and compared the results obtained with those of the English-to-Arabic pair. Table 5 below presents the results of 3000 TED Talks subtitle segments. It can be seen that there is an improvement when adding more fuzzy matches.

Context	spBLEU \uparrow	CHRF \uparrow	TER \downarrow
Zero-shot	44.26	64.72	51.82
Fuzzy 2-shot	44.68	64.99	51.12
Fuzzy 5-shot	45.15	65.34	50.29

Table 5: Evaluation results on TED Talks 2013 dataset composed of 3000 sentence pairs on the English-to-French language pair with GPT-3.5 Turbo.

As we have seen in the case of English-Arabic generic text and subtitle translation, we notice that the evaluation scores of the English-French subtitles are lower than those of the English-French generic text.

In order to compare the results obtained with GPT-3.5 Turbo for the translation of subtitles of the English-to-French language pair, we conducted experiments using the DeepL Encoder-Decoder model, used as API from their official website⁴. Table 6 below shows the results of experiments run on 3000 sentence pairs of the TED Talks 2013 dataset.

spBLEU \uparrow	CHRF \uparrow	TER \downarrow
44.33	64.12	49.91

Table 6: Evaluation results on TED Talks 2013 dataset composed of 3000 sentence pairs on English-to-French language pair with DeepL model.

When used with the zero-shot setting, the Encoder-Decoder slightly outperforms LLMs as can be seen in Tables 6 and 7. However, the results of our experiments demonstrate LLMs’ capability to adapt to In-Context learning with few-shot, outperforming Encoder-Decoder MT models. By way of illustration, with a 5-shot setting, GPT-3.5 Turbo achieves an increased BLEU score of 0.82 as shown in Tables 5 and 6.

In the previous experiments, we used the fuzzy matches from the ground-truth translations. In order to see the performance of the combination of LLMs and the Encoder-Decoder model (DeepL), with fuzzy matches constructed using the predicted sentences from DeepL, we conducted experiments

⁴<https://www.deepl.com/pro-api/>

on the subtitles dataset composed of 3000 segments. The experimental results are shown in Table 7 below.

Context	spBLEU \uparrow	CHRF \uparrow	TER \downarrow
Zero-shot	44.26	64.72	51.82
Fuzzy 2-shot	45.85	65.67	48
Fuzzy 10-shot	46.01	65.74	47.75

Table 7: Evaluation results on TED Talks 2013 dataset composed of 3000 sentence pairs on English-to-French language pairs with GPT-3.5 Turbo + DeepL model.

We can see that constructing the fuzzy matches from the DeepL Encoder-Decoder model’s predictions as a context to the GPT-3.5 Turbo model can improve the quality of the translation of the source segments. By way of illustration, an improvement of 1.17 and 0.86 in the BLEU score for 2-shot and 5-shot, respectively (cf. Tables 5 and 7). This can be explained by the use of the predicted sentences (from DeepL model) to compose the prompt for the GPT-3.5 Turbo model, which supports our previous hypothesis based on manual checks of the quality of the translation in the TED Talks 2013 dataset.

6. Conclusion

This work explored GPT-3.5 Turbo’s efficiency in adaptive MT with fuzzy matches. Experimental results were provided showing the effectiveness of our technique with respect to the prompt composition and the selection of the fuzzy matches. The results of our experiments indicate LLMs’ capability to adapt to context, outperforming Encoder-Decoder MT models. Our work on subtitles corroborated results from previous work on generic text that the combination of LLMs and Encoder-Decoder models improves the quality of the translation. It was also shown that LLMs and Encoder-Decoder models achieve better results with generic texts than with subtitles for the language pairs En \rightarrow Ar and En \rightarrow Fr. Experiments using GPT-3.5 Turbo on different data sizes of English-to-Arabic subtitles indicated that the bigger is not really the better. Further research is required to validate these results and also explore the use of other LLMs in MT, especially for low-resource languages.

7. Bibliographical References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022.

- In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, et al. 2020. Tico-19: the translation initiative for covid-19. *arXiv preprint arXiv:2007.01788*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Boris Bulte and Ayşe Aişe Tezcan. 2019a. Fuzzy matches for improving the consistency of neural machine translation. *arXiv preprint arXiv:1903.11534*.
- Bram Bulte and Arda Tezcan. 2019b. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *57th Annual Meeting of the Association-for-Computational-Linguistics (ACL)*, pages 1800–1809.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Kasra Hosseini, Federico Nanni, and Mariona Coll Ardanuy. 2020. Deezy-match: A flexible deep learning approach to fuzzy string matching. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pages 62–69.
- Daz Cintas Jorge and Aline Remael. 2007. *Audio-visual translation: subtitling*. Routledge.
- Rebecca Knowles and Philipp Koehn. 2018. Fuzzy match incorporation for neural machine translation. *arXiv preprint arXiv:1806.08117*.
- Rebecca Knowles and Patrick Littell. 2022. Translation memories as baselines for low-resource machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6759–6767.
- Rebecca Knowles, John Ortega, and Philipp Koehn. 2018. A comparison of machine translation paradigms for use in black-box fuzzy-match repair. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 249–255.
- Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. *arXiv preprint arXiv:2208.05909*.
- Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refined-web dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.
- Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. 2021. Language models are good translators. *arXiv preprint arXiv:2106.13627*.

Jitao Xu, Josep-Maria Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

INCLURE: a Dataset and Toolkit for Inclusive French Translation

Paul Lerner¹, Cyril Grouin²

Sorbonne Université, CNRS, ISIR¹, Université Paris-Saclay, CNRS, LISN²
75005, Paris, France¹, 91400, Orsay, France²
lerner@isir.upmc.fr, cyril.grouin@lisn.upsaclay.fr

Abstract

Inclusive French (gender-neutral language) is a variety of French that is used to highlight awareness of gender and identity against Standard French, which enforces the use of masculine for generic usage or plural. Although widely used and challenging to a set of NLP tools, Inclusive French was very little studied in NLP. Detractors of Inclusive French argue that it is difficult to read, while its supporters argue that it provides a fairer representation of women and gender minorities. We provide INCLURE, the first large-scale parallel corpus for Standard to Inclusive French translation, and vice-versa, thus providing a “bilingual” access to French, for both detractors and supporters of Inclusive French. This corpus comes with a toolkit that can be readily applied to larger French corpora and could be extended to other languages, for which the number of inclusive varieties is growing. We also provide Fabien.ne BARThez, a sequence-to-sequence model trained on INCLURE. Apart from its direct application to translation, this model could also be used in most NLP pipelines, either as a pre-processing step to improve downstream processing or as a post-processing according to the user’s preference.

Keywords: Inclusive French, Gender-neutral Language, Parallel Corpus, Neural Machine Translation

1. Introduction

Inclusive French (gender-neutral language) is a variety of French used to highlight awareness of gender and identity (Alpheratz, 2018, 2019). Indeed, Standard French, as other languages (Hellinger and Bußmann, 2015), enforces the use of masculine for generic usage (e.g., *un doctorant se doit de publier*¹) or plural (e.g., *mon frère et ma sœur sont des doctorants*²). Inclusive French would include women in these speeches mainly in two different manners (Grouin, 2022) (see Figure 1):

1. coordination of feminine and masculine forms: *un doctorant ou une doctorante*;
2. morphological combination of masculine and feminine flectional endings (colloquially known as inclusive writing or *écriture inclusive* in French): *un.e doctorant.e*.

Although Inclusive French is prone to controversy³, several studies have found that Standard French shadows women and impacts the mental representations of the speakers (Sczesny et al., 2016). To avoid this issue, Touraille and Allasonnière-Tang

(2023) argued generalizing gender-neutral words in French by proposing a new non-binary inflexional ending⁴. Other studies focus on the perception of sentences written in inclusive French, highlighting that feminization and coordination of feminine and masculine forms are better accepted than other processes (Delaborde et al., 2021). We choose not to choose. With the INCLURE dataset and toolkit, anyone should be able to translate⁵ from Standard to Inclusive French, and vice-versa, thus providing “bilingual” access to French.

Inclusive French was very little studied in the NLP community. To our knowledge, this is only the *second* study of Inclusive French, after the exploratory study of Grouin (2022), and the first for Inclusive French Translation. We propose:

- INCLURE, a dataset of 69K aligned sentences (bitext)⁶;
- Fabien.ne BARThez, a sequence-to-sequence model trained on INCLURE, able to translate from Standard to Inclusive French, and vice-versa⁷.

¹Meaning “a PhD Student must publish”. The feminine form of *un doctorant* is *une doctorante*.

²Meaning “My brother and sister are PhD students”.

³The *Académie Française* considers that Inclusive French puts the French language “in mortal peril” and wishes to ban its usage (Grouin, 2022). The *Rassemblement National* of Marine Le Pen shares this opinion and proposed another law to ban Inclusive French on October 12th, 2023 https://www.assemblee-nationale.fr/dyn/16/textes/l16b0777_proposition-loi.

⁴The authors proposed to use the final vowel “-i” to produce non-binary words: *li doctoranti est heurési* meaning “the Ph.D. student is happy”.

⁵We use the term *translate* for lack of a better one, but the problem is much simpler than translating from French to any other language. Standard and Inclusive French are but varieties of the same language, the grammar is identical. This will be further demonstrated in Section 5.

⁶https://huggingface.co/datasets/PaulLerner/oscar_inclure

⁷https://huggingface.co/PaulLerner/fabien.ne_barthez

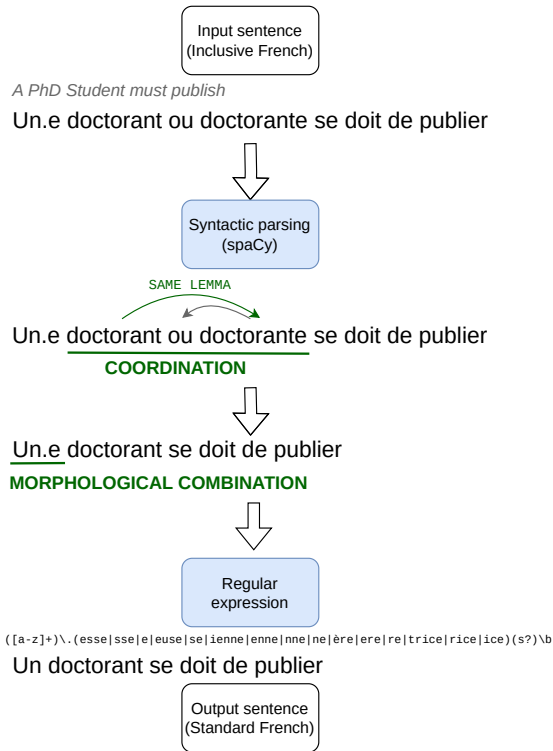


Figure 2: Simplified diagram of our rule-based system for Inclusive to Standard French translation, used to generate the INCLURE parallel corpus.

This task is best learned automatically from data, as described in Section 4.

More precisely, we focus on the two main processes of Inclusive French, which are easily detected automatically (see Figure 2):

1. coordination: e.g., *un doctorant ou une doctorante* is detected through a syntactic analysis: the head of *doctorante* is *doctorant*, but both share the same lemma;
2. morphological combination: e.g. *les doctorant.e.s* is detected through a regular expression.

The regular expression is built around common French feminine suffixes: (esse|sse|e|euse|se|ienne|enne|nne|ne|ère|ere|re|trice|rice|ice). Because Inclusive French is yet unstandardized, we see several variants of the same suffix, e.g., *trice|rice|ice*. These might occur in *auteur.trice*, *auteur.rice*, or *auteur.ice* (all meaning “author”). Likewise, the ordering of the masculine and feminine suffix is variable; both *auteur.trice* and *autrice.teur* are acceptable. Therefore, the core of our regex substitution method lies in two regexes:

- $\langle \text{FEM} \rangle s? \backslash . ([a-z]+) \backslash b$, when the feminine suffix comes before the separating dot;

	INCLURE		IFC	
	I	S	I	S
Length	33.0	29.4	33.9	32.1
Vocabulary	70,200	66,500	899	860
TTR	0.93	0.90	0.91	0.87

Table 1: Average sentence length, vocabulary size, and type-to-token ratio (TTR) of INCLURE and the Inclusive French Corpus (IFC), in the Inclusive (I) or Standard (S) version.

- $([a-z]+) \backslash . \langle \text{FEM} \rangle (s?) \backslash b$, when the feminine suffix comes after.

Where $\langle \text{FEM} \rangle$ stands for the feminine suffixes listed above. Parenthesis shows the captured sections of the string that are substituted back (e.g., *teur* in *autrice.teur* to obtain *auteur*, the masculine form). *s* marks the plural. Instead of $[a-z]$, we use all lowercase French letters, including accents and diacritics ($[a-z\grave{\text{a}}\grave{\text{a}}\acute{\text{e}}\acute{\text{e}}\grave{\text{e}}\grave{\text{i}}\grave{\text{i}}\grave{\text{o}}\grave{\text{u}}\grave{\text{u}}\grave{\text{y}}\grave{\text{ç}}\grave{\text{æ}}]$), but left them out above to improve readability.

Note that the interpunct (“.”, U+00B7) is frequently used as a separating sign instead of the dot (“.”, U+002E). However, the interpunct is absent of BARThez vocabulary (Eddine et al., 2021), which we use as a foundation model for our translation model (Section 4). Therefore, all interpuncts between two lowercase letters are replaced by dots in preprocessing.

3.2. Implementation

Syntactic dependency parsing, lemmatization, and morphological analysis are done using spaCy, more precisely the *fr_dep_news_trf* model, based on CamemBERT (Martin et al., 2020), which is pre-trained on OSCAR 2019 (Suárez et al., 2019) and fine-tuned on the Sequoia Corpus (Candito et al., 2014). We use a single NVIDIA V100 GPU with 32GB of memory to process a subset of OSCAR 22.01 in 20 hours.

Our code is available so that INCLURE can be easily extended to larger corpora and other languages.

3.3. Processing OSCAR

A random 1.3% of French OSCAR 22.01 was processed, that is 681K documents of a total 2.29M sentences. Our system estimates that 0.3% of these sentences are Inclusive French, yielding 69K aligned sentences (bitext) in Standard and Inclusive French. We denote the resulting dataset INCLURE.

The dataset has a total vocabulary of 70,200 different words in its original Inclusive French and a smaller 66,500 words in the translated Standard French, as words have fewer inflected forms in Standard French. Likewise, we find Standard French

sentences to be shorter and with a smaller type-to-token ratio. These statistics are summarized in Table 1.

The dataset is split randomly into three subsets: train (90%), validation (5%), and test (5%).

We show two random examples of the test set, for each Inclusive French process:

1. coordination: *Toutes les informations utiles sur la sécurité des données et les éventuels risques pour la sécurité, sur le type d'enregistrement des données, leur étendue et leur conservation, et sur les droits des clientes et clients, doivent être communiquées.* \iff *Toutes les informations utiles sur la sécurité des données et les éventuels risques pour la sécurité, sur le type d'enregistrement des données, leur étendue et leur conservation, et sur les droits des clients, doivent être communiquées.*¹¹
2. morphological combination: *Le message est clair : ces organisations et personnalités sont accusé.e.s de complicité dans les attentats commis ces dernières semaines.* \iff *Le message est clair : ces organisations et personnalités sont accusés de complicité dans les attentats commis ces dernières semaines*¹²

4. Inclusive French Translation with Fabien.ne BARThez

4.1. Method

We adopt the now-standard learning method to translate end-to-end with a sequence-to-sequence model (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), in either translation direction, while our main interest lies in the Standard to Inclusive direction.

The Transformer architecture, now more widely known for large language models, was originally proposed for translation and is well-suited for the task (Vaswani et al., 2017). We leverage the BARThez model of Eddine et al. (2021), a sequence-to-sequence model of 139M parameters¹³ pre-trained to reconstruct a corrupted input, in the manner of BART (Lewis et al., 2020), but

¹¹Meaning “All relevant information on data security and possible security risks, on the type of data storage, its scope and retention, and on customer rights, must be provided.”

¹²Meaning “The message is clear: these organizations and personalities are accused of complicity in the attacks of recent weeks.”

¹³Eddine et al. (2021) report 165M parameters but we find 139M in their released model. The embedding layer of 38M parameters is tied to the output layer, counting it twice would result in 178M parameters.

for French instead of English. BARThez was pre-trained on 66 GB of French raw text from diverse sources, mostly from CommonCrawl. It uses the SentencePiece tokenizer (Kudo and Richardson, 2018) trained on a 10 GB random sample from their pre-training corpus. We leave studies on the impact of the vocabulary and tokenizer for future work.

Although the training data differs, we fine-tune BARThez using the same loss function as for its pre-training, i.e., minimizing the cross-entropy between the predicted output and the ground truth. Each prediction is conditioned on the whole input and the preceding output tokens, using teacher forcing as systematically done with Transformers. We note this fine-tuned model Fabien.ne BARThez.

4.2. Implementation and Hyperparameters

We use the same hyperparameters for both translation directions. The model is trained using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5×10^{-5} linearly decreasing for a maximum of 10K steps if training is not interrupted before, according to the validation loss. At inference, decoding is done using greedy search as we have found that beam search decreased BLEU on the validation set.

We use a single NVIDIA V100 GPU with 32GB of memory holding a batch of 128 aligned sentences. In both translation directions, models start overfitting, and training is interrupted after 3K steps (\approx 6 epochs), after about an hour of training.

Our implementation is based upon Transformers (Wolf et al., 2020), itself built upon PyTorch (Paszke et al., 2019). Our code is freely available to ensure the reproducibility of our results.

5. Results

5.1. Evaluation Data and Metric

In addition to the IID test set of INCLURE, we evaluate the out-of-domain (OOD) performance of Fabien.ne BARThez using the Inclusive French Corpus of Grouin (2022). Indeed, this corpus mostly contains transcripts of political speeches, whose oral style differs from the text typically found in OS-CAR/CommonCrawl. Exceptions are six examples used to illustrate the use of the inclusive neutralization process described by Alpheratz (2019). These six examples were written by Grouin (2022) to complete the coverage of their corpus, as they could not find the natural occurrence of this process, which hints at its rareness. We will return to these examples in Section 6.

As for INCLURE, all separating signs of Inclusive French are normalized to use a standard dot (“.”),

Model	IID	OOD
Identity (baseline)	76.30	79.74
Fabien.ne BARThez	92.83	83.05

Table 2: Main results: BLEU scores from Standard to Inclusive French. IID: results on the test set of INCLURE, after training and tuning hyperparameters on the dedicated IID subsets. OOD: out-of-domain results, without fine-tuning or hyperparameter-tuning on the Inclusive French Corpus.

U+002E), to ease evaluation. Note that the corpus of Grouin (2022) originally contained various separating signs in addition to the dot and inter-punct, such as the slash, dash, and parenthesis. Moreover, Grouin (2022) kept the demonyms coordination (e.g. *les Martiniquaises et les Martiniquais*, which refers to Martinicans) in the Standard version of the corpus, as they are a kind of named entity. We remove them from the Standard version of the corpus as we are more interested in translation than named entity recognition. Additionally, we segment the corpus in sentences. This is easily done automatically as there is a 1-1 mapping between Standard and Inclusive French sentences, in the same order. We filtered out identical sentences in both varieties (as some documents contained mixed varieties) to arrive at 72 aligned sentences.

The dataset has a total vocabulary of 899 different words in its original Inclusive French and a smaller 860 words in the translated Standard French, similarly to INCLURE. Again, Standard French sentences are shorter and have a smaller type-to-token ratio. These statistics are summarized in Table 1.

Quantitative evaluation is done using BLEU (Papineni et al., 2002) implemented with SacreBLEU¹⁴ (Post, 2018). We leave the study of other metrics for translating Inclusive French to future work, as they would require collecting human judgments.

5.2. From Standard to Inclusive French

Our main results, translating from Standard to Inclusive French, are reported in Table 2. As both varieties of French are close, we use as a baseline the identity function, i.e., simply computing the BLEU score between the Standard French input and Inclusive French ground truth. This baseline, or lower bound, gives very high BLEU scores, between 76 and 80, depending on the evaluation corpus.

Fabien.ne BARThez nevertheless largely outperforms the baseline, on both the IID test set and the OOD corpus, although no fine-tuning or hyperparameter-tuning was done on the latter. We

¹⁴`nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

Model	IID	OOD
Identity (baseline)	77.12	79.89
Rule-based	–	86.63
Fabien.ne BARThez	96.07	94.60

Table 3: Additional results: BLEU scores from Inclusive to Standard French.

find, however, a 10 absolute BLEU point gap between the two corpora, which would suggest a poorer performance of our model on the OOD corpus. Our qualitative analysis reveals, however, that most OOD examples with relatively modest BLEU scores are semantically equivalent, because of the limitations of the surface metric that is BLEU. Take for example the ground-truth *Indemnités d’élue plafonnées au salaire médian.*¹⁵, for which our model provided *Indemnités d’élue et d’élue plafonnées au salaire médian.*, preferring the coordination process over the morphological combination process, and scoring only 51 BLEU. It is even worse for *Révocabilité des élu.e.s.*¹⁶ vs. *Révocabilité des élues et élus.*, which scores only 13 BLEU, despite being equivalent. Likewise, while the ordering of the feminine (*élues*) and masculine (*élus*) does not matter, *Révocabilité des élues et élus.* vs. *Révocabilité des élus et élues.* would only score 21 BLEU.

Furthermore, Inclusive French is sometime inconsistent, especially in its oral form present in the Inclusive French Corpus. For example, one speech begins with *Tous ceux que je n’ai pu voir au-cours de cette brève visite*¹⁷ while our model correctly predicts *Tous ceux et celles que je n’ai pu voir au-cours de cette brève visite.*

We will see in the next section that BLEU is better suited to evaluate Standard French outputs, where our model achieves nearly perfect BLEU scores on both the IID and OOD evaluation sets.

5.3. From Inclusive to Standard French

Although our main research interest lies in the Standard to Inclusive direction, we study in this section the opposite direction, both for completeness but also to demonstrate that our model generalizes beyond learning the inverse function of our rule-based system, which generated the training data (cf. Section 3.1). BLEU scores are reported in Table 3. In addition to the Identity baseline, we also report the performance of our rule-based system, which generated the INCLURE corpus. This system is, therefore, not evaluated on the IID subset where

¹⁵Meaning “Elected representatives’ allowances capped at median salary.”

¹⁶Meaning “Revocability of elected representatives.”

¹⁷Meaning “All those I didn’t get to see during this brief visit”.

	INCLURE	IFC
F M (<i>toutes et tous</i>)	51%	93%
M F (<i>tous et toutes</i>)	49%	7%

Table 4: Statistics of the gender ordering in coordinations, on both INCLURE and the Inclusive French Corpus (IFC).

it should get 100 BLEU. Because it was designed to be precise, sometimes at the expense of recall, it does not systematically detect Inclusive French in the OOD evaluation set. In this case, we fall back to the Identity baseline (i.e., compute the BLEU between the Inclusive French input and the Standard French ground truth).

The rule-based system outperforms the Identity baseline but is largely inferior to Fabien.ne BARThez, which achieves near-perfect BLEU scores on both the IID and OOD evaluation sets, thus demonstrating its generalization capacities. Unlike the Standard to Inclusive direction, BLEU is reasonably well-suited to compare Standard French outputs to the ground truth. Coming back to our earlier examples, our model correctly predicts *Indemnités d'élus plafonnées au salaire médian* and *Révocabilité des élus*, which perfectly match the ground truth.

Again, in the Inclusive to Standard direction, the irregularities of Inclusive French are smoothed out. For example, *Tous ceux que je n'ai pu voir au cours de cette brève visite [...]* is correctly predicted, which explains the high BLEU scores.

6. Discussion

Language fixation Since the inclusive French language is constantly evolving, offering a variety of processes, we have not yet observed a language fixation of phrases produced by coordinating feminine and masculine words. In the INCLURE corpus, we found about as many female-male coordinations as male-female coordinations (see Table 4). Nevertheless, we observed a majority of female-male coordinations (93%) in the IFC corpus. Despite its low number of examples, we hypothesize that political discourse mainly uses female-male coordination to highlight women for political reasons, fixing *de facto* those phrases. Adopting a linguistic point of view, we may consider that using female words first makes it more distinctive from standard French which uses male words to encompass both men and women (*bonjour à toutes et à tous* vs. *bonjour à tous*¹⁸).

¹⁸Respectively “Good morning to all (women) and to all (men)” vs. “Good morning to all (men, including women)”

Inferring Feminization We have focused on the two main phenomenons of Inclusive French, coordination and morphological combination, which counteract Standard French’s use of masculine for generic usage or plural. However, another aspect of Inclusive French is the feminization of nouns that refer to women, particularly job titles. The IFC corpus contains a few of these examples, where feminization must be inferred from the gender of the name, e.g., *Giorgia Marras, illustrateur et auteur de bande dessinée, est née à Gênes en Italie, en 1988*¹⁹ must be translated to *Giorgia Marras, illustratrice et auteure de bande dessinée, est née à Gênes en Italie, en 1988* because *Giorgia Marras* is a woman, which may be inferred from her name.

Our model cannot infer this, because such examples are absent from INCLURE. We leave this for future work. Wikidata may be a useful resource for this, as it currently holds 52K entities that have different feminine and masculine labels in French, e.g., Q644687 *illustrateur* or *illustratrice*²⁰.

Morphological Neutralization As mentioned in Section 5.1, the IFC corpus of Grouin (2022) contains six synthetic examples, based on the work of Alpheratz (2019), to cover another rare process of Inclusive French: morphological neutralization. It consists in creating new neutral lexical units (e.g. *frœur*, which means both *frère* or *sœur*) or new inflected forms (e.g. *députæs* instead of *député.es*). Our model did not learn those processes either, as they are absent from INCLURE. However, we believe it may be addressed as a post-processing step according to the user’s preference (e.g., replacing *é.es* with *æs*). The same could be said about non-binary markers (e.g. *député.e.x*²¹).

Rare words Another limitation of our model, which we have observed on the OOD evaluation set, is its brittleness to rare words. For example, a speech beginning with *Martiniquais [...]* (addressing to Martinicans) is automatically translated to *Martiniquais, Martiniciennes [...]* instead of *Martiniquaises*, as *ienne* is a common feminine suffix.

7. Conclusion

This paper tackles the translation from Inclusive French to Standard French, and vice-versa. Inclusive French is a gender-neutral language used to highlight an awareness of gender and identity against the generic use of masculine in Standard

¹⁹Meaning “Giorgia Marras, illustrator and comic strip author, was born in Genoa, Italy, in 1988”

²⁰<https://w.wiki/7k3d>

²¹According to <https://eninclusif.fr/>. The corpus of Grouin (2022) does not contain such examples.

```

>>> from inclure.x import exclure
>>> import spacy
>>> model = spacy.load("fr_dep_news_trf")
# exclure yields aligned sentences for each sentence in the input text
>>> list(exclure(model("Bonjour à toutes et tous")))
[('Bonjour à toutes et tous', 'Bonjour à tous')]

```

Listing 1: Generating parallel sentences using the INCLURE toolkit python interface

```

>>> from transformers import pipeline, AutoModelForSeq2SeqLM
>>> inclure = pipeline("text2text-generation", model="PaulLerner/fabien.ne_barthez")
# high-level pipeline to get the output directly
>>> inclure("Bonjour à tous")
[{'generated_text': 'Bonjour à toutes et à tous'}]
# or load model for complete control
>>> model = AutoModelForSeq2SeqLM.from_pretrained("PaulLerner/fabien.ne_barthez")

```

Listing 2: Translating from Standard to Inclusive French using Fabien.ne BARThez via the Transformers library

French. Inclusive French was shown to provide fairer representations to the speakers but is also criticized for being difficult to read. With INCLURE, we sought to provide a “bilingual” access to Standard and Inclusive French.

Despite being widely used and challenging to NLP tools, Inclusive French has been very little studied in NLP. We present the second study and the first for Inclusive French translation. We provide INCLURE, a dataset of 69K aligned sentences (bitext) as well as Fabien.ne BARThez, a model able to translate from Standard to Inclusive French, and vice-versa. This model generalizes very well to out-of-domain data, through experiments on the Inclusive French Corpus (IFC) of Grouin (2022).

INCLURE comes with a toolkit for automatic annotation, which can readily be applied to larger corpora and may be extended to languages other than French, as discussed in the next section. INCLURE comes with a CLI, which can generate new training data as `python -m inclure.x <input> <output>`, where `<input>` should contain JSONL files formatted as OSCAR. Listing 1 shows how to use the Python interface. The Fabien.ne BARThez translation models can be accessed directly through the Hugging Face prediction GUI²² or via the Transformers library, see Listing 2.

We discuss our perspectives for future work in the next section.

²²Upon acceptance of the paper, similarly to <https://hf.co/moussaKam/barthez>.

8. Future Work

8.1. Vocabulary and Tokenization

We adopted BARThez as the foundation model in this work and kept its SentencePiece tokenizer. This is, however, likely suboptimal because inclusive words (e.g., député.e.s) are over-tokenized (e.g. _député . e . s). We assume that morphological tokenization (e.g., _député + <inclusive plural>) would be beneficial. A first step would be training the SentencePiece tokenizer on an Inclusive French corpus such as INCLURE. Remember that the BARThez tokenizer does not contain the interpunct, which hints at how little Inclusive French it was trained on (e.g., député.e.s is tokenized into _député <unk> e <unk> s).

However, switching tokenizers would imply re-training the model from scratch, which would allow studying two additional factors:

- the model size: do we need 139M parameters?
- its pre-training: is BARThez’ pretraining (corrupted input reconstruction) beneficial to Inclusive French Translation?

8.2. More Processes for Inclusive French

In this work, we focused on two main processes used in Inclusive French, the coordination of feminine and masculine forms, and the combination of feminine and masculine flecional endings. We plan to add other existing processes to produce Inclusive French, such as feminization of job titles and neutralization of gendered forms in producing

new morphological forms (such as the controversial *iel* personal pronoun including both masculine *il* and feminine *elle* pronouns). Another emerging process is proximity agreement, where the adjective agrees with the closest noun instead of keeping the generic masculine (e.g., *les garçons et les filles sont belles* instead of *beaux*²³; Riban and Gerin, 2017). Such syntactic rules could be detected using a dependency parser, similarly to what is described in Section 3.1.

8.3. Beyond French

French is far from the only language with inclusive varieties (Sczesny et al., 2016). Spanish, for example, uses similar processes, e.g., using @ or x to mark neutral gender instead of o (masculine) and a (feminine), for example *latinx* (Lomotey, 2015). Our work could be easily extended to other inclusive languages, such as Inclusive Spanish.

8.4. Beyond BLEU

We found in Section 5.2 that BLEU was not always suited to evaluate Inclusive French generation, due to the irregularities of Inclusive French, and the semantic equivalence between its two main processes (coordination and morphological combination). The machine translation community is gradually moving away from surface metrics like BLEU in favor of neural metrics (Nakhlé, 2023), such as COMET (Rei et al., 2020) or BLEURT (Sellam et al., 2020). We should, however, be careful before using these metrics on Inclusive French, which may be out-of-domain of the underlying language model. We should first assess the correlation between these metrics and human judgments, which would need to be collected, e.g., for the corpus of Grouin (2022).

9. Acknowledgements

We thank the reviewers for their helpful feedbacks.

This work was partly funded by the French Agence Nationale de la Recherche (ANR) under grant ANR-22-CE23-0033 / MaTOS.

10. Bibliographical References

- My Alpheratz. 2018. [Français inclusif : conceptualisation et analyse linguistique](#). *SHS Web Conf.*, 46:13003.
- ²³Meaning “the boys and girls are pretty”. *filles* is the closest noun to the adjective *belles*, which therefore agrees with the feminine.
- My Alpheratz. 2019. [Français inclusif : du discours à la langue ?](#) *Le Discours et la Langue Revue de linguistique française et d'analyse du discours*, (111):53–74.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. [Neural Machine Translation by Jointly Learning to Align and Translate](#). In *ICLR 2015*.
- Marion Baranes and Benoît Sagot. 2014. Normalisation de textes par analogie: le cas des mots inconnus. In *TALN-Traitement Automatique du Langage Naturel*, pages 137–148.
- Farah Benamara, Diana Inkpen, and Maite Taboada. 2018. Introduction to the special issue on language in social media: exploiting discourse and other contextual information. *Computational Linguistics*, 44(4):663–681.
- Franck Burlot and François Yvon. 2018. [Using monolingual data in neural machine translation: a systematic study](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.
- Bruno Cartoni. 2008. De l’incomplétude lexicale en traduction automatique: vers une approche morphosémantique multilingue (université de Genève).
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Marine Delaborde, Auphémie Ferreira, Loïc Grobol, Gabrielle Le Tallec, Benjamin Fagard, and Olga Seminck. 2021. [Usages et perception du langage inclusif : des pratiques langagières clivantes ?](#) Colloque Entre féminin et masculin – langue(s) et société, Lisbonne, Portugal.
- Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. Barthez: a skilled pretrained french sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390.
- Atefeh Farzindar and Mathieu Roche. 2013. Les défis du traitement automatique du langage pour l’analyse des réseaux sociaux. *Revue TAL–Traitement Automatique des langues*, 54(3):7–16.

- Marlis Hellinger and Hadumod Bußmann. 2015. Gender across languages: The linguistic representation of women and men. *Gender across languages*, pages 1–26.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *ICLR (Poster)*.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. [Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Benedicta Adokarley Lomotey. 2015. On sexism in language and language change—the case of peninsular spanish. *Linguistik online*, 70(1).
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamel Seddah, and Benoît Sagot. 2020. [CamemBERT: a Tasty French Language Model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Denis Maurel. 2004. Les mots inconnus sont-ils des noms propres. *Actes des JADT*.
- Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, Henning Peters, Paul O’Leary McCann, jim geovedi, Jim O’Regan, Maxim Samsonov, Daniël de Kok, György Orosz, Marcus Blättermann, Madeesh Kannan, Duygu Altinok, Raphael Mitsch, Søren Lind Kristiansen, Edward, Lj Miranda, Peter Baumgartner, Raphaël Bournhonesque, Richard Hudson, Explosion Bot, Roman, Leander Fiedler, Ryn Daniels, kadarakos, Wannaphong Phatthiyaphaibun, and Schero1994. 2023. [explosion/spaCy: v3.7.1: Bug fix for ‘spacy.cli’ module loading](#).
- Mariam Nakhlé. 2023. L’évaluation de la traduction automatique du caractère au document: un état de l’art. *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, page 143.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [PyTorch: An Imperative Style, High-Performance Deep Learning Library](#). *Advances in Neural Information Processing Systems*, 32.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Christelle Rabary, Thomas Lavergne, and Aurélie Névéol. 2015. Etiquetage morpho-syntaxique en domaine de spécialité: le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 192–198.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Chloé Riban and Murielle Gerin. 2017. [Les garçons et les filles sont belles](#).
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, 1994*.
- Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, 7:25.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Thierry Spriet, Frédéric Béchet, Marc El-Bèze, Claude de Loupy, and Liliane Khouri. 1996. Traitement automatique des mots inconnus. In *Proceedings of TALN*, volume 96, pages 170–179.

Frederik Stouten, Irina Illina, and Dominique Fohr. 2010. Regroupement des occurrences des mots hors-vocabulaire répétés en vue de leur modélisation pour la transcription d’émissions radio. *Mons, Belgique*, page 173.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. [Sequence to Sequence Learning with Neural Networks](#). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Priscille Touraille and Marc Allassonnière-Tang. 2023. Idéer une catégorie épïcène et la matérialiser cohéremment dans la langue. Une nécessité épistémologique autant que politique. In Patricia Lemarchand, editor, *Qu’est-ce qu’une femme ? Catégories homme/femme : débats contemporains*, Essais, chapter 8, pages 167–233. Editions Matériologiques, Paris.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [HuggingFace’s Transformers: State-of-the-art Natural Language Processing](#). *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

11. Language Resource References

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric Villemonte de La Clergerie. 2014. Deep syntax annotation of the sequoia french treebank. In *International Conference on Language Resources and Evaluation (LREC)*.

Cyril Grouin. 2022. [Impact du français inclusif sur les outils du TAL \(Impact of French Inclusive Language on NLP Tools\)](#). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 126–135, Avignon, France. ATALA.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

BnPC: A Gold Standard Corpus for Paraphrase Detection in Bangla, and its Evaluation

Sourav Saha^{1*}, Zeshan Ahmed Nobin^{1*}, Mufassir Ahmad Chowdhury^{1*},
Md. Shakirul Hasan Khan Mobin^{1*}, Mohammad Ruhul Amin², Sudipta Kar³

Shahjalal University of Science and Technology, Bangladesh,¹
{sourav95, zeshan07, mufassir73, shakirul34}@student.sust.edu,¹
Fordham University, USA,² Amazon, USA³
mamin17@fordham.edu,² sudipkar@amazon.com³

Abstract

In this paper, we present a benchmark dataset for paraphrase detection in Bangla. Despite being the sixth most spoken language¹ in the world, paraphrase identification in Bangla is barely explored. Our dataset contains 8,787 **human-annotated** sentence pairs collected from 23 newspaper outlets' headlines in four categories. We explored several supervised modeling approaches to benchmark the dataset, including similarity metrics, linguistic features, and fine-tuned BERT models. We also conducted a zero-shot analysis to assess the performance of pre-trained BERT models, and we carried out both zero-shot and few-shot evaluations of the publicly accessible generative language model GPT 3.5 turbo. In the benchmark evaluations, when examining GPT-3.5 using a few-shot modeling approach, it becomes evident that the model can grasp paraphrases in a manner akin to fine-tuned mBERT language models with just a handful of example data points. Within the set of benchmarking trials, the fine-tuned BanglaBERT delivered the most remarkable performance, achieving a weighted-F1 score of 87.91. Noteworthy is that GPT-3.5 excelled in both zero-shot and few-shot experiments, attaining weighted-F1 scores of 51.51 and 80.53, in that order. We also performed a cross-dataset analysis and the outcomes suggest that the model trained in our dataset resembles both diversity and generalization when tested on the other dataset. Finally, we report a human evaluation experiment to obtain a better understanding of the paraphrasing task's limitations. We make our dataset and code publicly available.²

Keywords: Paraphrase Identification, Semantic Similarity, Benchmarking Dataset, Cross Dataset Analysis

1. Introduction

Paraphrase identification is considered to be one of the pivotal and fundamental tasks of Natural Language Processing (NLP). When two different sentences express the same meaning, they are called paraphrases. Paraphrase identification has many implications on tasks like question answering (Fader et al., 2013a), text summarization (Barzilay et al., 1999), plagiarism detection (Barrón-Cedeño et al., 2013), information retrieval (Wallis, 1993), first story detection (Petrović et al., 2012), and value alignment, etc. As a result, extensive research has been conducted on paraphrase identification, and numerous paraphrase corpora have been developed in various languages like English (Dolan and Brockett, 2005; Xu et al., 2015a; Lan et al., 2017; He et al., 2020a), Turkish (Demir et al., 2012), Russian (Prinoza et al., 2016), Arabic (Menai, 2019), Portuguese (Fonseca et al., 2016), Chinese (Zhang et al., 2019), among others.

A descendent of Sanskrit, Bangla is currently

*Authors have equal contribution

¹[w.wiki/Pss](https://en.wikipedia.org/wiki/List_of_languages_by_number_of_speakers)

²<https://github.com/Mufassir-Chowdhury/BnPC>

Paraphrases with slight lexical differences <ul style="list-style-type: none">• কাল মিয়ানমারে জাতীয় নির্বাচন, রোহিঙ্গারা বঞ্চিত <i>National elections in Myanmar tomorrow, Rohingyas deprived</i>• মিয়ানমারে কাল নির্বাচন : ভোট নেই রোহিঙ্গাদের <i>Tomorrow's election in Myanmar: Rohingyas do not have votes</i>
Paraphrases with significant lexical differences <ul style="list-style-type: none">• বিজিবি এখন জলে, স্থলে ও আকাশপথে বিচরণ করবে <i>The BGB will now operate on water, land and air</i>• বিজিবির এয়ার উইংয়ের যাত্রা শুরু, ত্রিমাত্রিক বাহিনী ঘোষণা <i>The BGB air wing begins its journey, announcing three-dimensional forces</i>
Non-paraphrases with significant lexical similarity <ul style="list-style-type: none">• পদ্মা সেতুর ৩২তম স্প্যান বসতে পারে আজ <i>The 32nd span of the Padma Bridge can sit today</i>• পদ্মা সেতুর ৩২তম স্প্যান বসতে পারে কাল <i>The 32nd span of the Padma Bridge may sit tomorrow</i>
Non-paraphrases with slight lexical similarity <ul style="list-style-type: none">• ফিটনেস টেস্টে সাকিবের বাজিমাত <i>Shakib's shines in fitness test</i>• এক বছরেও 'ফিট' হতে পারেননি নাসির <i>Nasir could not be 'fit' in a year</i>

Table 1: Examples of paraphrase and non-paraphrase pairs with different amount of lexical overlap.

spoken by over 260 million people in the world and is set to become the third most spoken language by 2050.³ Bangla is the language of the

³[washingtonpost.com/news/worldviews/wp/2015/09/24/the-future-of-language](https://www.washingtonpost.com/news/worldviews/wp/2015/09/24/the-future-of-language)

people of the Bengal region, now divided between Bangladesh and the Indian state of West Bengal, which are considered to be the region of fastest growing economies.⁴ Because of the technological advancements in Bangla speaking communities, the demand and usage of the Bangla language in the digital world continue to grow exponentially.

Despite such a growing demand and need for digital Bangla resources, the task of Bangla paraphrase identification has received limited attention. Akil et al. (2022) generated a synthetic Bangla paraphrase dataset consisting of 603,672 sentence pairs. Kumar et al. (2022) also experimented with six different NLG tasks across eleven Indic languages including the task of Bangla paraphrase generation. Meanwhile, Scherrer (2020) curated sentential paraphrases on 73 languages including Bangla, for which they considered only 1,440 Bangla sentences.

To address the scarcity of paraphrase detection dataset in Bangla language, we propose BnPC, a gold-standard Bangla paraphrase corpus. We outline the contributions of this study below:

- We propose BnPC, the largest gold standard paraphrase corpus in Bangla, consisting of 8,787 **human-annotated** pairs collected from 23 different newspaper outlets in Bangladesh. We present a few examples in Table 1.
- We report a benchmark evaluation on BnPC by exploiting several supervised learning approaches, such as the similarity metrics (BLEU, METEOR), bag-of-words approach (Word and Character n-grams), and fine-tuned language models.
- We carried out both the zero-shot and few-shot experiments over the publicly accessible GPT-3.5 turbo model using BnPC and present shortcomings we observed from GPT-3.5 responses.
- We performed a cross-dataset analysis by fine-tuning a monolingual and a multilingual BERT on BnPC and testing it on several other datasets. We show that models trained on BnPC resembles the capacity to provide better performance on diverse datasets.
- We also conducted a human evaluation experiment to get insights into the paraphrasing task's limitations.

2. Related Work

Over the recent years, a great deal of work has been accomplished in paraphrase detection. We

discuss some of the notable works in this section.

Datasets for Paraphrase Identification: MSRP (Dolan and Brockett, 2005; Dolan et al., 2004) is the pioneering hand-labeled dataset extracted using heuristic techniques instead of the traditional machine translation method. Their approach obtained high lexical divergent paraphrase pairs, opening up new dimensions in the paraphrase identification field. Twitter Paraphrase Corpus (PIT-2015) (Xu et al., 2015a) is a realistic and balanced dataset collected from trending topics on Twitter containing a high degree of variation due to the use of informal language as well as more naturally occurring non-paraphrases. Twitter URL Corpus (TUC) (Lan et al., 2017) is a shared URL based growing paraphrase corpus with both formal and informal texts, where the authors mitigate the complications of extracting highly variant natural paraphrase sentence pairs on a large scale. Quora Question Pair (Chen et al., 2017) is a dataset containing interrogative sentence pairs that benefit the Q&A community by assisting in the detection of duplicate questions. PARADE (He et al., 2020b) is a domain-specific dataset where authors formed clusters of definitions focusing same aspect indicated by overlapping term and matched every two definitions from the same cluster together.

Approaches used in Paraphrase Detection: The noteworthy approaches for the task of paraphrase identification are MT metric based classifiers (Eyecioglu and Keller, 2015) combining lexical and compositional features. The modeling approaches include referential and machine translations (Finch et al., 2005; Biçici and Way, 2014), feature based approaches (Zarrella et al., 2015), supervised learning (Vo et al., 2015; Karan et al., 2015) using SVM and logistic regression (Satya-panich et al., 2015; Madnani et al., 2012a; van der Goot and van Noord, 2015), deep learning and BERT based approaches (Zhao and Lan, 2015; Bertero and Fung, 2015; Chandra and Stefanus, 2020).

Bangla Paraphrase Detection: TaPaCo (Scherrer, 2020) is a paraphrase corpus generated by populating a graph from the Tatoeba database and finding equivalent links between the sentence pairs with everyday sentences. They used a crowd-sourced method of paraphrase generation without assessing the capability of the translators. BanglaParaphrase (Akil et al., 2022) curated sentences from a Bangla blogging website using a machine translation (back-translation) and a novel filtering process based on PINC score (Chen and Dolan, 2011) (a metric based on lexical dissimilarity). IndicNLG used pivoting approach (Kumar et al., 2022) to extract paraphrases from a parallel corpus using English as the pivot.

⁴[britannica.com/place/Bengal-region-Asia](https://www.britannica.com/place/Bengal-region-Asia)

In contrast, the BnPC dataset was created from human-generated text from newspaper headlines and labeled by three expert annotators validating all paraphrase pairs using a rigorous process to ensure the quality of the data.

3. Overview of BnPC Dataset

Data Collection: We constructed the BnPC corpus by gathering news headlines from 23 of the most popular⁵ Bangla news portals. This is because headlines for similar news tend to be paraphrases. Thus we gathered news on four broad categories: *national*, *international*, *sports*, and *entertainment* over the four months starting from September to December of 2020. Alongside visiting individual news websites, we also utilized Google News⁶ service to retrieve cluster of similar news, and a similar service from the Pipilika News⁷.

Through manual inspection, we formed a total of 145 national, 158 international, 139 sports, and 175 entertainment related news clusters by selecting similar news of identical events. Each cluster contained different headlines focusing on different aspects of the same event reported by various news agencies. We followed different methods of paraphrasing to select paraphrasing pairs. These methods are presented in Table 2.

Annotation: Three of the native Bangla-speaking authors annotated the pairs. Each annotator was trained on different methods of paraphrasing according to Table 2. We decided to use five different paraphrase scores on a scale from 0 to 1 to reach a better labeling consensus among the annotators at the end of the process.

We discuss our score assignment for each of the 5 different paraphrasing decision: (1) “Not Paraphrase”: Score 0; (2) “Not-Paraphrase with Slight Similarity”: Score 0.25; (3) “Undecided”: Score 0.5; (4) “Paraphrase with Lexical Differences”: Score 0.75; and (5) “Paraphrase”: Score 1.0.

During the annotation, we followed the guidelines described in Bhagat and Hovy (2013). We averaged the scores of three annotators. Sample above the threshold score (0.5) were considered as paraphrase and below it as non-paraphrase in the final dataset. We discarded the ones with an average score of 0.5 as the annotators could not agree on whether the pairs were paraphrase or not. These samples were mostly partial paraphrases or had ambiguous meanings. A Fleiss’ Kappa score

(Fleiss, 1971) of 0.61 indicates substantial inter-annotator agreement. We present some sample sentence pairs in Table 1.

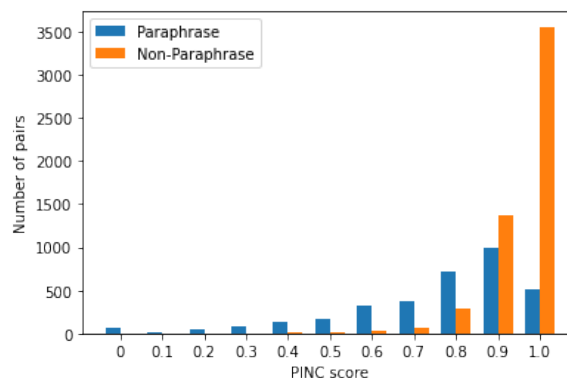


Figure 1: PINC score of paraphrase and non-paraphrase pairs of BnPC. PINC score denotes n-gram dissimilarity between two sentences. High PINC score denotes low lexical overlaps.

Statistics: As per Table 3, the class distribution of the dataset is slightly skewed towards the non-paraphrases, and non-paraphrase sentences tend to be a little longer than the paraphrase ones. There are 8,541 unique Bangla words (23.8%) in the dataset. We observe lexical diversity in the dataset as 35.19% sentence pairs have zero and 28.94% pairs have only one word in common. The high PINC score (Chen and Dolan, 2011) in Figure 1 for both paraphrase and non-paraphrase pairs indicates that the dataset contains more lexically diverse sentences. The diversity among the non-paraphrase pairs is more abundant.

Analysis: Paraphrase identification from real-world data is noisy and follows a wide range of methods compared to synthetically generated pairs. In our BnPC dataset, we analyzed various methods of paraphrases (Table 2). Often times, more than these methods are observed in paraphrase making in Bangla. This makes paraphrase detection in Bangla significantly more challenging for rule-based approaches.

4. Methodology

To develop a paraphrase classifier, we explore the metrics for machine translation evaluation, bag-of-words, zero-shot approaches and fine-tuning pre-trained language models.

4.1. Evaluation Metric Based Approach

Following Madnani et al. (2012b) and Kravchenko (2017), we investigate paraphrase classifiers using machine translation (MT) evaluation metrics

⁵[alexa.com/topsites/countries/BD](https://www.alexa.com/topsites/countries/BD)

⁶news.google.com/?hl=bn

⁷news.pipilika.com

Methods of Paraphrase	Explanation	Sentence1	Sentence2
Change Of Order	Change of order involves changing the order of a word or phrase in a sentence	নতুন আর্টার্নি জেনারেল এ এম আমিন উদ্দিন (Newly appointed attorney general is A M Amin Uddin)	আমিন উদ্দিন নতুন আর্টার্নি জেনারেল (Amin Uddin is newly appointed attorney general)
Synonym Substitutions	It involves the replacement of a word or phrase of the sentence with one of its synonyms	পিস্তল কিনে ফেসবুকে ছবি দিলেন এমপি! (M.P. posted a photo on facebook after purchasing a pistol)	পিস্তল কিনে ফেসবুকে ছবি দিলেন সংসদ সদস্য (The Member of Parliament posted a picture on Facebook after buying a pistol)
Verbatim	It is a type of plagiarism where a sentence is copied without changing any aspect of the sentence	লাইফ সাপোর্টে ব্যারিস্টার রাফিক-উল হক (Barrister Rafiq-ul-Haque on life support)	লাইফ সাপোর্টে ব্যারিস্টার রাফিক-উল হক (Barrister Rafiq-ul-Haque on life support)
Ellipsis	Ellipsis involves the omission of clauses that are understood from the context of the remaining sentence	করোনায় আরো ১৮ জনের মৃত্যু (18 more people die from Corona)	করোনায় আরো ১৮ মৃত্যু (18 more die from Corona)
Punctuation Changes	Punctuation changes involve the change of punctuation used in the sentence	গুগল, ফেসবুক ও ইউটিউব থেকে রাজস্ব আদায়ের নির্দেশ	গুগল-ফেসবুক-ইউটিউব থেকে রাজস্ব আদায়ের নির্দেশ
Emphasization	Emphasization is a type of paraphrase where the exact same	চলতি মাসে দেশে আঘাত হানতে পারে ঘূর্ণিঝড় (A cyclone is expected to strike our country later this month)	চলতি মাসেই আঘাত হানতে পারে ঘূর্ণিঝড় (A cyclone is set to hit the country this very month)
Abbreviation	It involves shortened form of a word or phrase in one of the pairs	ইউরোপীয় ইউনিয়নের সঙ্গে সম্পর্কচ্ছেদের হুমিয়ারি রাশিয়ার (Russia issues a threat to sever ties with the European Union)	ইইউ ছাড়ার হুমকি দিল রাশিয়া (Russia threatens to leave EU)

Table 2: This table presents different methods of paraphrasing in our BnPC dataset. Most of the definitions are picked from Zhou et al. (2022).[‡]

	T	P	W/S	C/S
Paraphrase	3,426	38.99%	6.97	46.95
Non-Paraphrase	5,361	61.01%	7.32	48.86
Total	8,787	100.00%	7.18	48.11

Table 3: Distribution of T (total number), P (percentage), W/S (word per sentence), and C/S (character per sentence) between paraphrase and non-paraphrase sentence pairs in the dataset.

Root Word	
Type	Example
Root	৭৮ দিনে করোনা শনাক্তের হার সর্বোচ্চ (Corona detection rate is highest in 7 days)
Prefix	প্রথমবার ওয়েব সিরিজে জুটি বাঁধছেন সোহম-শ্রাবন্তী দর্শকদের উপহার (উপ + হার) দেবেন থ্রিলার লাভস্টোরি (Soham-Sravanti to tie the knot for the first time in web series, thriller Love Story to present to viewers)
Suffix	তছেন সু চি, আবার হারছে (হার + ছে) রোহিঙ্গরা? (Suu Kyi is winning, Rohingyas are losing again?)
Concatination	২৪ ঘণ্টায় করোনায় মৃত্যুহার (মৃত্যু + হার) কমেছে (Mortality rate in Corona has decreased in 24 hours)

Table 4: Examples of prefix, suffix, and concatenation usage in Bangla from our dataset.[‡]

like BLEU (Papineni et al., 2002a) and METEOR (Lavie and Denkowski, 2009) as these metrics provide a notion of lexical similarity between a reference and a generated text. Given a candidate pair $X = (x_1, x_2)$ and a metric (e.g., BLEU), we classify the pair as a paraphrase or not paraphrase by the following equations:

$$f_{BLEU}(X) = \frac{BLEU(x_1, x_2) + BLEU(x_2, x_1)}{2}$$

$$\hat{y} = \begin{cases} \text{PARAPHRASE, IF } f_{BLEU}(X) \geq \alpha \\ \text{NOT PARAPHRASE, IF } f_{BLEU}(X) < \alpha \end{cases}$$

Here, α is a threshold, whose value was set by maximizing the performance on the training set ($\alpha=0.115$ for BLEU and $\alpha=0.136$ for METEOR).

4.2. Bag of Words (BOW)

For each text in a candidate pair, we extract word n-grams ($n=1, 2, 3$) and character n-grams ($n=2, 3, 4, 5$) and use the cosine similarity scores for each n-gram set as features to train a Support Vector Machine (SVM) classifier. Additionally, we investigate training the model by dividing the mean word embedding vectors of the pair, by its norm and taking the quotient as input feature. We use the pre-trained FastText (Bojanowski et al., 2016) Bangla embedding (coverage 91.77%) for this purpose.

4.3. Language Models

Pre-trained language models, particularly variants of BERT, have shown superior performance in a variety of natural language tasks. On the other hand, recent LLMs have shown superior quality in performing different NLP domain tasks. We use the Multilingual BERT (mBERT) (Devlin et al., 2018), RoBERTa (Liu et al., 2019c), XLM-RoBERTa (Conneau et al., 2019), and three different monolingual BERT models pre-trained on Bangla (Sarker, 2020; Bhattacharjee et al., 2021; Diskin et al., 2021)^{8,9,10} from HuggingFace transformers (Wolf et al., 2020) and fine tune the

⁸huggingface.co/csebuetnlp/banglabert

⁹huggingface.co/sagorsarker/bangla-bert-base

¹⁰huggingface.co/neuropark/sahajBERT

binary prediction layer. We reported the zero-shot performance of mBERT, XLM-RoBERTa, BanglaBERT. Additionally, we perform zero-shot and few-shot approaches on publically available GPT 3.5 turbo. BanglaBERT (Bhattacharjee et al., 2021) was trained on 27.5 GB data crawled from 110 Bangla websites, whereas bangla-bert-base (Sarker, 2020) was trained on wikidump and 11 GB web crawled data from OSCAR (Ortiz Suárez et al., 2020).

5. Experiments and Results

5.1. Experimental Setup

We use 70% of the data for training, and equally divide the rest for development and testing. For the metric-based approaches, we remove the punctuations and for BOW-based methods, we preprocess the data by removing punctuation and normalizing digits as it shows better results in the development set. As a set of simple baselines, we compare our results with a majority and a random baseline. We report our results using precision, recall, and weighted F1 score. We use Scikit-learn (Buitinck et al., 2013) implementations for SVM, cosine similarity, and n-gram extraction. For the pre-trained language models, we fine-tune ($\lambda=2*10^{-5}$, batch size 32) the models for 5 epochs with early stopping. For gpt-3.5-prompting we used the ChatGPT Platform API ¹¹ with the following parameters: temperature=0 (0 for deterministic output), max_tokens=256, top_p=1, frequency_penalty=0, presence_penalty=0.

5.2. Results & Analysis

Table 5 presents the precision, recall, and weighted F1 scores of different models on the test set. The MT metric-based approaches (BLEU, METEOR) perform relatively well compared to the baselines, with METEOR getting up to 77.08 F1 score. METEOR considers both unigram precision and recall, whereas BLEU solely measures precision when matching the sentence pairs. As a consequence, METEOR exhibits better performance for the task.

Unigram performs the best among the word n-grams with an F1 score of 74.93 and we notice a decline in F1 for the longer word n-grams. This pattern is consistent with the character n-grams as well. Character bigrams achieve a 77.97 F1 score and longer ngrams' F1 score decreases gradually. However, character n-grams show better performance than the word n-grams in general. Usage of prefixes, suffixes, and word concatenation is heavy in Bangla, which we believe is the reason for the

Model	P	R	F1
Baseline (Random)	50.56	50.67	49.62
Baseline (Majority)	34.86	59.04	43.83
BLEU	76.95	76.76	76.10
METEOR	77.28	77.40	77.08
Unigram (U)	76.67	75.97	74.93
Bigram (B)	74.59	73.67	72.21
Trigram (T)	73.88	66.36	59.46
U+B	76.30	75.82	74.90
U+B+T	76.42	75.90	74.95
Char-2-gram (C2)	79.07	78.62	77.97
Char-3-gram (C3)	78.61	78.41	77.87
Char-4-gram (C4)	78.06	77.76	77.12
Char-5-gram (C5)	77.52	76.97	76.12
C2+C3	78.72	78.41	77.80
C2+C3+C4	78.19	77.98	77.40
C2+C3+C4+C5	78.39	78.12	77.52
U+C2	79.22	78.77	78.11
U+C2+C3	78.73	78.34	77.68
U+C2+C3+C4	78.47	78.05	77.36
All n-grams	78.26	77.76	77.01
Word Embedding (E)	77.53	77.04	76.24
U+C2+E	78.83	78.19	77.41
bangla-bert-base (Zero-Shot)	51.54	58.68	45.02
mBERT (Zero-Shot)	26.39	48.87	23.82
XLM-RoBERTa (Zero-Shot)	34.86	59.04	43.83
sahajBERT (Zero-Shot)	55.29	48.85	46.85
BanglaBERT (Zero-Shot)	59.67	51.92	48.79
gpt-3.5-turbo (Zero-Shot)	71.69	62.27	51.51
gpt-3.5-turbo (Few-Shot)	80.53	80.63	80.53
bangla-bert-base (Sarker, 2020)	75.85	76.04	75.75
mBERT (Devlin et al., 2018)	82.54	82.42	82.47
XLM-RoBERTa (Conneau et al., 2019)	86.11	86.08	85.96
sahajBERT (Diskin et al., 2021)	86.55	86.37	86.19
BanglaBERT (Bhattacharjee et al., 2021)	87.92	87.95	87.91

Table 5: Results from different experiments of baseline, MT metrics, linguistic features, and pre-trained LMs are reported in Precision (P), Recall (R) and weighted-F1 score.

strength of character n-grams (Table 4). The combination of unigram and character bigram yields the highest F1 score of 78.11 among all the lexical feature combinations. We observe no improvement in this by integrating the embedding features.

Zero-shot performance of the models is significantly low (even compared to feature-based approaches). Among the zero-shot performance of the models, the GPT 3.5 turbo achieves the best results with an F1 score of 51.51. Interestingly, the GPT 3.5 turbo few-shot exhibits a significant performance boost. The few-shot (4-shot, two paraphrases, and two non-paraphrases) achieves an F1 score of 80.53 closer to the finetuning result of some LMs and surpassing all feature-based approaches indicating the paraphrase detection capabilities of large language models. We provide some interesting examples of LLM's failure in Table 7.

On our dataset, the best-performing model is BanglaBERT (Bhattacharjee et al., 2021), outperforming XLM-RoBERTa by a close margin. BanglaBERT is pre-trained on the highest volume of Bangla data (27.5 GB) to date. The competitive performance of XLM-RoBERTa results from its effective cross-lingual transfer learning.

To provide a performance comparison of the best-performing multilingual model with other

¹¹platform.openai.com/

Sentence 1	Sentence 2	Label	*Subject	**Model
প্রধানমন্ত্রীর সংবাদ সম্মেলন শনিবার (The Prime Minister's press conference is on Saturday)	প্রধানমন্ত্রীর সংবাদ সম্মেলন আজ (The Prime Minister's press conference is today)	0	0	1
জাপানে শক্তিশালী ভূমিকম্পে আহত শতাধিক (Hundreds injured in strong earthquake in Japan)	জাপানের উপকূলে ৭ দশমিক ৩ মাত্রার ভূমিকম্প (7.3 magnitude earthquake off the coast of Japan)	0	1	0
করোনায় মৃত্যু প্রায় ২৪ লাখ (About 24 lakh died in Corona)	মৃত্যু ২৩ লাখ ৬৭ হাজার, আক্রান্ত ১০ কোটি সাড়ে ৭৭ লাখের বেশি (23 lakh 67 thousand deaths, more than 10 crore 77.5 lakh affected)	1	1	0
জাপানের উত্তরাঞ্চলে ৭.৩ মাত্রার ভূমিকম্প (7.3 magnitude earthquake shakes northern Japan)	জাপানে ৭.১ মাত্রার ভূমিকম্প (7.1 magnitude earthquake shakes Japan)	1	0	1
আমেরিকার এই কুখ্যাত জেল বন্ধ করতে পারেন বাইডেন (Biden might close this infamous prison in America)	গুয়ানতানামো বে কারাগার বন্ধ করতে চান বাইডেন (Biden wants to close Guantanamo Bay prison)	1	0	0

Table 6: Disagreement among subject, model, and actual label. Here 1 represents paraphrase and 0 represents non-paraphrase sentence pairs. *Subject's prediction is taken using majority voting. **Prediction on BanglaBERT. ‡

Sentence 1	Sentence 2	Reason
খুলনায় ২৪ ঘণ্টা বন্ধ থাকবে পরিবহন (Transportation will be closed in Khulna for 24 hours)	খুলনায় পরিবহন চলাচল বন্ধ ঘোষণা (Transport closure announced in Khulna)	Unless it's direct syntactic similarity the LLM model fails in case of bangla. The broader context is easier for humans to comprehend.
চাঁদপুরে আগুনে পুড়ে স্কুল শিক্ষিকার রহস্যজনক মৃত্যু (Mysterious death of school teacher in fire in Chandpur)	আগুনে অঙ্গার শিক্ষিকা (Teacher turned into cinder in a fire)	LLMs struggle with idiomatic expressions, often misinterpreting them.
ব্রিটেনে আর ফিরতে পারবেন না শামীমা (Shamima will not be able to return to Britain)	শামীমার যুক্তরাজ্যে ফেরার আবেদন নাকচ করলেন আদালত (The court rejected Shamima's request to return to the UK)	LLMs may not detect paraphrases when two sentences convey the same news but use different subjects.
ফের নানা হলেন ডিপজল (Deepzal became grandfather again)	মা হলেন ডিপজল কন্যা ওলিজা (Deepzal daughter Oliza became a mother)	LLMs may struggle to follow logical syllogisms accurately.
১০০১ দিন পর জেল থেকে মুক্তি পেলেন সৌদি অধিকারকর্মী (Saudi rights activist released from jail after 1001 days)	৩ বছর পর সৌদির নারী অধিকার কর্মী লুজাইনের মুক্তি (Saudi women's rights activist Luzain released from jail after 3 years)	LLMs can be confused by changes in units when interpreting or processing information.

Table 7: Examples of sentence pairs where LLMs fail to classify using few-shot approach. ‡

datasets, we fine-tune XLM-RoBERTa on other substantial English datasets with the identical experimental setup. The F1 scores are 90.78 on MSRP (Dolan and Brockett, 2005), 75.01 on PARADE (He et al., 2020), and 88.31 on PIT (Xu et al., 2015a). 85.96 F1 on BnPC falls in between these scores and provides a competitive benchmark result.

5.3. Comparison of Datasets

Cross Dataset Generation: As the other datasets don't have any non-paraphrase pairs, we added the non-paraphrase from our dataset. To compare the quality of the contemporary datasets with the BnPC, we also maintained the paraphrase and non-paraphrase ratio of BnPC on the other datasets. For BanglaParaphrase and IndicNLG we randomly sampled the equivalent number of paraphrases as BnPC and appended all our non-paraphrase pairs to them. Since these two datasets are substantially larger than BnPC we repeated this process three times for brevity and experimented with each of these datasets and averaged the results. Since TaPaCo has a smaller size than BnPC, we appended only a random portion

of our non-paraphrase pairs to maintain the overall paraphrase and non-paraphrase ratio equivalent to BnPC. To ensure unbiased experiments, we include non-paraphrase pairs from our train, test, and validation sets into the corresponding sets of other datasets. (Fig: 2)

Results: To conduct cross-dataset testing, we implement both monolingual (sahajBERT) and multilingual (mBERT) models across various merged datasets. The models trained on BnPC consistently perform well across all datasets, achieving a minimum F1 score of 69.97 on IndicNLG. On the other hand, models trained on BanglaParaphrase excel across most datasets and face a downfall of performance on our gold standard BnPC dataset, scoring below 50%, while surpassing the 92% F1 score on other datasets. Models trained on TaPaCo demonstrate strong performance across most tests, with the notable exception of BnPC, where they yield the lowest F1 score of 44% among all the cross-dataset experiments. IndicNLG proves to be a strong performer across synthetic datasets, consistently achieving over 97%, and it delivers a respectable F1 score of 57.32 on our BnPC dataset. In summary, models trained on synthetic datasets display subpar performance when tested on our gold standard dataset.

We obtain the context of the paraphrase pairs by using BLEU (Papineni et al., 2002a) and ROUGE

‡denotes the sentences in these tables were translated using Google Translator for the clarity of the non Bangla speakers.

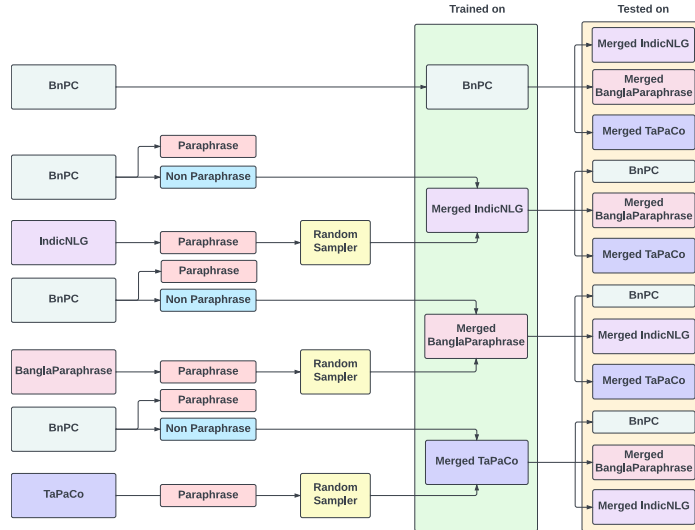


Figure 2: The workflow diagram of cross-dataset test. It shows the procedures for generating the merged datasets for cross-dataset experiments and the experimental procedures.

Model	Trained On	Tested On (BnPC)			Tested On (BanglaParaphrase)			Tested On (TaPaCo)			Tested On (IndicNLG)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
SahajBERT	BnPC	86.55	86.37	86.19	94.59	94.18	94.22	86.42	86.37	86.24	73.25	71.81	69.67
	BanglaParaphrase	73.21	61.31	49.04	99.52	99.54	99.54	98.84	98.83	98.82	93.56	92.87	92.73
	TaPaCo	75.84	59.11	44.00	84.16	78.35	76.14	100.00	100.00	100.00	90.08	88.07	87.60
	IndicNLG	71.02	64.94	57.32	97.61	97.55	97.56	98.29	98.22	98.22	97.18	97.15	97.15
mBERT	BnPC	80.61	80.63	80.62	89.85	88.61	88.70	84.88	84.54	84.62	65.14	65.71	63.75
	BanglaParaphrase	72.66	62.38	51.57	99.23	99.23	99.23	89.60	87.91	87.48	78.83	70.35	65.22
	TaPaCo	72.36	60.25	46.77	80.67	72.14	67.70	99.87	99.87	99.87	87.28	84.05	83.09
	IndicNLG	69.68	65.23	58.40	96.60	96.54	96.55	97.96	97.85	97.85	96.36	96.32	96.33

Table 8: The table shows the cross-dataset performance of monolingual (SahajBERT) and multilingual (mBERT) models. It contains precision (P), recall (R), and weighted-F1 scores of the models. The worst performances (row-wise) are shown in red and the best performances (row-wise) are shown in blue.

(Lin, 2004) metrics. We see that TaPaCo has the highest n-gram similarity since it mostly consists of simple and small sentences. IndicNLG shows the lowest n-gram similarity across all the metrics. BanglaParaphrase and BnPC have similar n-gram similarity across the metrics indicating a moderate n-gram overlap.

Analysis: From Table 8, we see that models trained on synthetic datasets show poor performance on human-generated data. On the other hand, models trained on BnPC show decent performance on synthetic datasets. Despite BnPC having moderate n-gram similarities, the failure of models trained on other datasets and tested on BnPC can originate from the wide distribution of paraphrases across the PINC Score spectrum. The BnPC paraphrases are spread across the spectrum from 0.0-1.0, which is absent in other datasets with the single highest being only 36.25% of the samples on 0.9. 82% of the data is within 0.6-1.0. and the other 18% data falls within 0.0-0.5 which is the highest among other datasets.

The monolingual Model trained on BanglaParaphrase did well except on BnPC and the Multi-

lingual model trained on BanglaParaphrase did moderate performance on BnPC and IndicNLG. This can stem from the fact that BanglaParaphrase has paraphrases (~98%) mostly spread within 0.6-0.9 PINC score with 44.48% data on 0.8. This makes it hard to perform well on a dataset with more distributed n-gram similarity and similar size. Models trained on other datasets and tested on BanglaParaphrase show a better performance except for TaPaCo which might originate from the smaller size of TaPaCo. TaPaCo has 42% smaller size than the other datasets. Models trained on other datasets show good performance on TaPaCo. This can be traced back to the smaller size of the dataset, sentences, and high n-gram similarity of TaPaCo paraphrase pairs shown in Figure 3 which is the highest among all the datasets. Making it easier to identify paraphrases in simple and small sentences. Models trained on TaPaCo show poor performances on all the datasets except on IndicNLG. IndicNLG has the lowest n-gram similarity among all the datasets. Because of this, models tested on IndicNLG show comparatively weaker performance.

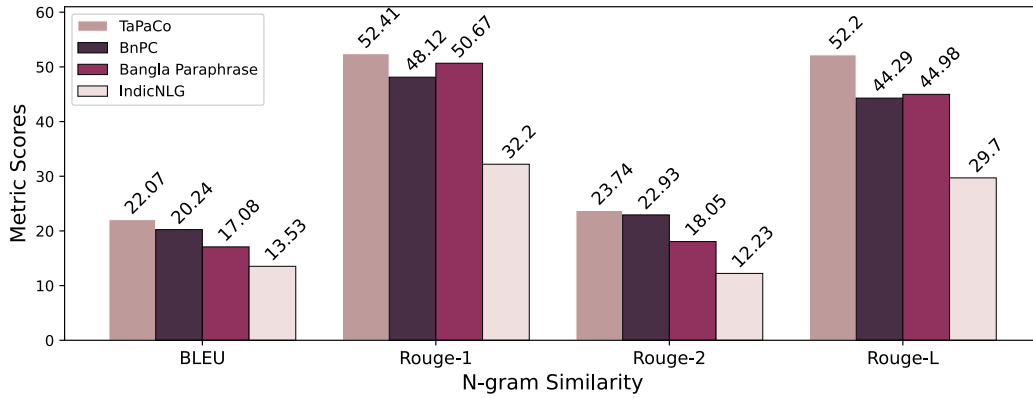


Figure 3: This Figure shows the N-gram similarity comparison of the datasets. For comparing the N-gram similarity we implement BLEU, [Rouge-N, Rouge-L](Lin, 2004) methods.

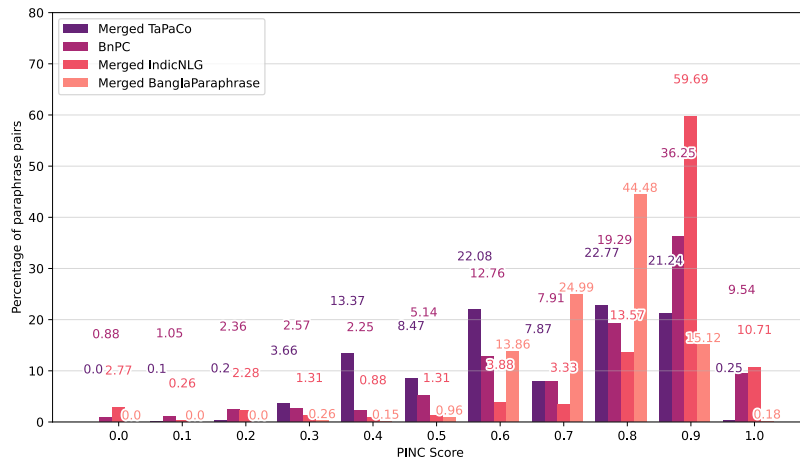


Figure 4: This Figure shows the PINC score comparison of the datasets. PINC score denotes n-gram dissimilarity between two sentences. High PINC score denotes low lexical overlaps between the sentence pairs.

Models trained on IndicNLG show a good performance except on BnPC. Figure 4 shows that almost 60% of their paraphrase pairs stem 0.9 PINC score. This is a probable reason for the IndicNLG’s poor performance on the BnPC dataset. We exhibit that models trained with lower n-gram similarity tend to do well on datasets that have higher n-gram similarity on paraphrase identification task.

5.4. Human Evaluation

We conduct a human evaluation study with 300 randomly selected examples from our test set to assess the human performance in the task. We take the help of five native Bangla-speaking undergraduate students from different majors on a voluntary basis to ensure diversity in subjects. After instructing them about the task, we asked them to classify each pair into either paraphrase

or non-paraphrase. Then we compare their assigned labels against the ground truth. The individual F1 scores of the five annotators are 69.48, 72.25, 74.37, 74.58, and 84.13, yielding an average F1 score of 74.96. Using Fleiss’ Kappa metric, we calculate the inter-annotator agreement of those pupils and get a score of 0.47. The best-performing model’s F1 score of 87.98 on this sample of data indicates that the job can be more difficult for humans to accomplish.

Analyzing the errors and interviewing the human subjects, we find that the main reasons are lack of domain knowledge, presence of numbers in the sentences, and pairs with long overlaps of spans. (Table 6).

6. Conclusion and Future Works

In this paper, we propose BnPC, the largest hand-crafted Bangla dataset for paraphrase detection. Through our investigations to develop a benchmark classifier, we find that lexical features like character n-grams show competitive performance in identifying paraphrases. Similar performance can be achieved by simply using the machine translation metric-based classifiers. From our experiments, we see that the monolingual model BanglaBERT slightly outperforms the multilingual model XLM-RoBERTa on the BnPC dataset. Also, we find the GPT-3.5 turbo performs almost as well as fine-tuned language models. Our cross-dataset analysis shows that models trained on our dataset generalize more compared to contemporary datasets and we provide some quantitative analysis differentiating the datasets. Our dataset comprises formal data from newspaper headlines. So, a good direction for future work can be extending this dataset with different domains and topics' data, for example, conversational data. We release the corpus publicly to foster further work in this area.

Limitations

The study has some potential limitations. One potential limitation is that our dataset is comprised of formal data from news headlines which is different from the noisy data on social media. Social media data generally contains misspellings, and slang words creating challenges for paraphrase detection tasks, which is absent in our dataset. Other potential sources for curating a paraphrase dataset include blogs, books, and various academic writings. Moreover, our dataset comprises roughly 9K data leaving the scope for extending the dataset in the future.

Ethical Considerations

Dataset Release: The Copyright Act, 2000¹² of People's Republic of Bangladesh allows reproduction and public release of copyright materials for non-commercial research proposals. We will release our BnPC dataset under a non-commercial license. Publicizing other supplementary materials like codes won't cause any copyright infringements.

Annotators' Compensation: All the annotators participated voluntarily in this research work.

¹²[http://copyrightoffice.portal.gov.bd/sites/default/files/files/copyrightoffice.portal.gov.bd/law/121de2e9_9bc9_4944_bfef_0a12af0864a5/Copyright,2000\(1\)%20\(2\).pdf](http://copyrightoffice.portal.gov.bd/sites/default/files/files/copyrightoffice.portal.gov.bd/law/121de2e9_9bc9_4944_bfef_0a12af0864a5/Copyright,2000(1)%20(2).pdf)

Quality Assurance of the Dataset: All the annotations were done by native Bangla speakers. The Fleiss' Kappa score of our dataset showed substantial agreement, ensuring the quality of our dataset.

7. Bibliographical References

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.
- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.
- Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. [BanglaParaphrase: A high-quality Bangla paraphrase dataset](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 261–272. Online only. Association for Computational Linguistics.
- Abdullah Al Hadi, Md. Yasin Ali Khan, and Md. Abu Sayed. 2016. [Extracting semantic relatedness for bangla words](#). In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 10–14.
- Firoj Alam, Shammur Absar Chowdhury, and Sheak Rashed Haider Noori. 2016. Bidirectional lstms—crfs networks for bangla pos tagging. In *19th International Conference on Computer and Information Technology (ICCIT), 2016*, pages 377–382. IEEE.
- American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.
- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. [Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection](#). *Computational Linguistics*, 39(4):917–947.
- Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*.
- Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. [Information fusion in the context of multi-document summarization](#). In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA. Association for Computational Linguistics.
- Dario Bertero and Pascale Fung. 2015. Hlhc-hkust: A neural network paraphrase classifier using translation metrics, semantic roles and lexical similarity features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 23–28.
- Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.
- Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md Saiful Islam, Wasi Uddin Ahmad, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. [Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla](#).
- Ergun Biçici and Andy Way. 2014. Rtm-dcu: Referential translation machines for semantic similarity.
- Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *RANLP*, volume 97.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.
- Andreas Chandra and Ruben Stefanus. 2020. [Experiments on paraphrase identification using quora question pairs dataset](#).
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](#). *Journal of the Association for Computing Machinery*, 28(1):114–133.
- David Chen and William Dolan. 2011. [Collecting highly parallel data for paraphrase evaluation](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.
- Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.
- Jianpeng Cheng and Dimitri Kartsaklis. 2015. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. *arXiv preprint arXiv:1508.02354*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Seniz Demir, İlknur Durgar El-Kahlout, Erdem Unal, and Hamza Kaya. 2012. [Turkish paraphrase corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 4087–4091, Istanbul, Turkey. European Language Resources Association (ELRA).
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitriy Popov, Dmitry V. Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobrelev, Yacine Jernite, Thomas Wolf, and Gennady Pekhimenko. 2021. [Distributed deep learning in open collaborations](#). *CoRR*, abs/2106.10207.
- William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.
- William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. *arXiv preprint arXiv:2101.08382*.
- Asli Eyecioglu and Bill Keller. 2015. Asobek: Twitter paraphrase identification with simple overlap features and svms in proceedings of 9th international workshop on semantic evaluation (semeval).
- Asli Eyecioglu and Bill Keller. 2016. Constructing a turkish corpus for paraphrase identification and semantic similarity. *Lecture Notes in Computer Science*, Computational Linguistics and Intelligent Text Processing. Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics.:562–574.
- Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013a. [Paraphrase-driven learning for open question answering](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria. Association for Computational Linguistics.
- Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013b. [Paraphrase-driven learning for open question answering](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618.
- Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics*, pages 45–52.
- Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the third international workshop on paraphrasing (IWP2005)*.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020a. [Parade: A new dataset for paraphrase identification requiring computer science domain knowledge](#).
- Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020b. [Parade: A new dataset for paraphrase identification requiring computer science domain knowledge](#). *arXiv preprint arXiv:2010.03725*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Kyuyeon Hwang and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5720–5724. IEEE.
- Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

- Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. data. quora. com.
- Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891–896.
- Rafael-Michael Karampatsis. 2015. Cdtts: Predicting paraphrases in twitter via support vector regression. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 75–79.
- Mladen Karan, Goran Glavaš, Jan Šnajder, Bojana Dalbelo Bašić, Ivan Vulić, and Marie-Francine Moens. 2015. Tklbliir: Detecting twitter paraphrases with tweetingjay. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 70–74. ACL; East Stroudsburg, PA.
- Dmitry Kravchenko. 2017. Paraphrase detection using machine translation and textual similarity algorithms. In *Conference on artificial intelligence and natural language*, pages 277–292. Springer.
- Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. [Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages](#).
- Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *arXiv preprint arXiv:1708.00391*.
- Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.
- Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#).
- Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012a. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190.
- Nitin Madnani, Joel R. Tetreault, and Martin Chodorow. 2012b. Re-examining machine translation metrics for paraphrase identification. In *NAACL*.
- Alaa Altheneyan; Mohamed Menai. 2019. [Arpc a corpus for paraphrase identification in arabic text](#).
- Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. [A monolingual approach to contextualized word embeddings for mid-resource languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.
- Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019a. Improving semantic similarity with cross-lingual resources: A study in bangla—a low resourced language. In *Informatics*, volume 6, page 19. Multidisciplinary Digital Publishing Institute.
- Rajat Pandit, Saptarshi Sengupta, Sudip Kumar Naskar, Niladri Sekhar Dash, and Mohini Mohan Sardar. 2019b. [Improving semantic similarity with cross-lingual resources: A study in bangla—a low resourced language](#). *Informatics*, 6(2).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. [Bleu: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02,

- page 311–318, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Diana Pérez and Enrique Alfonseca. 2005. Application of the bleu algorithm for recognising textual entailments. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment*, pages 9–12. Citeseer.
- Saša Petrović, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 338–346.
- Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.
- Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*, pages 997–1005.
- Ekaterina Pronoza, Elena Yagunova, and Anton Pronoza. 2016. *Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction*, volume 573, pages 146–157.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. *Yara parser: A fast and accurate dependency parser*. *Computing Research Repository*, arXiv:1503.06733. Version 2.
- Dwijen Rudrapal, Amitava Das, and Baby Bhattacharya. 2015. Measuring semantic similarity for bengali tweets using wordnet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 537–544.
- Vasile Rus, Rajendra Banjade, and Mihai C Lintean. 2014. On paraphrase identification corpora. In *LREC*, pages 2422–2429. Citeseer.
- Sagor Sarker. 2020. *Banglabert: Bengali mask language model for bengali language understanding*.
- Taneeya Satyapanich, Hang Gao, and Tim Finin. 2015. Ebiquty: Paraphrase and semantic similarity in twitter using skipgrams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 51–55.
- Yves Scherrer. 2020. *TaPaCo: A corpus of sentential paraphrases for 73 languages*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.
- Md Shajalal and Masaki Aono. 2018. *Semantic textual similarity in bengali text*. pages 1–5.
- Manjira Sinha, Tirthankar Dasgupta, Abhik Jana, and Anupam Basu. 2014. Article: Design and development of a bangla semantic lexicon and semantic similarity measure. *International Journal of Computer Applications*, 95(5):8–16. Full text available.
- Manjira Sinha, Abhik Jana, Tirthankar Dasgupta, and Anupam Basu. 2012. *A new semantic lexicon and similarity measure in Bangla*. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 171–182, Mumbai, India. The COLING 2012 Organizing Committee.
- Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural paraphrase identification of questions with noisy pretraining. *arXiv preprint arXiv:1704.04565*.
- Rob van der Goot and Gertjan van Noord. 2015. Rob: Using semantic meaning to recognize paraphrases. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 40–44.
- Ngoc Phuoc An Vo, Simone Magnolini, and Octavian Popescu. 2015. Fbk-hlt: An effective system for paraphrase identification and semantic similarity in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 29–33.
- P. Wallis. 1993. Information retrieval based on paraphrase.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan

- Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015a. *SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT)*. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.
- Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015b. *Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit)*. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.
- Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsoulis. 2011. *Linguistic redundancy in twitter*. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 659–669.
- Guido Zarrella, John Henderson, Elizabeth Merkhofer, and Laura Strickhart. 2015. *Mitre: Seven systems for semantic similarity in tweets*. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 12–17.
- Bowei Zhang, Weiwei Sun, Xiaojun Wan, and Zongming Guo. 2019. *Pku paraphrase bank: A sentence-level paraphrase corpus for chinese*. In *NLPCC*.
- Jiang Zhao and Man Lan. 2015. *Ecnu: Leveraging word embeddings to boost performance for paraphrase in twitter*. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 34–39.
- Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2009. *Extracting paraphrase patterns from bilingual parallel corpora*. *Nat. Lang. Eng.*, 15(4):503–526.
- Chao Zhou, Cheng Qiu, and Daniel E. Acuna. 2022. *Paraphrase identification with deep learning: A review of datasets and methods*.
- Donglai Zhu, Hengshuai Yao, Bei Jiang, and Peng Yu. 2018. *Negative log likelihood ratio loss for deep neural network classification*.
- ## 8. Language Resource References
- Akil, Ajwad and Sultana, Najrin and Bhattacharjee, Abhik and Shahriyar, Rifat. 2022. *BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset*. Association for Computational Linguistics. PID <https://doi.org/10.48550/arXiv.2210.05109>.
- Dolan, Bill and Brockett, Chris. 2005. *Automatically Constructing a Corpus of Sentential Paraphrases*. Asia Federation of Natural Language Processing, Third International Workshop on Paraphrasing (IWP2005). PID <https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/>.
- He, Yun and Wang, Zhuoer and Zhang, Yin and Huang, Ruihong and Caverlee, James. 2020. *PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge*. Association for Computational Linguistics. PID <https://doi.org/10.18653/v1/2020.emnlp-main.611>.
- Kumar, Aman and Shrotriya, Himani and Sahu, Prachi and Mishra, Amogh and Dabre, Raj and Puduppully, Ratish and Kunchukuttan, Anoop and Khapra, Mitesh M. and Kumar, Pratyush. 2022. *IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages*. Association for Computational Linguistics. PID <https://doi.org/10.18653/v1/2022.emnlp-main.360>.
- Scherrer, Yves. 2020. *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages v.1*. Zenodo. PID <https://doi.org/10.5281/zenodo.3707949>.
- Xu, Wei and Callison-Burch, Chris and Dolan, Bill. 2015. *SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)*. Association for Computational Linguistics. PID <https://doi.org/10.18653/v1/S15-2001>.

9. Appendix

9.1. Source Portals for Data Collection

Name	Global Ranking	Country Ranking
prothomalo.com	500	4
jugantor.com	1,193	5
kalerkantho.com	1,646	6
jagonews24.com	1,691	7
bdnews24.com	1,573	8
bd-pratidin.com	2,106	12
banglanews24.com	3,238	16
dhakapost.com	4,545	17
banglatribune.com	3,319	18
ittefaq.com.bd	3,652	21
samakal.com	7,497	27
24livenewspaper.com	7,811	35
rtvonline.com	8,901	36
somoynews.tv	5,275	37
newsbangla24.com	10,987	40
dainikshiksha.com	10,417	41
ntvbd.com	8,935	43
dailyinqilab.com	9,745	44
anandabazar.com	3,415	50
mzamin.com	12,376	63
priyo.com	33,966	169
abplive.com	2,353	227

Table 9: Alexa ranking of different news portals. (Collected on 08 October, 2021)

We used the Alexa ranking¹³ to gather news from the most popular sites in the national and international domains. The global ranking and ranking in Bangladesh of the news portals are shown in Table 9.

9.2. Discarded Sentence Pair Examples

While annotating the dataset, we found some sentence pairs where the annotators could not agree if it was a paraphrase or not. We called these sentence pairs debatable. After careful analysis, we found that these sentence pairs are usually partial paraphrases, have partial information of the other sentence, or have uncertain sentence pairs.

- **Partial Paraphrases:** Partial paraphrase occurs when a section of a complex sentence incorporates the paraphrase of another sentence.
- **Partial Information:** One sentence lacks some information, making it impossible to determine if it is a paraphrase or not.

- **Generalization:** Certain phrases is generalized in one sentence, while it is specific in the other one.

All these issues create a problem to properly classify a pair as a paraphrase or not. Some debatable sentence pairs are added in Table 10.

¹³<https://www.alexa.com/topsites/countries/BD>

Sentence 1	Sentence 2	Reason
কোহলির বেঙ্গালুরুর এবারও খালি হাতে বিদায় (Kohli's Bangalore left empty handed this time)	কোহলিদের বিদায়, টিকে থাকল হায়দরাবাদ (Farewell to Kohli, Hyderabad survived)	Partial Paraphrase
জরিপে এগিয়ে বাইডেন, এরপরও ট্রাম্প যেভাবে জিততে পারেন (Biden ahead in the polls, yet how can Trump win)	ট্রাম্প যেভাবে জয়ী হতে পারেন (The way Trump can win)	
সম্মাননা পেলেন অপূর্ব-মেহজাবীন (Apurba-Mehzabin got the honor)	মেহজাবীনের হাতে সম্মাননা (Honor in the hands of Mehzabin)	Partial Information
নতুন দায়িত্বে আফসানা মিমি (Afsana Mimi in new responsibilities)	শিল্পকলা একাডেমির পরিচালকের দায়িত্বে মিমি ও মিনি (Mimi and Mini are the directors of Shilpakala Academy)	
চাবির ঘা' ইউনিটের ভর্তি পরীক্ষা না নেয়ার সিদ্ধান্ত (Decision not to take admission test of DU D unit)	চাবির 'ঘা' এবং 'চ' ইউনিট থাকছে না (DU does not have 'D' and 'F' units)	Generalization
মুম্বাইয়ে হোটেলে অজি ক্রিকেটার ডিন জোসের মৃত্যু (Aussie cricketer Dean Jones dies at hotel in Mumbai)	ধরাভাষ্য দিতে এসে অকালেই হৃদরোগে আক্রান্ত হয়ে প্রয়াত প্রখ্যাত ক্রিকেটার (The late famous cricketer suffered a heart attack prematurely when he came to comment)	
১০০ ছুইছুই বেশিরভাগ সবজি (Most vegetables touches 100)	কমোনি পেয়াজের বাজি, সবজির বাজারও চড়া (The market for onions and vegetables is also booming)	
যুক্তরাষ্ট্র থেকে ২২৯০ কোটি রুপির অস্ত্র কিনছে ভারত (India is buying arms worth Rs 2,290 crore from the United States)	আমেরিকা থেকে অতিরিক্ত ৭২,০০০ অ্যাসল্ট রাইফেল কিনবে ভারত (India will buy an additional 62,000 assault rifles from the United States)	

Table 10: Examples of debatable sentence pairs.

Creating Clustered Comparable Corpora from Wikipedia with Different Fuzziness Levels and Language Representativity

Anna Laskina, Eric Gaussier, Gaelle Calvary

Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG

38000 Grenoble, France

{anna.laskina, eric.gaussier, gaelle.calvary}@univ-grenoble-alpes.fr

Abstract

This paper is dedicated to the extraction of clustered comparable corpora from Wikipedia, that is comparable corpora with labelled information corresponding to the topics associated to each document. Despite the importance of such corpora for evaluating text clustering and classification methods in the context of comparable corpora, there is a notable absence of automatic algorithms capable of creating them with adjustable fuzziness levels and language representativity. The methodology we propose here offers control over the cluster distribution across languages, enables fine-tuning of fuzziness levels, and facilitates customization to accommodate specific subject areas. Moreover, we have developed a dedicated tool specifically designed for our purpose and present 18 bilingual clustered comparable corpora spanning English, French, German, Russian, and Swedish languages. The analysis of these corpora demonstrates the effectiveness and flexibility of the approach in constructing corpora with varying levels of fuzziness and language representativity. Our results, tool and corpora, pave the way to construct various gold standard collections for future research in clustering and classification in comparable corpora.

Keywords: Comparable corpora, Gold standard collections, Text clustering/text classification, Wikipedia

1. Introduction

Our objective in this study is to provide a tool to automatically extract from Wikipedia comparable corpora with clustering information, each cluster corresponding to a meaningful Wikipedia category called *topic* afterwards. We refer in the remainder to such corpora as *clustered comparable corpora*. As such, this study is part of a broader initiative on clustering comparable corpora where gold standard collections are needed in order to compare different clustering approaches and methods. It has to be noted that such collections can also be used for studying, in a (semi-)supervised setting, text classification in comparable corpora.

Wikipedia stands out as a well-known and freely available public resource, offering a vast array of texts in multiple languages. Moreover, the texts in Wikipedia covering similar topics are intricately linked and categorized in the same high-level categories, facilitating the construction of coherent and comprehensive comparable corpora. In addition, as many articles cover different topics and belong to different Wikipedia categories, it is possible to construct clustered comparable corpora in which documents can have different *fuzziness levels*, *i.e.*, be assigned to one or more clusters, enabling more nuanced analysis and interpretation of the data. Lastly, in, for example, the context of bilingual comparable corpora involving two languages ℓ_1 and ℓ_2 , for a given set of topics, it is possible to extract from Wikipedia different clustered comparable corpora with different proportions of clusters containing only documents in ℓ_1 , only documents in ℓ_2 , or a mixture

of documents in ℓ_1 and documents in ℓ_2 .

Based on the above considerations, we aim in this study at developing a methodology and an associated suite of tools to extract clustered comparable corpora from Wikipedia while offering to researchers the possibility to:

- Tailor such corpora to specific subjects,
- Regulate their fuzziness levels,
- Control the proportions of monolingual and multilingual clusters.

Through the integration of these elements, researchers can access richer, more diverse datasets, thereby advancing the frontiers of data-driven inquiry and analysis in comparable corpora. In particular, they can use the collected datasets for evaluating clustering and/or classification methods. The described methodology can be applied to any other knowledge base with a similar structure to Wikipedia when the need arises to create collections with different knowledge from Wikipedia. For simplicity, we focus here on the construction of clustered *bilingual* comparable corpora. The extension to multilingual corpora is nevertheless direct.

The structure of this paper is organized as follows. Section 2 provides an overview of related work in the field of extracting comparable corpora from Wikipedia. Section 3 presents our proposed methodology, which consists of two main components: extracting a category tree from the Wikipedia category graph (Section 3.1) and building clustered bilingual comparable corpora (Section 3.2). Section 4 presents our results, consisting of a tool we

developed (Section 4.1) and an analysis of several collected corpora (Section 4.2). Finally, Section 5 concludes the paper by summarizing the key findings and outlining potential avenues for future research in this area.

2. Related Work

Wikipedia is widely used across different domains, making it a suitable primary data source for extracting comparable corpora. Several studies have utilised Wikipedia data for dictionary extraction (Chu et al., 2014; Erdmann et al., 2008; Yu and Tsujii, 2009) and machine translation tasks (Ramesh and Sankaranarayanan, 2018; Ruiter et al., 2019; Alegria et al., 2013). Wikipedia data is commonly used to train pre-trained models, in particular word embeddings and language models. Examples of pre-trained word representations that use Wikipedia text corpora include fastText (Mikolov et al., 2018), BERT (Devlin et al., 2019) and LASER (Artetxe and Schwenk, 2019). Recent advancements in large language models (LLMs), such as GPT-3 (Brown et al., 2020), mT6 (Chi et al., 2021), llama (Touvron et al., 2023), and LaMDA (Thoppilan et al., 2022), have been also incorporating Wikipedia data into their training processes.

There are several works in Wikipedia-based comparable corpora. Although some efforts concentrate on collecting parallel sentences (Plamada and Volk, 2012; Plamadă and Volk, 2013) or pairs of articles (Saad et al., 2013; Goyal et al., 2020) in multiple languages to create comparable corpora, these endeavors are mainly aimed at machine translation applications rather than clustering and classification tasks, which are aligned with our objectives.

When exploring methodologies for creating comparable collections from Wikipedia, various works strive to gather comparable collections for a specific language pair and a specific topic. These works vary mainly in their document selection process for the chosen topic. In (Otero and López, 2010; Otero et al., 2011), the authors align topics with specific Wikipedia categories and considered three possible options for the comparability of documents in different languages: documents belonging to the same topic because they have the same associated category (non-aligned), documents connected by an inter-language link (softly-aligned), and documents connected by an inter-language link and belonging to the same category (strongly-aligned). A limitation of this approach is that it focuses on documents directly related to the selected category, which limits the size of the corpus and poses challenges in assembling larger corpora. According to (Barrón-Cedeno et al., 2015), an alternative approach involves selecting documents that are not only directly related to a topic associated with

Wikipedia categories but also those associated with its subcategories. We intend to adopt this strategy. A recent study (España-Bonet et al., 2023) proposed an approach to improve the selection of documents from subcategories of the Wikipedia category associated with the topic. This was achieved by using a vocabulary that describes the topic and retaining only those subcategories whose titles contain at least one word from the vocabulary. While acknowledging its advantages, we have decided not to employ this approach in this paper. This is mainly caused by the topic vocabularies, which can number over a hundred and vary depending on the collection topic, fuzziness levels, and language representation. Nevertheless, we do intend to explore its potential inclusion in future work. That said, none of these methods aims at building clustered comparable corpora and the methodology we propose in this paper is the first one, as far as we know, to address this problem.

3. Methodology

We describe in this section the methodology followed to extract clustered bilingual comparable corpora from Wikipedia. It relies on a first step that creates a category tree from the Wikipedia category graph to determine appropriate topics for labeling a corpus. The second step involves creating the corpus according to the specified preferences regarding language representativity and fuzziness.

3.1. From a Category Graph to a Category Tree

Each page in Wikipedia typically has multiple categories, which are organised into a hierarchical graph known as the Wikipedia category graph (Hecht and Gergle, 2010). The Wikipedia category graph, which has been the subject of many studies (Zesch and Gurevych, 2007; Suchecki et al., 2012; Aspert et al., 2019), contains numerous cycles (España-Bonet et al., 2023; Barrón-Cedeno et al., 2015) in the sense that a category can refer to itself as a parent category after several generations. For instance, the category *Soil* serves as both a parent and a subcategory of the category *Soil science*, creating a cycle of *Soil* → *Soil science* → *Soil*. This said, it has been shown that the Wikipedia category graph can be hierarchized by identifying a root and organising the graph into hierarchy levels according to the length of directed paths from the root (Aouicha et al., 2016; Aghaebrahimian et al., 2022). Following this idea, the category *Main Topic Classifications* has often been chosen as the root category and has therefore been assigned a level of 0. Note that this category was selected because it includes the main Wikipedia

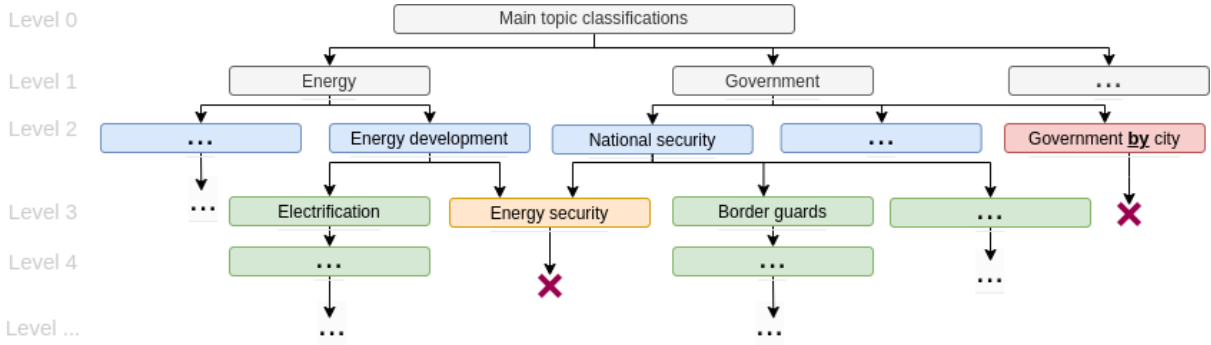


Figure 1: Part of the category tree showing the relationships between categories and their types: *insignificant* (red), *ambiguous* (orange) and *theme* (blue, green and gray). When the level of reference equals 2, the blue *theme* categories generate clusters in the corpus, while the green *theme* categories are used when searching for clusters in documents. The grey *theme* categories are not used in corpus generation because they are too general.

topics for categorisation, as provided by Wikipedia itself.

Several studies have further demonstrated that not all categories are adequate for creating a cluster (Aghaebrahimian et al., 2022). Relevant categories can either be selected manually from a list, typically limited to a few dozen categories (Plamadã and Volk, 2013; Barrón-Cedeno et al., 2015), or extracted automatically by filtering out non relevant categories. The latter option is the most promising (España-Bonet et al., 2023) and is the one we adopt here.

The first aspect which differentiates relevant and non relevant categories relates to the fact that some categories indeed represent specific topics (e.g., *Music*, *History*), while others mainly serve for organizing the whole Wikipedia collection (e.g., *Outlines of general reference*). The latter categories, precisely defined below, are not suitable for clustering documents into topics and are referred to here as *insignificant* categories.

Definition 3.1 A category, the name of which contains any of the words *by*, *in*, *from*, *about*, *and*, *after*, *list*, *award*, *image*, *quotation*, *event*, *outline*, *redirect*, *people* is called an *insignificant category*.

The second aspect relates to the fact that the Wikipedia category graph contains duplicates of categories with different parents at the same distance from the root. The distance considered here is the length of the shortest path from a given node to the root. These categories are *ambiguous* in that they are equally relevant to all of their parent categories equally distant from the root and do not exhibit a stronger association with any one of them.

Definition 3.2 A category that has more than one parent category equally distant from the root category is called *ambiguous*.

We focus here on all categories but insignificant

and ambiguous categories for creating clustered comparable corpora. Such categories are called *theme* categories in the remainder of the paper:

Definition 3.3 A category which is neither insignificant, nor ambiguous is called a *theme category*.

In Figure 1, the category *Government by city* is *insignificant* because it contains the word *by*, the category *Energy security* is *ambiguous* as it is both a subcategory of *Energy development* and *National security*, which are both level 2 categories, and the categories *Electrification* and *National security* are *theme* categories because they are neither *insignificant* nor *ambiguous*.

Finally, the process we rely on to create a category tree \mathcal{T} aims at filtering out the Wikipedia category graph by removing cycles and relying only on theme categories, thus obtaining a tree backbone of the Wikipedia category graph fully suited to topical clustering/classification. It goes as follows:

1. Set the root node c_0 of \mathcal{T} to the category *Main Topic Classifications*; set the level l to 0.
2. Recursively add to \mathcal{T} , in a breadth-first manner and at level $(l + 1)$, all subcategories of all theme categories present at level l in \mathcal{T} if they are not already in \mathcal{T} ; mark as insignificant and ambiguous the added subcategories complying with definitions 3.1 and 3.2.

Note that one can easily check if a category is ambiguous by verifying if it is already present in \mathcal{T} at the same level. This process naturally stops at level 22.

Table 1 displays the different types of categories at the different levels of the tree obtained by the above process. As one can note, there are 39 theme categories at level 1, 825 at level 2 and 5539 at level 3. To ensure homogeneity between clusters in the final corpus, we consider as original

Level	<i>insignificant</i>	<i>ambiguous</i>	<i>theme</i>
1	2	0	39
2	341	46	825
3	1542	921	5539
4	4814	2690	16914
5	12032	4769	38390
6	25251	12647	54742
7	29710	8896	69122
8	35695	10195	59671
9	28389	6392	53236
10	23272	5759	41600
11	19065	2527	27797
12	9767	1039	15472
13	4145	415	10345
14	4050	317	6440
15	1852	177	2541
16	1172	78	1275
17	342	19	332
18	83	9	54
19	4	0	9
20	0	0	3
21	0	0	4
22	0	0	0

Table 1: The amount of *insignificant*, *ambiguous* and *theme* categories in the category tree by level. From level 22 there are no more *theme* categories.

topics to construct clustered bilingual comparable corpora theme categories at the same level, which will be referred to as l_r for *level of reference*:

Definition 3.4 A theme category at level l_r in the category tree is called a topic. Furthermore, any Wikipedia category c as well as the Wikipedia documents assigned to it belong to a topic t if $c = t$ or c is a descendant of t in \mathcal{T} .

Different levels of reference can be used depending on the balance between coarse-grained and fine-grained clusters one is interested in. Lastly, the clusters (or classes if one is rather interested in text classification) we consider for constructing clustered (or categorized) bilingual comparable corpora are a subset of the topics defined above. As described below, we will rely on both primary and secondary topics to obtain our clusters.

3.2. Corpus Creation

We consider here the creation of a clustered bilingual comparable corpus where documents are written in either language l_1 or language l_2 . Such a corpus can display three types of clusters: clusters of type 1 (resp. 2) containing only documents written in l_1 (resp. l_2) and documents of type 1&2 containing both documents written in l_1 and documents written in l_2 . Of course, all types may not be represented in every clustered bilingual compara-

ble corpus; in addition, for simplicity, we focus here on clustered corpora in which a document can only belong to clusters of the same type.

In order to control the representativity of each language in the corpus to be created, we define three hyperparameters, denoted Nt_1 , Nt_2 and $Nt_{1\&2}$, which specify the number of *primary* topics of type 1, 2 and 1&2 one is interested in. Each number Nt_1 , Nt_2 and $Nt_{1\&2}$ can either be set directly or be randomly chosen from a set of values defined by the user (see Table 2 for example). Furthermore, we allow users the possibility to have clusters of different sizes by randomly selecting, from a given set of values, the number of documents Nd_{ij} associated to the j^{th} topic of type i ($i \in \{1, 2, 1\&2\}$).

In addition to controlling the language representativity, we also want to control the overall degree of fuzziness of documents across clusters. To this end, we introduce two additional hyperparameters, f_{min} and f_{max} , which respectively represent the minimum and maximum numbers of clusters a document should belong to. f_{min} is lower bounded by 1 and upper bounded by f_{max} , whereas f_{max} is lower bounded by f_{min} and upper bounded by the maximum number of topics a document can have in Wikipedia. Both f_{min} and f_{max} are defined by the user.

Collecting the Nd_{ij} documents for the j^{th} topic of type i with a fuzziness degree comprised between f_{min} and f_{max} can be done in a natural way by (a) recursively considering all theme sub-categories of the given topic in the constructed tree \mathcal{T} (see previous section), (b) randomly selecting all documents in the subcategory with at least f_{min} and at most f_{max} different topics till Nd_{ij} documents are collected, and (c) adding to the collected corpus the documents which belong to topics different from topics of types different from i . A question however arises when doing so when $f_{max} > 1$: for a given document, should one keep all the topics it belongs to or should one just disregard the ones different from the original j^{th} topic of type i ? Disregarding such topics can be detrimental to our purpose of constructing gold standard clustered bilingual comparable corpora as one may lose valuable information relating documents (through disregarded topics) which are finally placed in different topics while being strongly related. We thus propose here to keep them, referring to them as *secondary* topics, and consider them as new clusters of type i . The final set of clusters thus comprises both primary and secondary topics, the latter being added to the former when collecting documents. Because of this addition, Nt_1 , Nt_2 and $Nt_{1\&2}$ correspond to lower bounds of the actual number of clusters obtained, as illustrated in Table 2. However, as one can note, in most cases, as the number of topics per documents is limited in Wikipedia, one ends

ID	Language pair	Doc.	# of clusters	# of primary topics	T/D	D/T	f_{min}	f_{max}	Order	# of clusters per type	# of documents per cluster
Na01	De-Fr	6982	49	43	1.19	170.12	1	5	(2, 1&2, 1) †	{10, 15, 20}	{200, 500, 750}
Na02	Fr-Sw	6811	63	49	1.52	164.51	1	5	(2, 1&2, 1) †	{10, 15, 20}	{200, 500, 750}
Na03	De-En	4415	33	31	1.20	160.48	1	10	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 250, 500}
Na04	Fr-Ru	3386	35	32	1.16	111.80	1	10	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 250, 500}
Na05	Fr-Ru	5636	20	20	1.00	281.80	1	1	(2, 1&2, 1) †	{5, 10}	{100, 150, 250, 500}
Na06	En-De	5139	19	19	1.00	270.47	1	1	(2, 1&2, 1) †	{5, 10}	{100, 150, 250, 500}
Na07	En-Fr	2726	18	17	1.06	160.06	1	10	(1, 2, 1&2) †	{10, 15, 20}	{100, 150, 200, 250}
Na08	En-Fr	3255	21	19	1.14	176.10	1	10	(1&2, 2, 1) †	{10, 15, 20}	{100, 150, 200, 250}
Na09	En-Fr	2578	17	15	1.26	190.35	1	10	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 200, 250}
Na10	En-Fr	2677	122	34	2.15	47.25	2	10	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 200, 250}
Na11	En-Fr	3466	24	22	1.14	164.04	1	100	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 200, 250}
Na12	En-Fr	3411	17	17	1.00	200.65	1	1	(2, 1&2, 1) †	{10, 15, 20}	{100, 150, 200, 250}
Na13	En-Fr	14617	31	31	1.00	471.52	1	1	(1, 1&2, 2)	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
Na14	En-Fr	25813	119	71	1.73	374.82	1	10	(1, 2, 1&2)	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
Na15	En-Fr	6460	212	21	2.98	90.92	2	10	(1, 2, 1&2)	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
Na16	En-Fr	13544	70	63	1.13	218.74	1	10	(2, 1&2, 1)	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
Na17	En-Fr	20106	94	60	1.48	315.89	1	10	(1, 2, 1&2) †	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}
Na18	En-Fr	30932	113	57	2.00	547.71	1	10	(2, 1&2, 1) †	{10, 15, 20, 25, 30}	{100, 250, 500, 750, 1000, 1250, 1500, 2000}

Table 2: This table provides details for comparable corpus, including the language pair, number of documents, number of primary topics, overall number of topics, average number of topics per document (T/D), and average number of documents per topic (D/T). The section on the right-hand side provides information on creating a corpus. This includes the minimum and maximum number of topics in documents, the order in which topic types are collected, and the range for randomly selecting the number of topics of each type and the number of documents in a topic. A special sign (†) means alternating order, while no sign means consideration by type. The level of reference l_r is 2 for all corpora.

up with a number of clusters relatively close to the original number set by the user.

In the process described above, in accordance with our will to construct clustered corpora in which documents only belong to clusters of the same type, one has to check, for every selected document, whether it belongs to clusters of different types or not. If this verification is simple, it raises the question of the ordering in which the different types of clusters are considered. Indeed, the more versatile topics, *i.e.*, the topics being commonly assigned with other topics, are more likely to be encountered at the beginning of the above process than at its end. Such topics also impact the fuzziness degree as they are likely to be present in the documents selected. We thus allow the user to play with possible orderings, firstly by deciding in which order the different types should be considered¹, and secondly by deciding to either process all topics in a given type before moving to the other types, or alternate between types after each topic. These different configurations are also illustrated in Table 2.

4. Results & Discussion

This section presents the results of our study, which consists of two main components. Firstly, technical details about the tool used and its application are provided. Secondly, the bilingual comparable corpora created with the tool are analysed to identify whether control over the number of clusters represented in only one language or both languages, the

fuzziness, and the ability to adapt the corpora to a particular domain were achieved.

4.1. Tool

The code was implemented in Python (v.3.8.10), using requests (v.2.27.1), beautifulsoup4 (v.4.10.0), numpy (v.1.21.6) libraries and is freely available². Information from the Wikipedia pages was obtained through MediaWiki API³. The tool has three functions: creating a category tree, creating a clustered bilingual comparable corpus, and visualising an obtained corpus. During the creation of the category tree, two adjustable parameters are available: a root category and the level of reference l_r used for selecting topics and thus clusters. When initiating corpus creation, several parameters can be considered, including the language pair for the collection, the range of topics present in the documents through the parameters f_{min} and f_{max} , the order in which topics are considered, the number of topics of each type, and the number of documents in each cluster. The last two parameters may either be a specific number or a set of values from which one value will be randomly selected. Additional details can be found on the code repository.

4.2. Collected Corpora

Creating a comparable corpus using Wikipedia as a source enables the development of topic-specific

¹There are six possible choices for that: (1&2, 1, 2), (1&2, 2, 1), (1, 1&2, 2), (1, 2, 1&2), (2, 1&2, 1), and (2, 1, 1&2).

²https://github.com/anna-laskina/comparable_corpora_generator

³https://www.mediawiki.org/wiki/API:Main_page

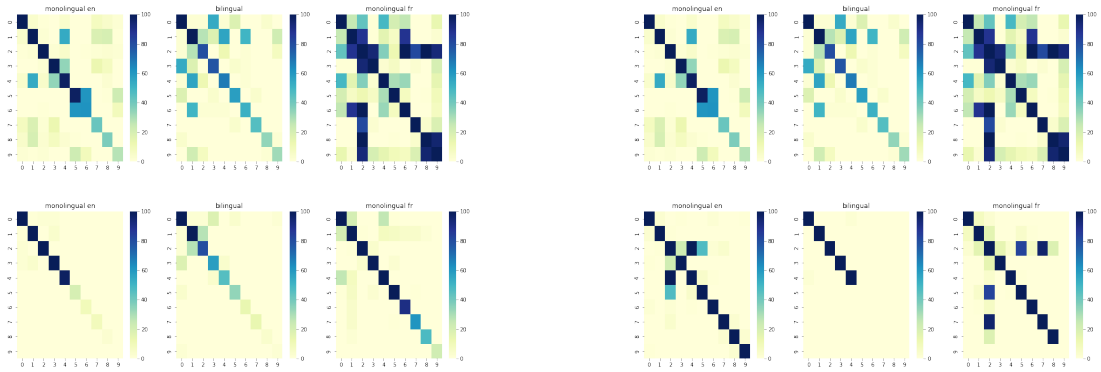


Figure 2: Heat map of the number of documents shared by the 10 largest topics of type monolingual En (left), bilingual (middle), and monolingual Fr (right) for corpus №10. Top: primary and secondary topics; bottom: primary topics only.

corpora, particularly in languages with a substantial representation in Wikipedia, such as English, German, French, and Swedish. This study focuses on the English-French language pair, but also includes corpora for other language pairs such as English-German, French-Russian, French-German, and French-Swedish. A series of corpora were generated to analyse the effectiveness of the corpus generation algorithm. In this paper, we provide detailed descriptions of 18 corpora, with pertinent information delineated in Table 2.

There are two types of obtained topics: primary and secondary. Primary topics are those initially selected when the collection began, while secondary topics are those that appeared during the collection process when an article with unreferenced topics was added to the collection; these topics were added as new clusters and acquired the type as the requested topic. Considering the top 10 topics of each type reveals that secondary topics exhibit a greater dispersion of documents beyond the main diagonal (Fig. 2). This observation suggests that primary topics tend to be more coherent, with fewer documents containing multiple primary topics. In contrast, secondary topics introduce fuzziness into documents, facilitating a higher incidence of multiple topics within a single document.

Subsequently, we examined three datasets initialized with different values of f_{min} and f_{max} alongside consistent remaining parameters to discern the varying degrees of fuzziness attainable. Fixing f_{max} at 1 yields completely non-fuzzy (hard) clustering, depicted in Figure 3 (bottom). Conversely, selecting f_{min} at 2 and f_{max} at 10 facilitates achieving fuzzy clustering, as illustrated in Figure 3 (top). The mean number of topics per document across corpora ranges from 1.00 to 2.98 (Table 2), whereas within a single corpus, the number of topics per doc-

Figure 3: Heat map of the number of documents shared by the 10 largest topics of type monolingual En, bilingual, and monolingual Fr (from right to left respectively) across corpora №10, №11, №12 (from top to bottom), run with the same parameters, except f_{max} , which is equal to 10, 100, 1 for these corpora respectively, and f_{min} , which is equal to 1 for corpus №11 and №12, and equal to 2 for corpus №10.

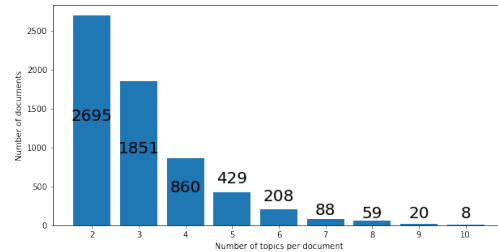


Figure 4: The distribution of documents in corpus №15 by the number of topics present within it.

ument can reach the maximum value defined by f_{max} (see Figure 4).

When executing corpus collection with parameters varying solely in the order of topic consideration, it becomes evident that when topics are arranged by type style (as depicted by (1, 2, 1&2) on the top and (2, 1&2, 1) in the middle of Figure 5), fuzziness becomes concentrated in the monolingual ℓ_1 and monolingual ℓ_2 categories, respectively, as they were the first types considered. Conversely, when topics are arranged by alternating style (bottom of Figure 5), fuzziness is more evenly distributed across different types. However, achieving precise control over the localization of the fuzzier

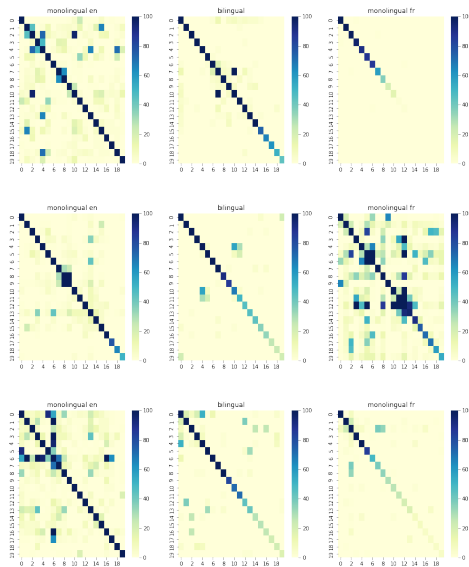


Figure 5: Heat map of the number of documents shared by the 20 largest topics of type monolingual En, bilingual, and monolingual Fr (from right to left respectively) across corpora №14, №16, №17 (from top to bottom), run with the same parameters, except the order in which topic types are considered: (1, 2, 1&2) *by type* for corpus №14, (1, 2, 1&2) *alternating* for corpus №17 and (2, 1&2, 1) *by type* for corpus №16.

segment becomes more challenging, as the second type of topics gains little advantage from being considered earlier than the final third type. Additionally, although the style of order in consideration influences the distribution of topics by types, a more significant correlation is observed initially from the selection of specific topics for each type.

Finally, customization of the category tree creation according to preferences is feasible. Ones have the option to select the root category and a set of topics for corpora. In our context, the category *Main topic classifications* was selected as the root category, as we did not have specific topic preferences. However, one can narrow the selected cluster to a particular area and choose, for example, the *Health* category as the root category (Fig. 6). The selection of the level of reference l_r in the obtained tree allows one to further focus on specific subcategories of, e.g., the Health domain.

5. Conclusion

We have presented in this paper a method to extract clustered bilingual comparable corpora from Wikipedia with different fuzziness levels and language representativity. Wikipedia is an excellent source for constructing such corpora because of its categorised articles and interlingual links, which

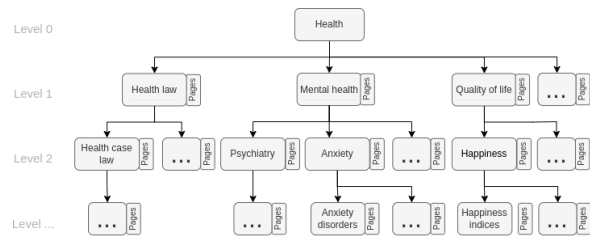


Figure 6: The category tree with the category *Health* as root.

facilitate the creation of bilingual links between articles. After extracting the topical tree backbone of the Wikipedia category system, we have proposed a construction process which allows one to somehow regulate the fuzziness level (*i.e.*, the fact that a document can be associated with more than one cluster) of the obtained corpus, as well the representativity of each language. Indeed, clustered bilingual comparable corpora are characterized by the fact that they contain three types of clusters: those consisting of documents in either language only, and those comprising documents from the two languages.

Our analysis has shown that it is possible to exert considerable influence over the above corpus characteristics, achieving significant control over fuzziness levels and language representativity, as well as determining the subject domain of the corpus. Future enhancements of the proposed methodology could include the method of collecting Wikipedia corpora on a particular topic proposed by [España-Bonet et al. \(2023\)](#). We also plan to extend our tool to directly construct clustered comparable corpora in more than two languages.

6. Acknowledgements

This work has been funded by the French projects ANR-20-IDES-0005 IDÉES@UGA and ANR-19-P3IA-0003 MIAI@Grenoble Alpes.

7. Bibliographical References

Ahmad Aghaebrahimian, Andy Stauder, and Michael Ustaszewski. 2022. Testing the validity of wikipedia categories for subject matter labelling of open-domain corpus data. *Journal of Information Science*, 48(5):686–700.

Iñaki Alegria, Unai Cabezón, Unai Fernandez de Betono, Gorka Labaka, Aingeru Mayor, Kepa Sarasola, and Arkaitz Zubiaga. 2013. Reciprocal enrichment between basque wikipedia and machine translation. *The People’s Web Meets*

- NLP: Collaboratively Constructed Language Resources*, pages 101–118.
- Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Malek Ezzeddine. 2016. Derivation of “is a” taxonomy from wikipedia category graph. *Engineering Applications of Artificial Intelligence*, 50:265–286.
- Mikel Artetxe and Holger Schwenk. 2019. [Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Nicolas Aspert, Volodymyr Miz, Benjamin Ricaud, and Pierre Vanderghenst. 2019. A graph-structured dataset for wikipedia research. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1188–1193.
- Alberto Barrón-Cedeno, Cristina España Bonet, Josu Boldoba Trapote, and Luís Márquez Villodre. 2015. A factory of comparable corpora from wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang, Saksham Singhal, Xian-Ling Mao, Heyan Huang, Xia Song, and Furu Wei. 2021. [mT6: Multilingual pretrained text-to-text transformer with translation pairs](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Computational Linguistics and Intelligent Text Processing*, pages 296–309, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from wikipedia. In *Proceedings of the 13th International Conference on Database Systems for Advanced Applications, DASFAA’08*, page 380–392, Berlin, Heidelberg. Springer-Verlag.
- Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Márquez. 2023. Tailoring and evaluating the wikipedia for in-domain comparable corpora extraction. *Knowledge and Information Systems*, 65(3):1365–1397.
- Vishal Goyal, Ajit Kumar, and Manpreet Singh Lehal. 2020. Document alignment for generation of english-punjabi comparable corpora from wikipedia. *International Journal of E-Adoption (IJE)*, 12(1):42–51.
- Brent Hecht and Darren Gergle. 2010. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 291–300.
- Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhresch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- P Otero, I López, S Cilenis, and Santiago de Compostela. 2011. Measuring comparability of multilingual corpora extracted from wikipedia. *Iberian Cross-Language Natural Language Processings Tasks (ICL)*, page 8.
- Pablo Gamallo Otero and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25.
- Magdalena Plamada and Martin Volk. 2012. Towards a wikipedia-extracted alpine corpus.
- Magdalena Plamadă and Martin Volk. 2013. [Mining for domain-specific parallel text from Wikipedia](#).

In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 112–120, Sofia, Bulgaria. Association for Computational Linguistics.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. [Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dana Ruiters, Cristina España-Bonet, and Josef van Genabith. 2019. [Self-supervised neural machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Motaz Saad, David Langlois, and Kamel Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences*, 95:40–47.

Krzysztof Suchocki, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnhorst. 2012. Evolution of wikipedia’s category structure. *Advances in complex systems*, 15(supp01):1250068.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kun Yu and Jun’ichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. In *Proceedings of Machine Translation Summit XII: Posters*.

Torsten Zesch and Iryna Gurevych. 2007. [Analysis of the Wikipedia category graph for NLP applications](#). In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 1–8, Rochester, NY, USA. Association for Computational Linguistics.

EuReCo: Not Building and Yet Using Federated Comparable Corpora for Cross-Linguistic Research

Marc Kupietz, Piotr Bański, Nils Diewald, Beata Trawiński, Andreas Witt

Leibniz Institute for the German Language (IDS)
R5 6-13, 68161 Mannheim, Germany
{kupietz, banski, diewald, trawinski, witt}@ids-mannheim.de

Abstract

This paper gives an overview of recent developments concerning the European Reference Corpus EuReCo, an open long-term initiative aimed at providing and using virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. Given the problems and shortcomings of other types of multilingual corpora – such as the shining-through effects in parallel corpora or the limitation to web material only in web-based comparable corpora – EuReCo constitutes a unique linguistic resource that offers new perspectives for fine-grained cross-linguistic research. The approach advocated here puts forward new solutions to notorious IPR and licensing issues, as well as to challenges of interoperability. It also addresses methodological questions concerning comparability and representativeness. While the focus of this paper is on EuReCo's implementation-based approach to ensuring interoperability in a feasible and maintainable way, it also presents preliminary results of pilot comparative studies on light verb constructions in German, Romanian, Hungarian, Polish and Bulgarian, and reports on recent extensions and plans.

Keywords: Reference Corpora, National Corpora, Federated Corpora, Multilingual Corpora, Cross-Linguistic Research, Comparability

1. Introduction

The challenge of comparability in multilingual studies relates both to the language data itself and to the methods applied. In this paper, we discuss the relevant features of the available corpus types from a linguistic perspective and point out their advantages and disadvantages, particularly for cross-linguistic research (Section 2). Against this background, we present an approach to using comparable corpora without having to build them: the European Reference Corpus EuReCo. EuReCo is an open long-term initiative that aims at providing and using virtual and dynamically definable comparable corpora based on existing national, reference or other large corpora. Section 3 presents the basic ideas behind EuReCo and the previous work. Section 4 introduces and discusses access to federated corpora and EuReCo's approach to interoperability, with the corpus analysis platform KorAP as a working implementation, and Section 5 presents recent developments within the EuReCo initiative, including applications in the area of cross-lingual studies of light verb constructions (Section 5.4). Section 6 summarizes the paper and sketches the next steps.

2. State of the Art

From the linguistic point of view, there exist several advantages and disadvantages of monolingual corpora, parallel corpora and the available comparable corpora. Based on Kupietz et al. (2020b)

and Trawiński and Kupietz (2021), we argue that there is a great need in cross-linguistic research for high-quality multilingual data whose degree and angle of comparability can be flexibly adjusted.

2.1. Monolingual Corpora

Monolingual corpora are, by definition, corpora that contain texts in a single language. They are characterized by a very high and controlled linguistic quality, as they typically contain (ideally only) original texts and thus reflect native language usage. There is currently a large number of monolingual corpora, including both (mostly smaller) specialized corpora and national or reference and other very large general corpora, such as the British National Corpus (BNC; Aston and Burnard, 1998; Brezina et al., 2018), the Corpus of Contemporary American English (COCA; Davies, 2011), the Czech National Corpus (CNC; Křen, 2020), the Romanian Contemporary Language Reference Corpus (CoRoLa; Barbu Mititelu et al., 2018), the German Reference Corpus (DeReKo; Kupietz et al., 2010, 2018), the Hungarian National Corpus (HNC; Váradi, 2002; Oravecz et al., 2014), and the Polish National Corpus (NKJP; Przepiórkowski et al., 2012) — of which the last four are already, at least partially, integrated into EuReCo.¹

¹Numerous corpora, both monolingual and multilingual, are also provided by Sketch Engine (see, e.g., Kovář et al., 2016, <https://www.sketchengine.eu>), but they are not freely available to the full extent.

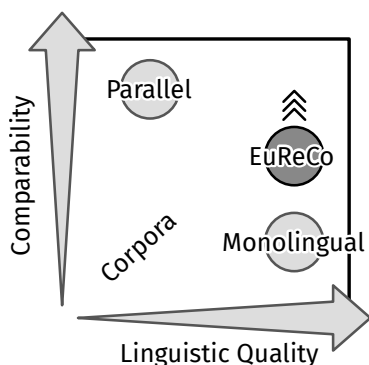


Figure 1: Comparability and Linguistic Quality

The high linguistic quality of monolingual corpora is one of the main reasons why they are used not only for single-language studies but also for cross-language research, both as a source of evidence and for advanced quantitative analyses (see, for example, the numerous contributions in [Trawiński et al., 2023](#)). However, while the high linguistic quality of monolingual corpora is a major advantage, their use as a basis of data for cross-linguistic research has obvious shortcomings, leading to the question of whether and to what extent the results of studies performed on different languages are comparable with one another. This is due to the large differences between the individual monolingual corpora in terms of size, composition and annotation (see, e.g., [Kupietz et al., 2020b](#); [Trawiński and Kupietz, 2021](#)).

Since the low comparability of monolingual corpora (despite their high linguistic quality as illustrated in Fig. 1) poses a serious empirical and methodological problem for language comparison, multilingual corpora, and especially parallel corpora, which are discussed in the following section, are predominantly used in cross-linguistic research.

2.2. Parallel Corpora

Parallel corpora consist of original texts in one language (source language) and their translations in other languages (target languages), which is why they are sometimes called translation corpora (e.g. in translation studies). The parallel texts are usually aligned at sentence level and are sometimes linguistically annotated. There are now a number of electronic parallel corpora that are freely accessible and can be searched using various web-based research and analysis systems. Among the largest and most popular are currently The Open Parallel Corpus OPUS ([Tiedemann and Nygaard, 2004](#); [Tiedemann, 2012](#)), the multilingual parallel corpus InterCorp ([Čermák and Rosen, 2012](#); [Rosen et al., 2019](#)), and The European Parliament Proceedings Parallel Corpus Europarl ([Koehn, 2005](#)). In addition, there exist numerous smaller parallel corpora,

which are either bilingual or consist of only a few languages, but often contain more detailed and accurate linguistic information due to (partly) manual annotation. Examples are the Stockholm MULTilingual TReebank SMULTRON ([Volk et al., 2015](#)) or the the CroCo corpus ([Hansen-Schirra et al., 2006](#)).

Parallel data, as provided by parallel corpora, represent linguistic units (words, phrases, sentences) in two or more languages that are translation equivalents of each other (based on functional equivalence) and as such convey the same (or similar) meanings. It is also important that these linguistic units can be viewed in-context in the respective source and target languages and within the same text types in relation to exactly the same topics, time periods, etc. Because of these properties, parallel data provide a perfect basis for determining functional equivalence between linguistic structures in a cross-linguistic context. In other words, they can be used as a perfect *tertium comparationis* (see also [James, 1980](#); [Chesterman, 1998](#)). In addition, parallel data provide insights into cross-linguistic similarities and divergences that can easily be overlooked when working with monolingual corpora. These properties of parallel data have been recognized early in cross-linguistic research and have been utilized in numerous studies in the fields of contrastive linguistics (see, e.g., [Altenberg and Granger, 2002](#); [Granger, 2010](#); [Trawiński et al., 2023](#)), language typology (see [Cysouw and Wälchli, 2007](#), and other articles in the containing volume) and translation studies (see, e.g., [Granger et al., 2003](#); [Granger and Lefer, 2022](#)).

However, despite the high degree of comparability in terms of content and size, parallel corpora provide a relatively small and undifferentiated database. In general, the more languages are used for comparison, the more the number and differentiation of parallel texts decreases. In addition, there is often a strong disproportion of original texts as opposed to translated texts (cf. the discussion in [Kupietz et al., 2020b](#); [Trawiński and Kupietz, 2021](#)).

Due to their special properties, translation texts are considered as a *third code*, i.e. a special type of text that differs from both the source language and the target language (cf. [Frawley, 1992](#); [Baker, 1993](#)). [Baker \(1995\)](#) observes that translations tend to use simpler language (*simplification*), to clarify things (*explication*), and to overuse typical patterns of the target language (*normalization*). [Laviosa \(1998\)](#) further identifies the following properties of translated texts: relatively low proportion of lexical words compared to functional words, relatively high proportion of high-frequency words compared to low-frequency words, frequent repetition of frequent words, and low variety of frequent words. In addition to *normalization*, [Teich \(2003\)](#) defines and

investigates the phenomenon of *shining-through* empirically on the basis of German-English and English-German corpora, using various grammatical constructions (such as passive and relative clauses) as examples. *Shining-through* occurs when translations are closer to the source language than to the target language. *Normalization* in terms of Teich (2003) occurs when translations are more closely oriented to the target language than would be expected.

To conclude, parallel corpora are highly comparable in terms of size and content, which is crucial for language comparison. In contrast, the quality of the linguistic material is poor compared to monolingual corpora (see Fig. 1).

2.3. Comparable Corpora

As explained above, monolingual and parallel corpora alone are suitable for contrastive linguistic research of finer granularity only to a limited extent, since, in short, they lack either comparability or linguistic quality. One way to avoid these limitations is to combine the parallel or monolingual corpora in question and to form hypotheses based on the parallel corpora, and afterwards to test them against the monolingual corpora. The disadvantage of this approach, however, is that it is time-consuming. This disadvantage can be decisive, especially in a corpus-led, explorative approach, where it is important to derive the most promising hypotheses and test them quickly in order to ultimately gain linguistic knowledge. In order to be able to assess the comparability and generalizability of corpus findings, further manual and argumentative work is also necessary. The situation is even more difficult if the corpora are only used indirectly via a language model in distributional analyses. It would therefore be better in most cases to be able to start from comparable corpora (McEnery and Xiao, 2007) of high quality.

To our knowledge, the only available comparable corpus with a broader coverage spectrum is Aranea – the family of comparable Gigaword web corpora (Benko, 2016). Aranea contains corpora of more than 20 languages, including corpora of German from Switzerland and from Austria, with controlled sizes of 120M and 1.2G words respectively. They can be queried online using the NoSketch engine (Rychlý, 2007) or KonText (Machálek, 2020). However, their limitation is that the comparability of the composition is not controlled and cannot be easily verified, since the Aranea corpora are fed exclusively from web texts that do not systematically contain the necessary metadata.

3. The European Reference Corpus EuReCo

3.1. Basic Assumptions

The European Reference Corpus EuReCo (Kupietz et al., 2017) is an open initiative founded around 2012 by the Leibniz Institute for the German Language (IDS) and the Academies of Sciences in Poland, Romania and Hungary. EuReCo is based on two fundamental assumptions. First, the creation of a significant number of new comparable corpora in Europe is unlikely to be feasible in the foreseeable future, also for reasons concerning research funding policy. The idea of EuReCo was therefore from the outset not to create new corpora, but rather to draw exclusively on the existing national and reference corpora, this way ensuring sufficient size and high linguistic quality. The second fundamental assumption of EuReCo is that general comparability of corpora is not an achievable and therefore not a particularly sensible goal (Kupietz and Trawiński, 2022).

EuReCo follows an approach that is complementary to the International Comparable Corpus (ICC) initiative (Kirk and Čermáková, 2017; Čermáková et al., 2021; Kupietz et al., 2023), which uses small corpora of predefined composition. In contrast to the ICC, no static extracts are copied from the source corpora of EuReCo – instead, the entire relevant corpora are linked virtually by means of the appropriate research software. Four reasons motivated this decision: firstly, this seemed to be the only way to fundamentally solve future copyright and licensing problems; secondly, it ensured that EuReCo would automatically benefit from future extensions of the corpora involved; thirdly, it seemed essential to use a common research platform anyway and to distribute its further development and maintenance across as many shoulders as possible. The fourth reason is the failure to establish a universal set of criteria for general comparability of corpora.

3.2. Comparability and Representativeness

Kupietz and Trawiński (2022) point out that corpora of reasonable size and diversity cannot in general be perfectly comparable, as there will always be some criterion by which the corpora differ. Whether an uneven distribution of a variable is relevant depends on the specific question being asked. Moreover, also monolingual corpora cannot be generally representative either, since their population (=language) cannot be generally defined (Evert, 2006; Kopleinig, 2017). Thus, whether a pair of corpora is *sufficiently* comparable and representative cannot be decided a priori, but depends

on the research question and the target language domain. For these reasons, a *primordial sample* approach (Kupietz et al., 2010) was chosen for EuReCo. This approach, which has been used since the 1990s for the German Reference Corpus (Teubert, 1998), invites users to use either predefined (comparable) virtual corpora or to define suitably representative and comparable corpora for the respective research question on the basis of metadata, roughly in accordance with stratified sampling. This construction of virtual comparable corpora can typically be understood as an iterative optimization process (Cosma et al., 2016). First, subcorpora are sampled from the monolingual corpora in such a way that they have similar text / token distributions with respect to relevant metadata variables, such as subject area, text type, year of publication. Then the investigations are carried out and the virtual comparable corpus definitions (or, if necessary, the research hypotheses) are iteratively refined until it can be ruled out that the findings are only due to inadequate comparability criteria or other confounding factors or artifacts. In this way, the comparability of the corpora can be effectively optimized specifically for individual research questions, as sketched in Fig. 1 (see Kupietz, 2015, for a more comprehensive description).

3.3. Previous Work

The idea of reusing existing large corpora and making subsets of them comparable is not new and, as far as we know, was first attempted by Bekavac, Osenova, Simov, and Tadić (2004), who built a Bulgarian-Croatian comparable corpus on the basis of two newspaper subcorpora from larger reference corpora of Bulgarian and Croatian.

As part of the EuReCo initiative, two large pilot projects have been carried out so far: DRuKoLA (2016–2018) and DeutUng (2017–2021)². As part of DRuKoLA, the Contemporary Reference Corpus of the Romanian Language CoRoLa (Barbu Mititelu et al., 2018; Tufiş et al., 2019) was made searchable via KorAP (Bánski et al., 2013).³ In addition, the first German-Romanian comparable corpora were defined in the project. For these, only the topic domain variable was controlled, and a random sample was drawn from DeReKo so that it contains the same token and text quantities as CoRoLa for each topic domain (see Kupietz et al., 2020b, for details). A corresponding virtual subcorpus of DeReKo also has a very similar token distribution with regard to the year of publication (Trawiński and Kupietz, 2021, p. 223) and can be

²Both funded by the Alexander von Humboldt Foundation as Institute Partnerships

³See <https://korap.racai.ro/>

publicly queried via KorAP.⁴ Several smaller pilot studies have also already been conducted on the basis of the German-Romanian comparable corpora (Kupietz et al., 2020b).

As part of the DeutUng project, the Hungarian National Corpus HNC with over one billion words was made searchable via KorAP.⁵ Individual small contrastive studies were also carried out.

4. Access to Federated Corpora for Cross-Linguistic Research

The use of already existing, large national or reference corpora for cross-linguistic studies means that, on the one hand, the rights to the data are held by separate institutions and therefore data cannot be provided centrally by a single instance (especially for legal reasons; see Fig. 2a). On the other hand, the use of different corpus analysis platforms provided by these institutions (with different feature sets, different frontends, and different API methods for accessing the separate corpus data) means reduced methodological comparability and increased demands placed on the user's skills when it comes to operating multiple systems (see Fig. 2b). A technical solution to access these corpora for contrastive research must therefore offer both geographical distribution of the data, and parallel searchability and analyzability using comparable methods.

4.1. Specification-Based vs. Implementation-Based Interoperability

In recent years, the CLARIN Federated Content Search⁶ (FCS; Trippel, 2013) has proven to be the most important technical initiative for decentralized cross-linguistic research. The FCS specifies protocols and formats that corpus providers have to implement in order to make their data accessible for comparison (see Fig. 2c). This form of specification-based interoperability (comparable to other Internet specifications such as HTML or email) has some advantages in a heterogeneous corpus landscape. The most prominent advantage is certainly the autonomy of the data providers, who can decide to what degree they want to be interoperable and who can provide not only existing corpora but also ex-

⁴The following link leads to a modifiable search within a predefined virtual DeReKo subcorpus, which is comparable to CoRoLa in terms of topic domain composition: https://korap.ids-mannheim.de/?q=%3Cbase/s=t%3E&cq=referTo%20drukola.20180909.1b_words

⁵See <https://korap.nlp.nytud.hu/>

⁶<https://contentsearch.clarin.eu/>

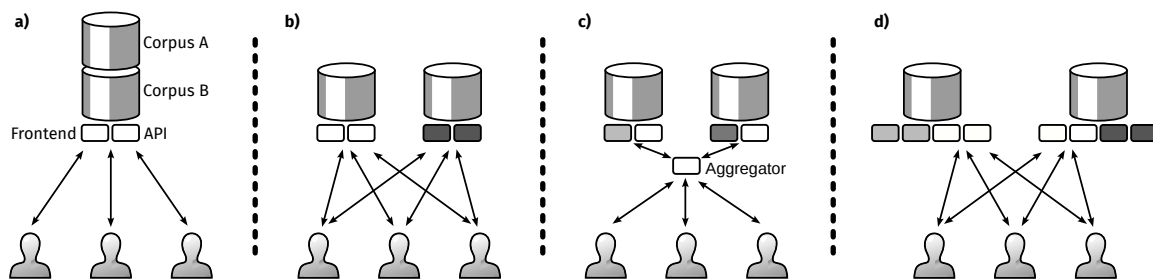


Figure 2: Querying comparable corpora: a) provided by a central instance; b) provided by different instances and interfaces; c) provided by different instances but comparable interfaces; d) provided by different instances but identical interfaces

isting corpus analysis platforms, optimized for their data and their users.

However, specification-based interoperability also has some disadvantages that can be a hindrance to the primary application scenario of EuReCo, namely to allow detailed language comparison studies:

- The scope of features provided is limited to the intersection of the feature sets provided by all participating systems (and is therefore often pretty basic);
- Innovations in the specification to extend or adapt the scope of features require new implementations and maintenance work at multiple locations and can only be used once this work has been carried out on all participating systems.

For this reason, an implementation-based approach to interoperability was chosen for EuReCo (comparable to, e.g., Shibboleth⁷), in which corpus providers deploy a special platform that is developed openly and can be used in parallel with existing corpus analysis systems (see Fig. 2d) with as little maintenance and cost as possible.

4.2. KorAP as a Tool for Implementation-Based Interoperability

While it has yet to be decided which software solution (or solutions) are going to be used for EuReCo in the future, the corpus search and analysis platform KorAP has been applied for the previous pilot projects. KorAP (Bański et al., 2012; Diewald et al., 2016) has initially been developed primarily as an access point to DeReKo, but is suitable for most corpora⁸ with arbitrary metadata and arbitrary annotations due to its agnostic approach regarding data and research questions. KorAP is also under active development as part of a standing project at

⁷<https://www.shibboleth.net/>

⁸Restrictions may concern, e.g., word segmentations.

IDS Mannheim, and is adaptable to various usage scenarios (e.g. with localization, plugins, an extensible set of query languages, and due to its open development⁹). Corpora that conform to the TEI-P5 guidelines (TEI Consortium, 2024) can be converted to the required target format and enriched with annotations in the CoNLL-U format¹⁰ (which is supported by numerous existing annotation tools) using an open source conversion pipeline.¹¹

The supported definition of virtual corpora based on metadata (Kupietz et al., 2010) makes it possible to create sub-corpora for search and analysis that can be referenced beyond instances according to certain criteria and are thus comparable in a decentralized scenario (as *virtual collections*; cf. Broeder et al., 2008).¹² KorAP also supports a complex and finely granular rights management system, which gives corpus providers exclusive control over the data to be made available, even in the case of decentralized access.

5. Recent Developments

5.1. Addition of Further Languages

This section reports on the steps taken towards two planned extensions to the coverage of EuReCo.

5.1.1. National Corpus of Polish

A pilot conversion project from a 1M sample of the National Corpus of Polish (NKJP; [Przepiórkowski](#)

⁹<https://github.com/KorAP/>; provided under a BSD-2-clause License

¹⁰<https://universaldependencies.org/format.html>

¹¹See, e.g., <https://github.com/KorAP/KorAP-XML-TEI> and <https://github.com/KorAP/KorAP-XML-CoNLL-U>

¹²This sampling procedure, as described in Sec. 3.2, can already be implemented using KorAP. However, the API interface or the R (Kupietz et al., 2020a) or Python libraries (Kupietz et al., 2022) are still required for down-sampling parts of the defined virtual corpora to their intended sizes.

et al., 2012) to the native format of KorAP was successfully concluded in the autumn of 2023. The project targeted a dataset published in May of that year as part of the Morfeusz test data suite¹³, referred to as NKJP-SGJP, where the latter part of the name stands for “grammatical dictionary of the Polish language”. This dataset is based on the original NKJP1M v. 1.2, published under CC-BY, and includes a format modification in the morphological layer that makes it more suitable for mass conversion. The additional advantage is that this version receives, on a nearly monthly basis, manual improvements of the POS and morphosyntactic annotation, according to the tagset defined for the Morfeusz tagger (Woliński, 2014), which is currently a *de-facto* standard tagger for numerous projects developing Polish language resources.

The converted tagset, apart from a layer of morphosyntactic annotation and NER information, includes also information on all possible morphological parses of its segments, before the phase of morphosyntactic disambiguation. This makes it possible to test a potential extension to the Poliqarp+ parser used in KorAP in order to handle the ~-operators (Janus and Przepiórkowski, 2007).

5.1.2. Bulgarian National Reference Corpus

Spassova (2023) has adapted the Bulgarian National Reference Corpus (BNRC; Simov et al., 2004) for use with KorAP and carried out a pilot comparative study. However, the metadata for the BNRC has not yet been mapped, and it is not yet publicly available for querying.

5.2. EuReCo as a CLARIN Project

At the EuReCo Kick-Off Workshop held on 18 October as part of the CLARIN Annual Conference 2023, the ideas underlying EuReCo were discussed with 26 invited representatives of different countries, regions and languages.

The main topics of discussion were the clarification and viability of the EuReCo solution for IPR and licensing issues, the challenge of metadata mapping, and the implementation-based approach to solving the interoperability problem with its additional costs of data conversion and of setting up and maintaining an additional corpus analysis tool. Following the discussion, which also touched on the issue of desirability versus feasibility, the final, unanimous decision was to propose a new joint CLARIN project to implement EuReCo.

¹³<http://morfeusz.sgjp.pl/download/>

5.3. Harmonization of Text Classification Metadata

The biggest challenge for the EuReCo approach is that the existing text type and domain classification systems differ among the national and reference corpora, so that these must either be mapped to a common taxonomy or to each other.

To address the issue of common domain classification, we are currently experimenting with fine-tuning multilingual Large Language Models using the English Wikipedia top-level domain as well as the standard library domain classification systems established in the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC).

5.4. Ongoing Work on Light Verb Constructions

Ongoing contrastive linguistic applications of EuReCo focus on comparisons of syntagmatic patterns in German with Romanian, Hungarian and Polish, and their variation depending on the context. Inspired by an approach by Taborek (2020), collocation analyses have been carried out in order to explore light verb constructions and their variation depending on text-external variables (text type, topic domain). These studies also serve to evaluate the properties of the respective comparable corpus definitions, KorAP’s support for contrastive analyses¹⁴, and the viability of the EuReCo approach, in general.

So far, the individual results of these studies were not particularly surprising. However, the overall results were surprising in that they supported our assumptions to a greater extent than we had anticipated.

The studies show, for example, that the results of collocation analyses vary greatly with the composition of the corpus and are particularly dependent on the proportion of certain topic domains (see Kupietz and Trawiński, 2022, p. 429ff). The type and strength of the effects differ depending on the language and on the light verb constructions analyzed. The richness of the results and the strong dependence on the composition of the comparable corpora show that even simple lexicological-syntagmatic analyses benefit greatly from an approach that allows for the dynamic definition of (comparable) corpora. Furthermore, the pilot studies, including those using the ICC, have also shown that corpora (samples) with a size of 1 million words or less are not sufficient for the study of even relatively frequent light verb constructions (Bański

¹⁴Contrastive collocation analyses are not yet possible via the KorAP web interface. Instead, we used KorAP’s R library. This also facilitated replication when analyzing the effects of different corpus compositions.

et al., 2023; Kupietz et al., 2023), so that the size of national and reference corpora, with several 100 million words, seems to be a good minimum for conducting finer-grained cross-linguistic research.

6. Conclusions and Outlook

The provision of comparable corpora for cross-linguistic research is associated with scientific, technical, legal and sometimes political challenges. With an implementation-based model for federated access to these corpora, we are pursuing an approach that is as cost-effective and low-maintenance as possible while still ensuring a high level of variability and methodological rigor.

In the next steps, further national and reference corpora are going to be integrated into EuReCo. Meanwhile, different approaches to mapping metadata (in particular topic domain and text type) to common classification systems are going to be evaluated.

7. References

- Bengt Altenberg and Sylviane Granger, editors. 2002. *Lexis in Contrast: Corpus-based approaches*, volume 7 of *Studies in Corpus Linguistics*. Benjamins, Amsterdam.
- Guy Aston and Lou Burnard. 1998. *The BNC Handbook*. Edinburgh University Press.
- Mona Baker. 1993. Corpus linguistics and translation studies – Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In honour of John Sinclair*, pages 233–250. Benjamins, Amsterdam.
- Mona Baker. 1995. *Corpora in Translation Studies: An overview and some suggestions for future research*. *Target*, 7(2):223–243.
- Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. *The Reference Corpus of the Contemporary Romanian Language (CoRoLa)*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1178–1185, Miyazaki, Japan. European Language Resources Association (ELRA).
- Piotr Bański, Nils Diewald, Marc Kupietz, and Beata Trawiński. 2023. *Applying the newly extended European reference corpus EuReCo. Pilot studies of light-verb constructions in German, Romanian, Hungarian and Polish*. In *Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10)*, 18-21 July, 2023, Mannheim, Germany, pages 274–276, Mannheim. IDS-Verlag.
- Piotr Bański, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. *The New IDS Corpus Analysis Platform: Challenges and Prospects*. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2905–2911, Istanbul, Turkey. European Language Resources Association (ELRA).
- Božo Bekavac, Petya Osenova, Kiril Simov, and Marko Tadić. 2004. *Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian*. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Vladimír Benko. 2016. *Two Years of Aranea: Increasing Counts and Tuning the Pipeline*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4245–4248, Portorož, Slovenia. European Language Resources Association (ELRA).
- Vaclav Brezina, Robbie Love, and Karin Aijmer, editors. 2018. *Corpus Approaches to Contemporary British Speech: Sociolinguistic studies of the Spoken BNC2014*. Routledge, New York.
- Daan Broeder, Thierry Declerck, Erhard Hinrichs, Stelios Piperidis, Laurent Romary, Nicoletta Calzolari, and Peter Wittenburg. 2008. *Foundation of a Component-based Flexible Registry for Language Resources and Technology*. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1433–1436, Marrakech, Morocco.
- Piotr Bński, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr P zik, Carsten Schnober, and Andreas Witt. 2013. *KorAP: the new corpus analysis platform at IDS Mannheim*. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human language technology challenges for computer science and linguistics. 6th language & technology conference*, pages 586–587. Uniwersytet im. Adama Mickiewicza Poznanu, Poznań.
- Andrew Chesterman. 1998. *Contrastive Functional Analysis*. Number 47 in *Pragmatics & Beyond*. Benjamins, Amsterdam.
- Ruxandra Cosma, Dan Cristea, Marc Kupietz, Dan Tufiş, and Andreas Witt. 2016. *DRuKoLA – towards contrastive German-Romanian research*

- based on comparable corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pages 28–32, Portorož, Slovenia. European Language Resources Association (ELRA).
- Michael Cysouw and Bernhard Wälchli. 2007. [Parallel texts: using translational equivalents in linguistic typology](#). *Language Typology and Universals*, 60(2):95–99.
- Mark Davies. 2011. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, 25:447–465.
- Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. [KorAP Architecture - Diving in the Deep Sea of Corpus Data](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3586–3591, Portorož, Slovenia. European Language Resources Association (ELRA).
- Stefan Evert. 2006. [How Random is a Corpus? The Library Metaphor](#). *Zeitschrift für Anglistik und Amerikanistik*, 54(2).
- William Frawley. 1992. *Linguistic semantics*. Lawrence Erlbaum Associates, Hillsdale.
- Sylviane Granger and Marie-Aude Lefer, editors. 2022. *Extending the Scope of Corpus-Based Translation Studies*. Bloomsbury Advances in Translation. Bloomsbury, London, UK.
- Sylviane Granger. 2010. Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. *Contemporary Foreign Language Studies*, 10(2):14–21.
- Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson. 2003. *Corpus-based approaches to contrastive linguistics and translation studies*, volume 20. Rodopi, Amsterdam & Atlanta.
- Silvia Hansen-Schirra, Stella Neumann, and Michaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the 5th workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pages 35–42, Stroudsburg. ACL.
- Carl James. 1980. *Contrastive Analysis*. Longman, London.
- Daniel Janus and Adam Przepiórkowski. 2007. [Poliqarp: An open source corpus indexer and search engine with syntactic extensions](#). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 85–88, Prague, Czech Republic. Association for Computational Linguistics.
- John Kirk and Anna Čermáková. 2017. From ICE to ICC: The new International Comparable Corpus. In Piotr Bański, Marc Kupietz, Harald Lungen, Paul Rayson, Hanno Biber, Evelyn Breiteneder, Simon Clematide, John Mariani, Mark Stevenson, and Theresa Sick, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section*, pages 7 – 12. IDS, Mannheim.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.
- Alexander Koplein. 2017. [Against statistical significance testing in corpus linguistics](#). *Corpus Linguistics and Linguistic Theory*, 15(2):321–346.
- Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. 2016. [Sketch Engine for Bilingual Lexicography](#). *International Journal of Lexicography*, 29:ecw029.
- Marc Kupietz. 2015. [Constructing a Corpus](#). In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, pages 62–75. Oxford University Press.
- Marc Kupietz, Adrien Barbaresi, Anna Čermáková, Małgorzata Czachor, Nils Diewald, Jarle Ebeling, Rafał L. Górski, Eliza Margaretha, John Kirk, Michal Křen, Harald Lungen, Signe Oksefjell Ebeling, Mícheál Ó Meachair, Ines Pisetta, Elaine Uí Dhonnchadha, Friedemann Vogel, Rebecca Wilm, Jiajin Xu, and Rameela Yadhige. 2023. [News from the International Comparable Corpus. First launch of ICC written](#). In *Book of Abstracts of the 10th International Contrastive Linguistics Conference (ICLC-10)*, pages 45–48, Mannheim, Germany. IDS-Verlag; Leibniz-Institut für Deutsche Sprache (IDS).
- Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. [The German Reference Corpus DeReKo: A primordial sample for linguistic research](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).
- Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2020a. [RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo](#)

- via KorAP. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, pages 7015–7021, Marseille, France. European Language Resources Association.
- Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2022. [Building paths to corpus data: A multi-level least effort and maximum return approach](#). In Darja Fišer and Andreas Witt, editors, *CLARIN. The Infrastructure for Language Resources.*, pages 163–189. deGruyter, Berlin. Section: number x.
- Marc Kupietz, Nils Diewald, Beata Trawiński, Ruxandra Cosma, Dan Cristea, Dan Tufiş, Tamás Váradi, and Angelika Wöllstein. 2020b. Recent developments in the European Reference Corpus EuReCo. *Translating and Comparing Languages: Corpus-based Insights. Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference. Louvain-la-Neuve: Presses universitaires de Louvain*, pages 257–273.
- Marc Kupietz, Harald Lungen, Paweł Kamocki, and Andreas Witt. 2018. [The German reference corpus DeReKo: New developments – new opportunities](#). In *Proceedings of the Eleventh International Conference on language resources and evaluation (LREC '18)*, pages 4353–4360, Miyazaki, Japan. ELRA.
- Marc Kupietz and Beata Trawiński. 2022. [Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo](#). In Laura Auteri, Natascia Barrale, Arianna Di Bella, and Sabine Hoffmann, editors, *Wege der Germanistik in transkultureller Perspektive. Akten des XIV. Kongresses der Internationalen Vereinigung für Germanistik (IVG) (Bd. 6)*, Jahrbuch für Internationale Germanistik - Beihefte - 6, pages 417–439. Peter Lang, Bern.
- Marc Kupietz, Andreas Witt, Piotr Bański, Dan Tufiş, Dan Cristea, and Tamás Váradi. 2017. [EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research](#). In *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017*, pages 15–19, Mannheim. Institut für Deutsche Sprache.
- Michal Křen. 2020. [Czech National Corpus in 2020: Recent Developments and Future Outlook](#). In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 52–57, Marseille, France. European Language Resources Association.
- Sara Laviosa. 1998. Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, 43(4):557–570.
- Tomáš Machálek. 2020. [KonText: Advanced and Flexible Corpus Query Interface](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.
- Anthony M. McEnery and Richard Zhonghua Xiao. 2007. [Parallel and comparable corpora: What is Happening?](#) In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: The Linguist and the Translator*. Multilingual Matters, Clevedon, UK.
- Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. [The Hungarian Gigaword Corpus](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, pages 1719–1723, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- Alexandr Rosen, Martin Vavřín, and Adrian J. Zaslava. 2019. *The InterCorp Corpus – Czech1, 12. Version*. Institute of the Czech National Corpus/Charles University, Prague.
- Pavel Rychlý. 2007. Manatee / Bonito - A modular corpus manager. In Petr Sojka and Aleš Horák, editors, *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70. Masaryk University, Brno.
- Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. [A Language Resources Infrastructure for Bulgarian](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Lora Spassova. 2023. *Integrating a Large Bulgarian Corpus into the European Reference Corpus EuReCo*. Bachelor Thesis, Heinrich-Heine University Düsseldorf.
- Janusz Taborek. 2020. [Kookkurrenz und syntagmatische Muster der Funktionsverbgefüge aus kontrastiver deutsch-polnischer Sicht am Beispiel in Not geraten](#). In Sabine Knop and Manon Hermann, editors, *Funktionsverbgefüge im Fokus:*

- Theoretische, didaktische und kontrastive Perspektiven*, pages 211–234. De Gruyter, Berlin.
- TEI Consortium. 2024. [TEI P5: Guidelines for Electronic Text Encoding and Interchange](#).
- Elke Teich. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.
- Wolfgang Teubert. 1998. [Korpus und Neologie](#). In Wolfgang Teubert, editor, *Neologie und Korpus*, number 11 in *Studien zur deutschen Sprache*, pages 129–170. Narr, Tübingen.
- Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. ELRA.
- Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus – parallel & free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, pages 1183–1186, Lisbon, Portugal. ELRA.
- Beata Trawiński and Marc Kupietz. 2021. [Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo](#). In Henning Lobin, Andreas Witt, and Angelika Wöllstein, editors, *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*, number 2020 in *Jahrbuch / Leibniz-Institut für Deutsche Sprache (IDS)*, pages 209–234. de Gruyter, Berlin, Germany.
- Beata Trawiński, Marc Kupietz, Kristel Proost, and Jörg Zinken, editors. 2023. *10th International Contrastive Linguistics Conference (ICLC). Book of abstracts*. IDS, Mannheim, Germany.
- Thorsten Trippel. 2013. [Minutes to the Workshop on Federated Content Search](#). Technical report, University of Copenhagen, Copenhagen.
- Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. 2019. Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. *On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*, 64(3). Place: Bucharest, Romania Publisher: Editura Academiei Române.
- Martin Volk, Anne Göhring, Annette Rios, Torsten Marek, and Yvonne Samuelsson. 2015. *SMULTRON (4. Version) — The Stockholm MULTilingual parallel TReebank*. Institute of Computational Linguistics, University of Zurich, Zurich.
- Tamás Váradi. 2002. [The Hungarian National Corpus](#). In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 385–389, Las Palmas, Spain. European Language Resources Association (ELRA).
- Marcin Woliński. 2014. [Morfeusz Reloaded](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1106–1111, Reykjavik, Iceland. European Language Resources Association (ELRA).
- František Čermák and Alexandr Rosen. 2012. [The case of InterCorp, a multilingual parallel corpus](#). *International Journal of Corpus Linguistics*, 17(3):411–427.
- Anna Čermáková, Jarmo Jantunen, Tommi Jauhiainen, John Kirk, Michal Křen, Marc Kupietz, and Elaine Uí Dhonnchadha. 2021. [The International Comparable Corpus: Challenges in building multilingual spoken and written comparable corpora](#). *Research in Corpus Linguistics: Special issue "Challenges of combining structured and unstructured data in corpus development"*, 9(1):89 – 103.

Building Annotated Parallel Corpora Using the ATIS Dataset: Two UD-style treebanks in English and Turkish

Neslihan Cesur^{1,2}, Aslı Kuzgun¹, Mehmet Köse³, Olcay Taner Yıldız²

Starlang Yazılım Danışmanlık¹, Işık University³, Özyeğin University²,
Istanbul, Turkey

{neslihan, asli}@starlangyazilim.com, mehmetkse@gmail.com, olcay.yildiz@ozyegin.edu.tr

Abstract

In this paper, we introduce the annotation process of the Air Travel Information Systems (ATIS) Dataset as a parallel treebank in English and in Turkish. The ATIS Dataset was originally compiled as pilot data to measure the efficiency of Spoken Language Systems and it comprises human speech transcriptions of people asking for flight information on the automated inquiry systems. Our first annotated treebank, which is in English, includes 61.879 tokens (5.432 sentences) while the second treebank, which was translated into Turkish, contains 45.875 tokens for the same amount of sentences. First, both treebanks were morphologically annotated through a semi-automatic process. Later, the dependency annotations were performed by a team of linguists according to the Universal Dependencies (UD) guidelines. These two parallel annotated treebanks provide a valuable contribution to language resources thanks to the spontaneous/spoken nature of the data and the availability of cross-linguistic dependency annotation.

Keywords: Universal Dependencies, ATIS, Annotated Corpus, Parallel Corpora

1. Introduction

Large natural language corpora, whether it includes spoken or written data, are a crucial asset to natural language processing (NLP) research when it comes to building intelligent systems which can understand, manipulate and produce human language. Manually and systematically parsed, gold-standard treebanks provide important resources especially for the training and evaluation of parsers.

As most of the available corpora are monolingual, parallel corpora which contain the same content in two or more languages constitute valuable linguistic resources for supervised machine learning applications. Thanks to parallel corpora, we can build state-of-the-art multilingual parsers and evaluate parser quality using multiple languages. Parallel corpora are also beneficial for building tools such as machine translation systems and multilingual question answering systems. The ATIS parallel treebank will be part of four datasets which we hope to use in training a bilingual parser. Two of these treebanks are already available online: the PUD treebank and the Penn Treebank in English and Turkish (Kuzgun et al., 2020). The third dataset, a parallel QuestionBank is currently being annotated by our team.

The parallel ATIS treebank¹ is built as a dependency treebank in English and Turkish, in accordance with the Universal Dependencies (UD) guidelines. The treebank is comprised of annotated data from the Air Travel Information System

(ATIS) Dataset (Hemphill et al., 1990). This dataset was originally collected as a pilot corpus to evaluate the progress in Spoken Language Systems. It comprises transcripts of spoken data in which customers are inquiring about flight information. As a strictly domain-specific corpus, the data mostly contains names of cities, airports, airlines and flight numbers. As the vast majority of natural language corpora are made up of samples of written language, the main advantage of the ATIS Dataset is that it contains samples of spontaneous speech. As the data is not pre-written, the corpus contains incomplete sentences and errors in speech, which differentiates it from most corpora comprising written natural language data.

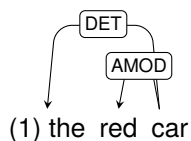
Both of our treebanks include 5,432 sentences (the treebank in English has 61,879 tokens while the treebank in Turkish has 45,875 tokens due to the agglutinative nature of Turkish). In this paper, we outline the steps of our annotation process and present quantitative results through a comparison of our treebanks. The annotation of the English ATIS treebank is made up of two main stages: automatic POS-tagging (later controlled by human annotators) and dependency annotation. The Turkish treebank, however, required five stages: the semi-automatic translation of the dataset, automatic morphological analysis, morphological disambiguation, automatic POS-tagging, and dependency annotation. The morphological and syntactic annotations were carried out by three annotators with a background in linguistic studies while the translation team included three translators with a background in linguistics and translation studies. The anno-

¹Both versions of the treebank can be accessed on the website of Universal Dependencies Project.

tation decisions specific to the data were made prior to the annotation process through open discussions.

2. The Universal Dependencies Project

In terms of the syntactic framework they follow, treebanks can either be annotated using phrase structure grammar or dependency grammar. Phrase structure grammar, whose foundations were laid by Noam Chomsky in the 1950s (Chomsky, 1957), consists of branching trees which group together constituents under labeled nodes. Dependency grammar, which was first popularized by the French linguist Lucien Tesnière in the 1950s² (Tesnière, 1959), is a framework which aims to mark one-to-one syntactic relations between the constituents of a given phrase or sentence. In DG, each element in a sentence is considered a node and is linked to another element through head-dependent relations. The example in (1) demonstrates the head-dependent relationship in a noun phrase. The element *car* is the head of the phrase. The determiner (DET) and the adjectival modifier (AMOD) are linked to the head *car* as dependents.



We have used the tags and rules of the Universal Dependencies (UD) project for our annotation. UD (Nivre et al., 2016) is a project which aims to provide a consistent and cross-linguistic annotation scheme for parts of speech (POS) tagging and dependency syntax. With more than 100 languages currently available for open-access, the Universal Dependencies project provides great resources for cross-lingual learning and multilingual parser development.

The first annotated treebank in Turkish is the METU-Sabancı Treebank (Ofazer et al., 2003; Atalay et al., 2003; Sulubacak et al., 2016). The Turkish Penn Treebank (Kuzgun et al., 2020) corpus is the largest Turkish dependency treebank currently available with 183,555 tokens. Moreover, The Penn Treebank corpus and the PUD treebank are the only multi-lingual treebanks which include Turkish. Amongst these annotated corpora, ATIS constitutes a crucial contribution in that it not only introduces the first treebank which is comprised of spoken natural language data but it is also another parallel treebank.

²Tesnière's work on syntax and dependency grammar was published posthumously.

3. Annotation Process

3.1. Translation

The ATIS dataset was loaded from an open-source repository, currently available on GitHub³. Before the annotation process, the ATIS Dataset was translated into Turkish by a team of seven translators. The translation was carried out on Google Sheets, which allowed the team to work simultaneously on an online platform. English sentences were listed in one column, with their corresponding Turkish translations added to the adjacent row. Figure 1 illustrates some English sentences with their Turkish counterparts. The translators adopted a semi-automatic translation strategy by translating the sentences with the help of different machine translation tools. Then, the outputs were checked and corrected by the human translators to ensure that the correspondence between the two languages was accurate. This process was important to keep the originality of the English data, including the absence of punctuation marks and the use of discourse particles such as *now* and *okay* at the beginning of sentences. As Figure 1 illustrates, the original sentences do not include question marks or periods at the end of sentences contrary to most written natural language corpora.

3.2. Morphological Analysis

Both morphological and syntactic annotations were carried out with the same interface called StarDust, introduced in (Yenice et al., 2022). StarDust is packaged as a JAR (Java ARchive) file and is implemented using the Java programming language. We opted for this interface because it provides a user-friendly interface for annotators and it can run different annotation programs such as POS-tagger, morphological analyzer and dependency annotator.

The English dataset only required POS-tag annotation whereas Turkish was a lot more complicated to analyze due to its agglutinating morphological structure. The morphological annotation of the English data consisted of POS-tagging the tokens, using the Penn Part of Speech Tags⁴ (Marcus et al., 1993). Within the interface used for annotation, the POS-tag detection took place automatically. After the tags were determined, the roots of the tokens were automatically selected by the analyzer through a rule-based algorithm. The rules consisted of removing the inflections found on the token and marking the remaining part as the root. For instance, if the plural noun *flights* is marked with the tag *NNS* (used to indicate plural nouns), the plural marker *s* is omitted and the remaining part is selected as

³https://github.com/howl-anderson/ATIS_dataset

⁴<https://www.cis.upenn.edu/~bies/manuals/tagguide.pdf>.

	A	B	C	D
1	TYPE	NO	ENGLISH	TURKISH
2	train	1	what is the cost of a round trip flight from pittsburgh to atlanta beginning on april twenty fifth and returning on may sixth	Pittsburgh'tan Atlanta'ya 25 Nisan'da gidiş 6 Mayıs'ta dönüşü olan bir gidiş dönüş uçuşunun maliyeti nedir
3	train	2	now i need a flight leaving fort worth and arriving in denver no later than 2 pm next monday	Şimdi Fort Worth'dan ayrılan ve en geç gelecek pazartesi akşam 2'ye kadar Denver'e varacak bir uçağa ihtiyacım var
4	train	3	i need to fly from kansas city to chicago leaving next wednesday and returning the following day	Önümüzdeki çarşamba gidiş ertesi gün dönüşü olan Kansas City'den Chicago'ya giden bir uçuşa ihtiyacım var

Figure 1: The translation sheet with English sentences appearing in Column C and their corresponding Turkish translations in Column D.

the root of the token. For exceptional cases such as suppletive forms (like *are* and *were* having the root *be*), separate rules were implemented for the selection of the root. In Figure 2, black words indicate the tokens and blue words indicate the root of the tokens. According to our rules, the root of the token *are* is automatically determined as *be* and the root of *flights* is determined as *flight*. POS-tags are indicated in red and can be modified by the annotators by clicking on the token.

After the tags were checked and manually corrected by our annotators, the Penn POS-tags were automatically converted into UD-style tags, called Universal POS-tags⁵. This was also done by a rule-based algorithm. For instance, these rules automatically convert noun tags such as *NN*, *NNS* and *NNP* into a *NOUN* UD tag. The *PRP* (personal pronoun) tag is converted to *PRON* tag and so on. These UD POS-tags are visible to the annotators during the dependency annotation process as shown in Figure 3.

actually	what	are	the	nonstop	flights
actually	what	be	the	nonstop	flight
RB	WP	AUX:VBP	DT	JJ	NNS

Figure 2: A view of the POS-tagger, showing the tokens in black, the roots in blue and the Penn POS-tags in red.

As for the morphological analysis of the Turkish treebank, a rule-based morphological analyzer by Yıldız et al. (2019) was implemented. This open-source morphological analyzer works with a lexicon and a finite state transducer. It lists out the derivations for every possible root of a given token along with every possible morphological tag of a given suffix. After this automatic morphological analysis which separated the tokens into possible roots and affixes, a manual morphological disambiguation

⁵<https://universaldependencies.org/u/pos/>

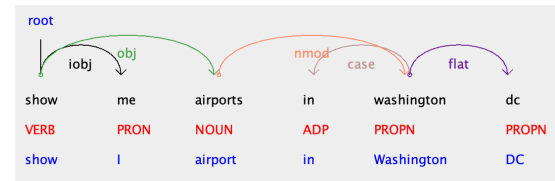


Figure 3: A view of the dependency annotator with the UD POS-tags and roots indicated below the tokens.

was carried out by our annotators in order to select the correct analysis for each token. One reason why this step is crucial for Turkish is because the same form can correspond to different morphological tags depending on the context. For instance, when a word receives the suffix *-i*, one needs to decide whether it is an accusative marker found on direct objects or a third person possessive marker. Another example is shown in Figure 4, in which the token *sabahın* receives two possible analyses due to the suffix *-(n)ın* which is added to the root *sabah* 'morning'. The suffix can either correspond to a second person possessive marker (as in *your morning*) or a genitive marker (as in *early hours of the morning*). The second option is selected according to the context in the given example.

sabahın	erken	saatlerinde
sabah+NOUN+A3SG+P2SG+NOM	erken+ADJ	saat+NOUN+A3PL+P3SG+LOC
sabah+NOUN+A3SG+PNON+GEN		

Figure 4: A view of the Turkish morphological analyzer, showing two possible analyses for the token *sabahın*.

After the manual morphological disambiguation, the tokens are automatically assigned their UD POS-tags according to their final morphological tags. As with the sentences in English, the interface makes the UD POS-tags visible to annotators

during the dependency annotation stage, as shown in Figure 5.

3.3. Dependency Annotation

After the morphological analysis/disambiguation of Turkish tokens and the assignment of Universal POS-tags of both treebanks, the dependency annotations were carried out by the same three annotators. The annotations took place on the same open-source interface (Yenice et al., 2022) which was used for the morphological analysis and POS-tagging. During this stage, the annotators determined the heads and dependents in a given sentence or phrase and labeled them with the appropriate UD dependency tags. Images 3 and 5 show how the arrows depart from the head and point to the dependent. Each relation is marked with a separate color and the corresponding tag is shown in the arch of the arrow. Moreover, Figure 6 shows a larger overview of the interface including the tag box. When the annotator drags the dependent towards the head, the tag box pops up. As the interface allows for a connection between the layers of the POS-tagger and dependency annotator, the possible UD tags which are available for a specific relation are automatically restricted to enable a faster selection. Moreover, errors in annotation violating the UD rules are automatically detected and indicated at the bottom of the screen.

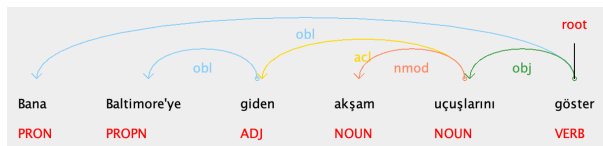


Figure 5: A view of the dependency annotator, showing how heads are connected to their dependents with arrows.

A total of 32 UD tags were used in the syntactic annotation process. Table 1 and 2 illustrate the 10 most frequently used UD tags in the Turkish and English ATIS treebanks, respectively.

We observe that the frequency for the NMOD (nominal modifier) tag is higher than the ROOT tag in both languages. This shows that most of the sentences contain more than one nominal modifier. Even though nominal modifiers are common in most treebanks, the significant number in our treebank points to the frequency of phrasal elements such as *from Burbank*, *in Washington*, etc. Also, the fact that the CASE tag (which marks adpositions) is more common than the ROOT tag in English points to the abundance of prepositions indicating location and direction such as *in*, *to* and *from*, which is also shown in Table 4. The dependency representations below show examples from

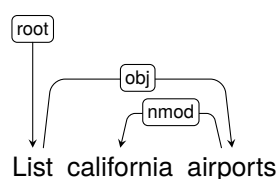
Dependency Relation	Frequency
NMOD	11.626
ROOT	5.431
OBL	4.193
FLAT	3.266
AMOD	2.928
OBJ	2.744
ACL	2.038
NSUBJ	1.824
COMPOUND	1.470
DET	1.305

Table 1: Top 10 most frequent dependency tags and their frequencies in the Turkish ATIS treebank

Dependency Relation	Frequency
CASE	13.131
NMOD	8.568
ROOT	5.432
DET	4.738
NSUBJ	3.323
OBJ	3.274
FLAT	3.148
OBL	3.130
COMPOUND	2.377
AMOD	1.787

Table 2: Top 10 most frequent dependency tags and their frequencies in the English ATIS treebank

the English ATIS treebank. The first example shows a case where the NMOD tag is used within a noun phrase. The nominal modifier *california* -which is the dependent- is linked to the head of the phrase, *airports*. The verb *list* is marked as the ROOT and the head of the noun phrase *airports* is marked as the object (OBJ) of the main verb. The second example shows a noun phrase with the head noun *flights*. The phrases *from Las Vegas* and *to Burbank* are attached to the head noun as nominal modifiers. We also see the use of the two most common words in the English dataset, *to* and *from*, attached to different noun heads with the CASE tag.



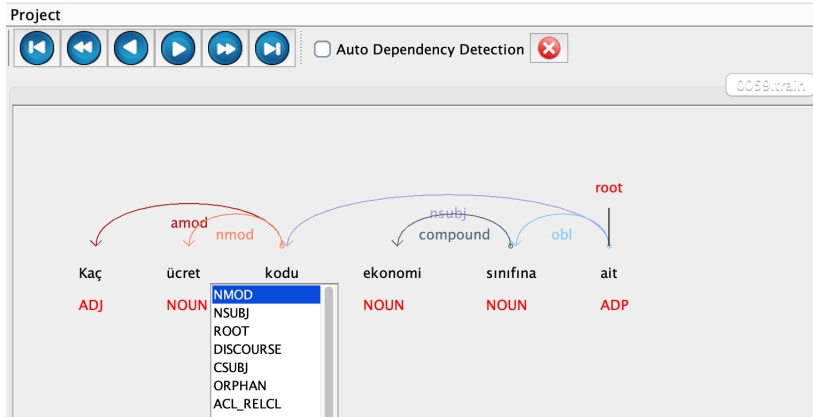
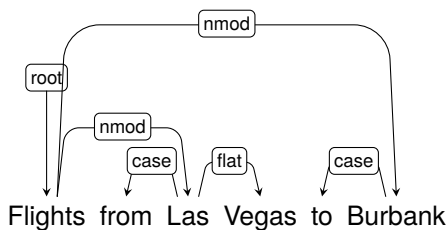


Figure 6: An overview of the dependency annotation interface, illustrating the control buttons and the tag box.



What is interesting is that we do not find the PUNCT tag in neither of the lists even though it usually appears quite frequently in general-purpose datasets as each sentence usually ends with a punctuation mark. As we have indicated in Section 3.1 and Figure 1 the original dataset lacks the usual punctuation marks. Their absence in our treebanks points to the fact that ATIS is a dataset of spoken natural language data. As the transcripts do not include punctuation, we observe a discrepancy compared to treebanks with written language data. The table below shows inter-annotator agreement scores for both treebanks. DEP shows the percentage of dependencies linked to the correct head with the correct tag. TO shows the percentage of dependencies linked to the correct head. TAG shows the percentage of dependencies which were marked with the correct tag.

	DEP	TO	TYPE
Turkish	78	84	82
English	82	91	86

Table 3: Inter-annotator agreement scores for both languages

Overall, the percentages for English are higher for each score type. The linking between heads and their dependents is more accurate than the selected tag in both languages.

4. A Quantitative Analysis of the Treebanks

As we have already stated, the Turkish ATIS Treebank comprises a total of 45,875 tokens while the English ATIS Treebank has 61,879 tokens. Considering that the number of annotated sentences are the same, the significant difference between token numbers point to the distinct morphological nature of the two languages.

Moreover, due to the same morphological pattern, we observe that the English dataset has less unique surface forms⁶ (932 unique surface forms) than the Turkish dataset (2,133 unique surface forms), despite containing more tokens. This means that 4.64% of the Turkish dataset consists of unique forms while for English, this percentage is around 1.5. One reason for this discrepancy can be found in prepositional phrases. A state/city name in English can only appear in its bare form in English (such as *Denver* or *Boston*). The directionality information is conveyed through prepositions such as *to* and *from*. However, in Turkish, the directionality is expressed using nominal case markings such as the dative form (*Denver'a*), the locative form (*Denver'da*) and the ablative form (*Denver'dan*). If we consider that these case markings are suffixed to each location name, we end up with a greater number of unique forms in Turkish. For each location name, only 1 unique word is added in English (the bare form of the location name) while in Turkish, four unique forms (considering only the nominative, locative, dative and ablative forms) are created.

Another significant comparison can be made regarding the domain-specific nature of the treebanks. Compared to a dataset including a wider range of topics, a domain-specific treebank is ex-

⁶Each occurrence of a distinct word form is counted as a *unique surface form*. For example, *flight* and *flights* are two unique surface forms in English.

pected to contain less unique surface forms. We can clearly observe this fact when we compare the Turkish ATIS treebank to the KeNet dependency treebank which is also a part of the Universal Dependencies Project. KeNet contains a total of 149,524 tokens⁷ amongst which 49,156 are unique forms. This means that while 4.64% of the Turkish ATIS treebank is comprised of unique forms, the KeNet data comprises up to 32.84% of unique forms. This significant difference indicates that as a domain-specific treebank, ATIS shows much less variety in terms of words and word forms. Another domain-specific dependency treebank in Turkish, the Tourism treebank, contains a total of 71,322 tokens and 4,961 unique surface forms which makes up the 6.96% of the dataset. This number is slightly larger than what we have found for the Turkish ATIS treebank. This shows that amongst the Turkish treebanks, the new ATIS dataset is the most specific one with the least amount of diversity in words and word forms.

The effects of domain-specificity can also be observed in the most frequent surface forms. Word frequency lists of more generic datasets usually pattern with the most frequently used words of the given language. These usually include determiners, prepositions, auxiliaries and conjunctions. However, due to their restricted content, domain-specific treebanks include content words relating to the topic of the dataset. Table 4 is a list of the most common 15 words in the ATIS treebanks. We observe different forms of the word *flight* (*uçuşlar* which means *flights* and its accusative form *uçuşları*) in both treebanks. We also find several state/city names (*Boston*, *Denver*, *San Francisco*) and question words. Such specific content words and proper names would not appear as frequently in a dataset containing more generic content. The rest of the words include pronouns (*ben*, *bana*, *I*, *me*), determiners (*the*, *bir*) and prepositions expressing directionality (*to* and *from*).

5. Conclusion

This paper was an overview of the morphological and syntactic annotation process of a parallel treebank in English and in Turkish.

Our two annotated treebanks constitute a valuable contribution to the Universal Dependencies project as they are the only annotated dependency treebanks which include solely spoken language data. They also show certain distinct characteristics regarding their domain-specific nature, including a decreased variety in unique forms and a more

⁷The number of tokens indicated here does not include punctuation marks considering that the KeNet dataset includes a great number of punctuation while ATIS does not make use of a significant amount.

	Turkish ATIS	English ATIS
1	uçuşları	to
2	San	from
3	olan	flights
4	göster	the
5	uçuş	on
6	bir	what
7	istiyorum	flight
8	uçuşlar	me
9	var	I
10	ve	San
11	bana	Boston
12	Boston'dan	show
13	hangi	a
14	en	Denver
15	Francisco'ya	in

Table 4: Top 10 most frequent surface forms in both ATIS Treebanks

specific set of most frequent words compared to generic datasets.

Another valuable aspect of our treebanks is that they are bilingual. As we have seen above, this type of treebanks allow for a typological comparison between languages. We have discussed the gap between the number of tokens and the percentage of unique words in order to show that such treebanks offer quantitative measures which point to morphological distinctions between languages. In addition to typological analysis, parallel treebanks can be used for the training of multilingual parsers. In this regard, the ATIS treebanks would be especially useful in training parsers for the analysis of spoken natural language and interpreting simple commands.

6. Bibliographical References

- Nart B Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the turkish treebank. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*.
- Noam Chomsky. 1957. *Syntactic structures*. Mouton § Co.
- Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
- Aslı Kuzgun, Neslihan Cesur, Bilge Nas Arıcan, Merve Özçelik, Büşra Marşan, Neslihan Kara, Deniz Baran Aslan, and Olcay Taner Yıldız. 2020.

- On building the largest and cross-linguistic turkish dependency corpus. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.
- Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a turkish treebank. *Treebanks: Building and using parsed corpora*, pages 261–277.
- Umut Sulubacak, Gülşen Eryiğit, and Tuğba Pamay. 2016. Imst: A revisited turkish dependency treebank. In *Proceedings of TurCLing 2016, the 1st international conference on Turkic computational linguistics*. Ege University Press.
- Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck.
- Arife B Yenice, Neslihan Cesur, Aslı Kuzgun, and Olcay Taner Yıldız. 2022. Introducing stardust: A ud-based dependency annotation tool. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 79–84.
- Olcay Taner Yıldız, Begüm Avar, and Gökhan Ercan. 2019. An open, extendible, and fast turkish morphological analyzer. In *International Conference Recent Advances in Natural Language Processing*.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. [CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Bootstrapping the Annotation of UD Learner Treebanks

Arianna Masciolini

Språkbanken Text

Department of Swedish, Multilingualism, Language Technology

University of Gothenburg

arianna.masciolini@gu.se

Abstract

Learner data comes in a variety of formats, making corpora difficult to compare with each other. Universal Dependencies (UD) has therefore been proposed as a replacement for the various *ad-hoc* annotation schemes. Nowadays, the time-consuming task of building a UD treebank often starts with a round of automatic annotation. The performance of the currently available tools trained on standard language, however, tends to decline substantially upon application to learner text. Grammatical errors play a major role, but a significant performance gap has been observed even between standard test sets and normalized learner essays. In this paper, we investigate how to best bootstrap the annotation of UD learner corpora. In particular, we want to establish whether Target Hypotheses (THs), i.e. grammar-corrected learner sentences, are suitable training data for fine-tuning a parser aimed for original (ungrammatical) L2 material. We perform experiments using English and Italian data from two of the already available UD learner corpora. Our results show manually annotated THs to be highly beneficial and suggest that even automatically parsed sentences of this kind might be helpful, if available in sufficiently large amounts.

Keywords: second language acquisition, learner corpora, dependency parsing, universal dependencies

1. Introduction

In recent years, Second Language Acquisition (SLA) has become more and more reliant on corpus studies, to the point of Learner Corpus Research becoming a well-established field, as attested by the founding of the Learner Corpus Association¹ and the institution of a dedicated journal². Learner data, however, comes in a variety of formats depending on each corpus' original purpose. This makes such datasets difficult to reuse and hardly comparable with each other. In this sense, linguistic annotation in Universal Dependencies (UD) (de Marneffe et al., 2021) is an appealing alternative to the various existing *ad-hoc* annotation schemes. UD would in fact provide a uniform annotation layer not only across datasets, but also across languages.

In particular, Lee et al. (2017) proposed *L1-L2 parallel dependency treebanks*, consisting of UD-annotated learner sentences paired with *correction* or *target hypotheses* (henceforth THs) as a replacement for explicitly error-tagged corpora.³ The key idea is that systematic cross-linguistically consistent morphosyntactical annotation is sufficient for retrieving grammatical errors via tree queries, as demonstrated in Masciolini (2023). In addition, UD-annotated data lends itself to comparative cross-language studies and other types of analyses, both

quantitative and qualitative. L1-L2 treebanks of different sizes have been released for English (Berzak et al., 2022), Chinese (Lee et al., 2023) and Italian (Di Nuovo et al., 2023), and we have the medium-term goal of releasing a fourth one based on the Swedish Learner Language (SWELL) corpus (Volodina et al., 2019).

Nonetheless, building a high-quality UD corpus requires in-depth knowledge of the annotation guidelines and remains a time consuming task even for expert annotators. For this reason, most treebanks are, rather than annotated from scratch, the result of a process where the output of an automatic parser is used as a basis for manual validation and editing. The performance of off-the-shelf UD parsers, however, is often unsatisfactory on learner text, independent of the L2 and parser in question (Huang et al., 2018; Di Nuovo et al., 2022; Volodina et al., 2022; Sung and Shin, 2023).

In this paper, we address the problem of how to best bootstrap the annotation of UD learner corpora. More specifically, we hypothesize that part of the decline in performance observed upon evaluating standard tools on L2 material is due to differences between the training and test domain that go beyond grammaticality. Learner sentences, for instance, may be unidiomatic without necessarily containing an error (cf. Table 1 for examples in English and Italian). Our research question therefore becomes whether utilizing THs in the training of a dependency parser is helpful for parsing original learner sentences and, if so, whether automatically annotated THs suffice for this use case.

To find out, we fine-tune an array of parsers on

¹learnercorpusassociation.org

²benjamins.com/catalog/ijlcr

³In the expression “L1-L2 parallel dependency treebank”, “L2” indicates original learner material, while “L1” refers to THs, assumed to be native-liked.

	LEARNER SENTENCE	TARGET HYPOTHESIS
EN	For electrical goods, there will be no any kind of electrical products except computer.	Regarding electrical goods, there will not be any kind of electrical product except computers.
IT	in quello momento era lei, che diventa furiosa!	In quel momento era lei che diventava furiosa!

Table 1: Example sentence-correction pairs from the two datasets used in our experiments, the ESL and VALICO-UD treebanks. The Italian sentence can be translated as “In that moment, she was the one who was getting furious!”. Note how both THs are grammatically correct but might be perceived as unidiomatic: a more proficient English speaker would probably use the word *electronics* rather than the expression *electrical goods/products*, while native-like Italian speakers tend to use the inchoative verb *infuriarsi* more than the construction *diventare furiosi* (literally “becoming furious”).

both manually and automatically annotated THs from two largest available L1-L2 treebanks, the English as Second Language (ESL) treebank (Berzak et al., 2022), and the VALICO-UD treebank of learner Italian (Di Nuovo et al., 2023). We then evaluate their performance on normative data, unseen THs and, crucially, original learner sentences, comparing it with that of baselines trained on large-scale reference treebanks.

2. Related work

Nonstandard language in general and learner language in particular still pose significant challenges for automatic annotation tools. Early experiments using the Turbo parser (Martins et al., 2013) on L1-L2 English data showed that grammatical errors negatively affect parser performance (Berzak et al., 2016). This was confirmed by a systematic study on dependency parsing for learner English, which concluded that, despite often misleadingly high overall scores, all tools considered were vulnerable to grammatical errors (Huang et al., 2018).

More recently, Di Nuovo et al. (2022) evaluated a UDPIPE 2 model trained on standard Italian on an L1-L2 treebank. They reported a substantial decline in performance on L2 originals, but also a more modest one on THs. Similarly, Volodina et al. (2022) assessed the accuracy of the Sparv annotation pipeline (Borin et al., 2016) on both original and normalized L2 Swedish sentences from the Swedish Learner Language corpus (SWELL) (Volodina et al., 2022) as well as on a corpus of Swedish course books, COCTAILL (Corpus of CEFR⁴-based Textbooks as Input for Learner Levels’ modelling) (Volodina et al., 2017). They observed both an 11-percentage-points performance gap between the original L2 Swedish sentences and the course-book material, and a significant - although smaller - discrepancy between the latter and normalized learner data. In addition, they reported a strong correlation between the parsers’ performance on

L2 texts - both normalized and not - and their authors’ CEFR-based proficiency level.

Work on parsers specifically meant for L2 material is limited, although notably Sakaguchi et al. (2017) combined dependency parsing with Grammatical Error Correction (GEC), building an error-repairing parser for learner English. To the best of our knowledge, however, all previous studies have focused specifically on dealing with ungrammatical input, while no attempts have been made to adapt parsers to the broader domain of learner essays.

3. Parsing experiments

As mentioned in the introduction, our goal is to find out whether corrections are suitable data for fine-tuning a parser aimed for original learner texts. To do that, we use MACHAMP (Massive Choice, Ample tasks) (van der Goot et al., 2021) to train and compare an array of models on both manually annotated (gold) and automatically parsed (silver) THs from two of the available L1-L2 treebanks.

MACHAMP is a toolkit that allows easy fine-tuning of deep contextualized embeddings for a variety of linguistic annotation tasks. It has been shown to be especially beneficial in cases where multiple datasets are available for the same task. This is exactly our case, as we want to combine large-scale UD treebanks of standard language with smaller, domain-specific training sets derived from the aforementioned learner treebanks.

In a nutshell, our approach consists of selecting a suitable BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2018) and fine-tuning it for dependency parsing on the largest UD treebank available for the language at hand until its performance is comparable to that of off-the-shelf tools. This leaves us with strong baselines to compare our specialized models with. We then continue training on silver- and, when available, gold-annotated THs. In this further fine-tuning step, the baseline pre-trained dependency parser is specialized on the specific domain of normalized learner essays. An alternative to this kind of sequential training would be building a single, larger

⁴Common European Framework of Reference for languages.

treebank	language	# sentences		
		train	dev	test
EWT	standard en	12544	2001	2077
ESL	learner en	2×5124	2×100	2×5024
ISDT	standard it	13121	564	482
VALICO-UD	learner it	2×1613	2×233	2×398

Table 2: Summary of the datasets used in our experiments. Note that ESL and VALICO-UD consist of L1-L2 sentence pairs and that VALICO-UD’s development set was sampled from its training set.

training set by mixing the reference treebanks with the THs. Creating and experimenting with different mixes, however, requires training multiple models largely on the same reference data, with the energy and time costs this implies. Our approach, on the other hand, only adds a few epochs of domain-specific tuning to the more resource-intensive training of the baselines, which is only carried out once.

It must be kept in mind that our current aim is not to build a general-purpose parser robust to learner language, but to develop a simple method to maximize parsing performance on a highly specific domain, even at the cost of a significant performance drop on standard language. This is because the resulting parser is meant to be used to speed up a single annotation effort. At the same time, however, we are interested in assessing whether and to what degree the introduction of THs negatively affects model performance on the standard test sets. We also want to compare the results obtained on original L2 sentences, which remain at least partially out-of-domain, with the performance on unseen THs. For these reasons, we test all of our models on all three evaluation sets at our disposal: that of the reference treebank and, when it comes to the learner corpora, both the L1 (TH) and L2 portions of their respective test splits.

Even though our models are trained in a multi-task setting,⁵ we focus on dependency annotation in its strictest sense. This is both for the sake of compactness and due to the fact that, when it comes to learner language, dependency parsing has been shown to be more problematic than most other linguistic annotation tasks (Volodina et al., 2022). We therefore evaluate our models only in terms of Labelled and Unlabelled Attachment Scores (henceforth LAS and UAS) (Kübler et al., 2009), computed with the official CoNLL-18 evaluation script (Zeman et al., 2018).

⁵The Italian model produces complete CoNLL-U files, while the English one is only trained for dependency parsing and POS tagging, as the ESL treebank does not provide any information regarding lemmatization or morphological features.

3.1. English

In our first experiment, we fine-tune the original monolingual English BERT model (Devlin et al., 2018). We train our baseline on the UD English Web Treebank (EWT), the gold standard dependency corpus for English (Silveira et al., 2023), using MACHAMP’s default hyperparameters. As can be seen in Table 3, the resulting performance even slightly surpasses the LAS and UAS scores reported for the UDPIPE 2.12 model trained on EWT we use for comparison (Straka, 2023).⁶

The THs used in the additional fine-tuning passes come from the English as a Second Language (ESL) treebank (Berzak et al., 2022),⁷ which is in turn based on the First Certificate in English (FCE) corpus (Yannakoudakis et al., 2011). The latter is a collection of short essays produced by learners with widely different language backgrounds, all taking the FCE exam, which assesses English at an upper-intermediate level (B1 in terms of the CEFR). As indicated in Table 2, the ESL treebank consists of 10000+ manually annotated L1-L2 sentence pairs, pre-split in a roughly same-sized training and test set and a smaller development set. Unlike most medium- to large-scale treebanks, ESL is manually annotated completely from scratch, with the goal of avoiding any potential annotation biases. Annotation is however limited to dependency labels and Part-of-Speech tags.

The default 20 training epochs were enough for the baseline to learn from the standard-language treebank. Consequently, on account of the training set sizes, we do not expect this further fine-tuning step to require more than 8 epochs. As MACHAMP allows for epoch-wise monitoring of development set performance scores, as well as because overfitting is not the main concern for our use case, however, we set the limit to 10. We then train our first specialized model, FT-GOLD, using the THs from the gold-annotated train and development splits of the ESL treebank. Indeed, most of the learning happens during the first 7 training epochs and scores start oscillating slightly after epoch 8, but peak performance on the development set is reached after training for all 10 epochs. As a consequence, we use the same settings for the FT-SILVER model. The only difference between the two is the training data: the latter uses automatically parsed versions of the same sentences, obtained by re-annotating them

⁶Note, however, that the comparison between MACHAMP UDPIPE 2 scores is not exact, as the UDPIPE 2 model was trained and evaluated on the latest versions of the treebank, which is in a format not yet fully supported by MACHAMP. For this reason, all data was preprocessed with the cleanup script provided as part of the MACHAMP toolkit before using it with our models.

⁷This treebank is also known as the Treebank of Learner English (TLE).

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	91.79	93.64	86.43	90.18	85.21	89.38
FT-GOLD	84.32	88.67	98.92	99.65	95.28	97.05
FT-SILVER	86.61	90.55	90.70	93.44	89.32	92.46
UDPIPE 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 3: LAS and UAS scores for all three evaluation sets for the full-scale English experiment.

with the same UDPIPE 2 model used as a reference for the baseline.

Results, summarized in Table 3, clearly show gold-annotated THs to produce an important performance improvement on ESL data over both the MACHAMP baseline and the UDPIPE pretrained model. Fine-tuning on automatically annotated THs results in a more modest improvement over the MACHAMP baseline, but is substantially equivalent to using the UDPIPE 2 EWT model. The latter, in fact, seems to have much better cross-domain generalization capabilities than our baseline, to the point that it performs slightly better on the THs than on its own test set. Finally, we note that the scores on the EWT evaluation set are higher for the FT-SILVER model than for FT-GOLD. This is unexpected, but possibly due to the fact that the silver-annotated THs follow the exact same annotation conventions as the EWT, as the UDPIPE 2 model has been trained on the EWT itself.

3.2. Italian

We repeat the same experiment with Italian data. This time, the starting pretrained model is an Italian BERT (MDZ Digital Library team at the Bavarian State Library, 2021) and the baseline trained on the Italian Standard Dependency Treebank (ISDT) (Simi et al., 2023).

Learner data comes from the VALICO-UD corpus (Di Nuovo et al., 2022), a UD-annotated subset of the VALICO (*Varietà Apprendimento Lingua Italiana Corpus Online*, “online corpus of learner varieties of the Italian language”), an L2 Italian learner corpus elicited by comic strips (Corino et al., 2017). VALICO-UD comprises 237 texts written by L2 Italian learners, all native speakers of one of four Western European languages (English, French, German and Spanish). While there is no mention of CEFR levels, proficiency can be to some extent inferred from reported years of study, ranging from 1 to 4. VALICO-UD is therefore more homogeneous than the ESL treebank in terms of L1 backgrounds, but much more heterogeneous when it comes to proficiency. As displayed in Table 2, VALICO-UD is over four times smaller than its English counterpart in terms of total size. Furthermore, only its test set is manually validated, while the rest of the data is

	ISDT		VALICO-UD L1		VALICO-UD L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	93.64	95.21	89.25	91.86	85.99	89.94
FT-SILVER	89.96	93.15	88.49	91.46	85.59	89.77
UDPIPE 2	93.34	94.96	90.22	92.86	87.69	91.61

Table 4: LAS and UAS scores for all three evaluation sets for the Italian experiment.

automatically parsed with the UDPIPE 2.12 ISDT model, meaning that it is not impossible to fine-tune on gold-annotated THs.

Nonetheless, we are interested in seeing whether the improvement over the MACHAMP baseline observed upon fine-tuning on silver THs in the English experiment can be replicated on a different dataset and, most importantly, with less training instances at our disposal. Since VALICO-UD does not come with a development set, we build one by randomly sampling sentences from the training data. The resulting development set is 10% of the total size of the corpus. In terms of hyperparameters, we stick to the same values used for the English experiment.

Unsurprisingly, results for this smaller dataset are less conclusive. Table 4 shows a pattern that is only partially similar to that of Table 3. On the one end, the performance of the fine-tuned model does decrease on the standard-language treebank while staying relatively high on the L1 and L2 evaluation sets. At the same time, however, none of the MACHAMP-based models outperforms UDPIPE on learner data, even if the MACHAMP baseline is marginally better on standard Italian.

3.3. Reducing the training set size

A simple explanation for the differences observed between the ESL and VALICO-UD-tuned silver models could be that the size of the Italian training set is too small to learn from THs. To test this hypothesis, we rerun the English experiment on a smaller sample of the ESL treebank, identical to the VALICO-UD training set in terms of number of sentences. Results, reported in Table 5, support this

	EWT		ESL L1		ESL L2	
	LAS	UAS	LAS	UAS	LAS	UAS
BASELINE	91.79	93.64	86.43	90.18	85.21	89.38
FT-GOLD-SAMPLE	84.11	88.70	94.53	96.50	92.21	94.82
FT-SILVER-SAMPLE	86.27	90.32	90.46	93.24	88.95	92.18
UDPIPE 2	90.56	92.62	90.70	93.44	89.42	92.51

Table 5: Scores for a smaller-scale English experiment, conducted by fine-tuning on a 1613-sentence ESL sample. We invite the reader to compare these results with those reported in Table 3 (same language, different training set size) and Table 4 (different language, same training set size).

only in part. If, as expected, both fine-tuned models are negatively affected by the lower amount of training instances, with `FT-SILVER-SAMPLE` never reaching `UDPIPE 2` performance, the `FT-GOLD-SAMPLE` model still performs better than `UDPIPE 2` on ESL data, although by a smaller margin than its fully-tuned counterpart, `FT-GOLD`. Furthermore, the difference between `FT-SILVER-SAMPLE` and `FT-SILVER` is almost negligible, suggesting that 1613 sentences should be sufficient to observe an improvement at least over the `MACHPAMP` baseline.

We therefore speculate that the differences observed between the full-scale English and Italian experiments may also depend on the fact that `VALICO-UD`, includes even beginner-level written productions, making the gap between `ISDT` and `VALICO-UD` generally wider than that between `EWT` and `ESL`. The more significant performance gap between standard and learner data observed when testing the `UDPIPE` model on Italian data seems to confirm this second hypothesis.

4. Concluding remarks

In this paper, we tried to establish whether fine-tuning a dependency parser on THs results in better performance on learner language. This was based on the hypothesis that the performance drop usually observed when applying an off-the-shelf parser on L2 data might not be exclusively due to the presence of grammatical errors, but also to the fact that standard tools are generally not trained on learner essays, which are therefore out-of-domain even when grammatically correct.

The results of our experiments on ESL data strongly suggest that gold-annotated THs are indeed helpful, although the generalizability of this finding can only be confirmed by repeating them on a different, fully manually annotated dataset, which is however not available at the time of writing.

Based on the results of this first experiment, in any case, we recommend initiating the annotation of a new L1-L2 corpus by validating the THs (or, if time allows it, by manually annotating them from scratch). While still requiring skilled UD annotators, this is a relatively straightforward task compared to annotating actual learner language, as the latter requires the development of new guidelines to deal with grammatical errors consistently. The resulting gold-annotated THs can then be used to fine-tune a parser that should help bootstrap the more challenging process of analyzing L2 originals. In the best of cases, this would leave the annotators with a treebank where only the ungrammatical segments require manual editing.⁸ In the near future, we plan

⁸As long as a good GEC pipeline is in place to generate the THs, this strategy should also be applicable to L2-only treebanks.

on testings this strategy on the Swedish data at our disposal.

Whether silver THs are useful is unclear. While the English experiments seem to indicate that automatically annotated corrections can benefit a `MACHPAMP` model and therefore help in the absence of a good pretrained parser, the results on `VALICO-UD` seem to contradict this finding in a way that cannot be explained solely by differences in dataset size. In this sense, further experiments with other L1-L2 treebanks are necessary, but not immediately possible. The aforementioned CFL (Chinese as a Foreign Language), the only other manually annotated treebank of this kind, consists of a mere 451 sentences, which makes it too small to generate training, development and test splits. At the same time, none of the larger learner corpora with target hypotheses comes with any extent of manual UD-annotation, which is however crucial for experiments like the ones described in this paper at least for the evaluation step. This further motivates us to proceed with the creation of a high-quality Swedish L1-L2 treebank.

An interesting byproduct of our parser evaluation is the observation that the ability to generalize to out-of-domain data appears to be much better for `UDPIPE 2` models than for `MACHPAMP`-based parsers, even if no overfitting is observed when evaluating the latter on an in-domain test set. This deserves further investigation, possibly in the context of a more systematic comparison of the cross-domain generalization capabilities of several mainstream UD parsers. When training a highly domain-specific tool, however, `MACHPAMP`, is a powerful, easy-to-configure alternative, as exemplified by the excellent performance obtained with the `FT-GOLD EWT + ESL` model, whose training did not even require a hyperparameter search. Building development sets that combine standard and non-standard language should also make it possible to train more robust `MACHPAMP` models.

5. Acknowledgements

This work is preparatory to the development of a treebank and parser for L2 Swedish, both of which are intended to enrich the Swedish national research infrastructure. As such, the research presented in this paper is supported by the Swedish national research infrastructure *Nationella Språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. Special thanks go to several colleagues at *Språkbanken Text* and to the anonymous reviewers for their constructive and attentive feedback at different stages of the writing process.

6. Bibliographical References

- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. [Universal Dependencies for learner English](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.
- Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. [Sparv: Språkbanken’s corpus annotation pipeline infrastructure](#). In *The Sixth Swedish Language Technology Conference (SLTC)*, Umeå University, pages 17–18.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. [VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies](#). *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).
- Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. [Dependency parsing of learner English](#). *International Journal of Corpus Linguistics*, 23(1):28–54.
- Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. [Dependency Parsing](#), chapter 6. Springer International Publishing, Cham.
- John Lee, Keying Li, and Herman Leung. 2017. [L1-L2 parallel dependency treebank as learner corpus](#). In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.
- André Martins, Miguel Almeida, and Noah A. Smith. 2013. [Turning on the turbo: Fast third-order non-projective Turbo parsers](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria. Association for Computational Linguistics.
- Arianna Masciolini. 2023. [A query engine for L1-L2 parallel dependency treebanks](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 574–587, Tórshavn, Faroe Islands. University of Tartu Library.
- Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. [Error-repair dependency parsing for ungrammatical texts](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195, Vancouver, Canada. Association for Computational Linguistics.
- Hakyung Sung and Gyu-Ho Shin. 2023. [Towards L2-friendly pipelines for learner corpora: A case of written production by L2-Korean learners](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 72–82, Toronto, Canada. Association for Computational Linguistics.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. [Massive Choice, Ample tasks \(MaChAmp\): A toolkit for multi-task learning in NLP](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.
- Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala, and Daniela Helena Piipponen. 2022. [Reliability of automatic linguistic annotation: native vs non-native texts](#). In *Selected papers from the CLARIN Annual Conference 2021*. Linköping University Electronic Press (LiU E-Press).
- Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. [The SwELL language learner corpus: From design to annotation](#). *Northern European Journal of Language Technology*, 6:67–104.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. [CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

7. Language Resource References

- Berzak, Yevgeni and Kenney, Jessica and Spadine, Carolyn and Wang, Jing Xian and Lam, Lucia and Mori, Keiko Sophie and Garza, Sebastian and Katz, Boris. 2022. *English-ESL/TLE-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.11. PID <http://hdl.handle.net/11234/1-5150>.
- Elisa Corino, Carla Marello, and Simona Colombo. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*, volume 6. Guerra. The corpus can be queried at valico.org/index.html.
- Di Nuovo, Elisa and Sanguinetti, Manuela and Bosco, Cristina and Mazzei, Alessandro. 2023. *Italian-VALICO-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID <http://hdl.handle.net/11234/1-5287>.
- Lee, John and Leung, Herman and Li, Keying. 2023. *Chinese-CFL-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID <http://hdl.handle.net/11234/1-5287>.
- MDZ Digital Library team at the Bavarian State Library. 2021. *Italian BERT*. Bavarian State Library. Distributed via HuggingFace.
- Silveira, Natalia and Dozat, Timothy and de Marnette, Marie-Catherine and Bowman, Samuel and Connor, Miriam and Bauer, John and Manning, Chris. 2023. *English-EWT-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID <http://hdl.handle.net/11234/1-5287>.
- Simi, Maria and Bosco, Cristina and Montemagni, Simonetta. 2023. *Italian-ISDT-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID <http://hdl.handle.net/11234/1-5287>.
- Straka, Milan. 2023. *Universal Dependencies 2.12 models for UDPipe 2*. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.12. PID <http://hdl.handle.net/11234/1-5200>.
- Volodina, Elena and Granstedt, Lena and Matsson, Arild and Megyesi, Beáta and Pilán, Ildikó and Prentice, Julia and Rosén, Dan and Rudebeck, Lisa and Schenström, Carl-Johan and Sundberg, Gunlög and Wirén, Mats. 2022. *SweLL-gold*. Språkbanken Text. Distributed via SBX/CLARIN. PID <https://hdl.handle.net/10794/846>.
- Volodina, Elena and Pilán, Ildikó and Eide, Stian Rødven and Heidarsson, Hannes. 2017. *COCTAILL*. Språkbanken Text. Distributed via SBX/CLARIN. PID <https://hdl.handle.net/10794/130>.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. *A new dataset and method for automatically grading ESOL texts*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics. The dataset can be downloaded at ilexir.co.uk/datasets/index.html.

SweDiagnostics: A Diagnostics Natural Language Inference Dataset for Swedish

Felix Morger

University of Gothenburg

felix.morger@gu.se

Abstract

This paper presents SweDiagnostics, a natural language inference dataset for Swedish based on the GLUE Diagnostic dataset. It is the largest, manually corrected NLI dataset in Swedish to date and can be used to evaluate models on NLI in Swedish as well as estimate English-Swedish language transfer capabilities. We present the dataset, the methodology used for translation, compare existing implementations and discuss limitations of the dataset, in particular those related to translationese.

1. Introduction

Natural language inference (NLI) is the task of determining the logical relationship between two sentences. More specifically, whether a hypothesis entails, is neutral to or contradicts a given premise. For example, the hypothesis “John walks down the street” entails the premise “John is moving”, but contradicts the premise “John is sitting” and is neutral to the premise “John is listening to music”. NLI datasets have been created and studied extensively in natural language processing (NLP), based on the assumption that inferential reasoning is needed for all kinds of NLP tasks, such as question-answering, reading comprehension and sentiment analysis.

For English, several NLI datasets have come out, most notably the Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018) and Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015), which have 570K and 433K sentence pairs respectively. For cross-lingual evaluation, the Cross-lingual Natural Language Inference (XNLI) dataset extends these datasets with a separate 5K test split and 2.5K validation split of sentence pairs using the same data collection procedure, but is also translated into 15 different languages (Conneau et al., 2018). While these larger datasets were produced to provide enough training data to train large deep learning models, they are usually contrasted with smaller NLI datasets aimed to evaluate some specific phenomenon of interest (Poliak, 2020)[p. 94], usually called *test suites* or *diagnostics*. These consist of handpicked examples meant to target some phenomenon of interest. For example, Winogender (Rudinger et al., 2018) targets gender pronoun resolution and FraCaS (Cooper et al., 1996) covers a range of semantic phenomena from generalized quantifiers to temporal reference.

In this paper, we present a Swedish post-edited translation of the GLUE Diagnostic dataset. To

date, this is the largest, manually corrected NLI dataset in Swedish (1106 samples), surpassing the only other Swedish NLI dataset SweFraCaS, which has 305 samples. This dataset allows for evaluating large language models (LLMs) on NLI for Swedish and English-Swedish language transfer similar to those done with the XNLI dataset. Such evaluation can be done by using existing English-Swedish machine translated NLI datasets as training data, such as those in Superlim and Overlim (Kurtz, 2022), an entirely English-Swedish machine translated version of SuperGLUE (Wang et al., 2019b).

The rest of the paper is structured as follows: We give an overview of the creation of the dataset (Section 2), compare existing implementations (Section 3), discuss potential limitations (such as with translationese and post-edited) and the development of future NLI datasets for Swedish (Section 4).

The dataset is available on HuggingFace as part of the Superlim project¹ and independently on the Språkbanken website.²

2. Dataset description

The GLUE Diagnostic dataset, which SweDiagnostics is based on, was released with the original SuperGLUE (Wang et al., 2019). It was hand-crafted by linguistic experts with the aim to create a dataset for diagnosing a system’s ability to solve a wide variety of language phenomena. The idea is to construct a hypothesis/premise sentence pair, where the entailment relationship depends on one or more targeted phenomena of interest. Table 1 illustrates this with two examples from the dataset. In the top sentence pair, the only difference be-

¹<https://huggingface.co/datasets/sbx/superlim-2>

²<https://spraakbanken.gu.se/resurser/swediagnostics>

tween the sentences is the added negation “did not” in the premise, which causes a contradiction. In the sentence pair below, the only difference is the added word “quietly” after “whispering” which is redundant since “whispering” (in most cases) implies talking quietly. Since the sentences express the same thing (i.e. talking quietly), the premise entails the hypothesis.

By way of this setup, NLI is used as a proxy to analyze specific language phenomena (negation and redundancy in the examples given). If the system can correctly predict the entailment relationship between the hypothesis/premise sentence pair, the conclusion is that the system encodes the targeted phenomenon.

The GLUE Diagnostic dataset has 33 different fine-grained language phenomena organized into four different coarse grained categories: lexical semantics, predicate-argument structure, logic and knowledge. Although the entailment relationship usually hinges on one particular fine-grained category, a sentence pair can be annotated with more than one category if the phenomenon is present in the text. Table 3 in the Appendix gives an exhaustive list of these categories as well as how many times they have been annotated in the dataset. For a detailed description of these categories, we refer to the latest documentation on the SuperGLUE website.³

2.1. Translation methodology

To create the equivalent SweDiagnostics, the sentence pairs were first machine translated using the Google Translate API. They were then post-edited by a native speaker of Swedish with a Master’s degree in linguistics (the author of this paper). Besides adapting the translations to sound fluent and coherent, the translator also strove to uphold the two following criteria.

1. The entailment relationship remains the same after translation.
2. The annotated language phenomena remain the same after translation.

Although these criteria could not be fulfilled for every category due to morphological differences, such as in expressing double negation, in general this was not a problem. This is because (a) Swedish and English are closely related languages and, thus, share many of the morphological and syntactical features which are used to construct the contrasting sentence pairs and (b) the majority of the targeted linguistic phenomena of GLUE Diagnostic dataset are high-level natural

³<https://super.gluebenchmark.com/diagnostics>

language understanding features, which are not dependent on the particularities of English grammar.

The choice of post-editing over translating from scratch was done for efficiency reasons (cf. [Plitt and Masselot \(2010\)](#); [Daems et al. \(2017\)](#)). During translation the translator had the option of adding notes to document ambiguous or difficult parts of translation. Only 6.7% included notes, indicating a generally light post-editing effort.

3. Implementations

At the time of writing, SweDiagnostics has been evaluated in two separate projects. Firstly, as SweDiagnostics is a part of Superlim ([Berdicevskis et al., 2023](#)) it has been evaluated on multiple language models, both monolingual Swedish models and multilingual models. Secondly, a more fine-grained analysis has been done by [Morger \(2023\)](#), comparing English-Swedish language transfer capabilities of Swedish monolingual and multilingual models. In both of these projects, an English-Swedish machine translated version of MultiNLI was used for training.

In the discussion below as well as in Table 2 and Figure 1, model names are shortened for space reasons with the following abbreviations: *mt* for the “megatron” model ([Shoeybi et al., 2019](#)), *sw* for “Swedish”, *l* for “large”, *c* for “cased” and *b* for “base”.

Table 2 shows the results on SweDiagnostics of [Berdicevskis et al. \(2023\)](#). Non-neural, supervised machine learning models are clearly outperformed by LLMs. The highest performing one is the multilingual model *xlm-roberta-large* ([Conneau et al., 2019](#)), outperforming the largest monolingual Swedish model *KBLab/mt-bert-l-sw-c-165k* ([Malmsten et al., 2020](#)). These results suggest that the amount of Swedish training data does not translate into increased performance. *KBLab/mt-bert-l-sw-c-165k*, for example, was trained on 70GB of only Swedish training data while *xlm-roberta-large* on 2.5TB of which only 12GB is in Swedish. The discrepancy in performance could also be explained by the difference in trainable parameters and language modeling objective. The fact that the sentences are originally English sentences could make it easier for the multilingual *xlm-roberta-large* model (see discussion in Section 4).

The results by [Morger \(2023\)](#), as reported in Figure 4, further compare the original GLUE Diagnostic dataset to SweDiagnostics. They concluded that a complete English-Swedish language transfer can be achieved using the English-

	Swedish	English	
P	Katten satt på mattan.	The cat sat on the mat.	contradiction (negation)
H	Katten satt inte på mattan.	The cat did not sit on the mat.	
P	Tom och Adam viskade i teatern.	Tom and Adam were whispering in the theater.	entailment (redundancy)
H	Tom och Adam viskade tyst i teatern.	Tom and Adam were whispering quietly in the theater.	

Table 1: Two premise (**P**) and hypothesis (**H**) sentence pair examples from SweDiagnostics. The outermost column indicates the entailment relationship between the sentences. The annotated linguistic phenomenon which determines the relationship is in parentheses and marked **bold** in the text.

Swedish machine translated dataset of MultiNLI (cf. `bert-b-c` on GLUE Diagnostic dataset (blue bar) and `KB/bert-b-sw-c (mt-sv)` on SweDiagnostics (orange bar)). However, training on the original English data and only relying on multilingual pretraining (`bert-b-ml-c`) did not reach the same level of performance. Comparing this to `xlm-roberta-large` in Table 2, this gap could possibly be filled by pre-training on more Swedish data or having larger architectures, but this remains speculative until a complete comparison has also been made to the `xlm-roberta-large` fine-tuned on English-Swedish machine translated data.

Overall, the fact that no model achieves higher than 0.44 Krippendorff’s α (see `KB/bert-b-sw-c (mt-sv)` in Figure 1) shows that this task is still difficult for Swedish LLMs. Only a score above 0.67 is considered moderate agreement between the predicted and golden labels (Marzi et al., 2024). However, LLMs have made great headway towards solving this task when compared to non-neural, supervised machine learning models, which have scores close to 0 (i.e. no agreement) (see Table 2).

4. Concluding remarks

This paper has presented SweDiagnostics, an NLI dataset for Swedish, which is a post-edited, manually corrected version of the GLUE Diagnostic dataset.

As we see it, this resource provides three main contributions. Firstly, given the scarcity of NLI datasets for Swedish, this resource is an important addition in order to get *any* insights into the performance on NLI in Swedish, in particular monolingual Swedish language models. This is especially important given the release of multiple new monolingual Swedish language models in recent years, such as `KB-BERT` (Malmsten et al., 2020) and `GPT-SW3` (Ekgren et al., 2023). However, as the original authors of GLUE Diagnostic dataset are careful to point out, GLUE Diagnostic dataset is a *test suite* and, thus, one should be careful not

Model	Krippendorff’s α
<code>xlm-roberta-large</code>	0.415
<code>KBLab/mt-bert-l-sw-c-165k</code>	0.393
<code>KBLab/mt-bert-b-sw-c-600k</code>	0.363
<code>KB/bert-b-sw-c</code>	0.349
<code>AI-Nordics/bert-l-sw-c</code>	0.347
<code>KBLab/bert-b-sw-c-new</code>	0.338
<code>xlm-roberta-base</code>	0.318
<code>NbAiLab/nb-bert-base</code>	0.314
Decision tree	0.037
SVM	0.026
Random forest	0.010
Random	0.004
MajLab/Avg	-0.404

Table 2: Evaluation results on SweDiagnostics as reported in Berdicevskis et al. (2023). They are reported in Krippendorff’s α (Krippendorff, 2011), the metric of choice for Superlim. These are the results on eight different pretrained language models (upper part of the table) and five non-neural machine learning models (lower part of the table).

to generalize over all language usage as it does not attempt to represent a natural language distribution. Secondly, SweDiagnostics’s parallelity to GLUE Diagnostic dataset enables the comparison of English-Swedish cross lingual representations, which complements other multilingual resources, most notably XNLI (Conneau et al., 2018), which does not include Swedish. Thirdly, given the annotation of language phenomena in the dataset (see Section 2), further comparison can be made on the performance between different linguistic categories.

Creating a new resource by machine translating and post-editing an existing resource has both advantages and disadvantages. One of the most obvious advantages is that it is a cheap and efficient way to create a new resource, while another advantage is the resulting parallel corpora, which enables a close comparison between the languages. A disadvantage is that the samples

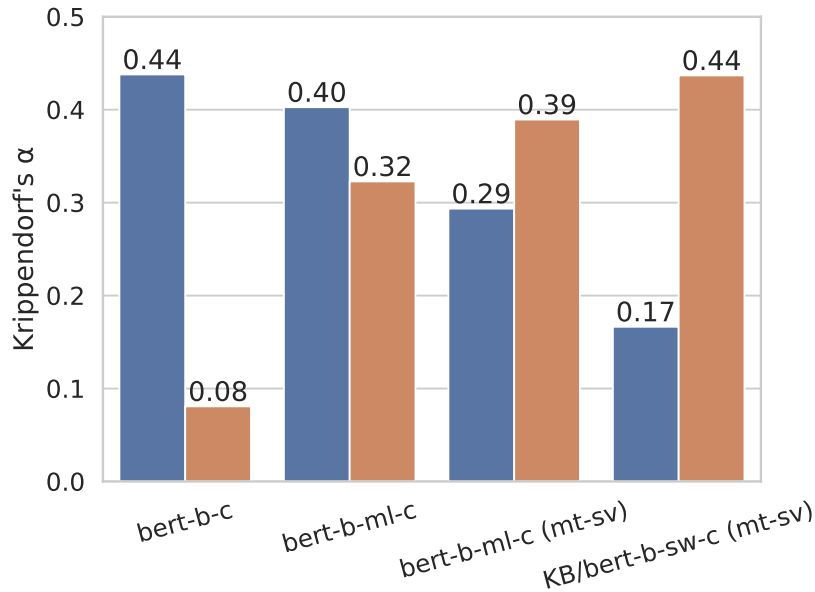


Figure 1: Model performance (Krippendorff's α) on the GLUE Diagnostic dataset (blue bars) and SweDiagnostics (orange bars) by (Morger, 2023). "mt-sv" refers to the model having been trained on the English-Swedish machine translated version of MultiNLI while the other ones are trained on the original English MultiNLI. Results can vary when compared with the original paper, which instead used the R_3 (Gorodkin, 2004) a three-class generalization of Matthews correlation coefficient (MCC).

are not taken from naturally occurring instances of the target language and will potentially not be a fair representation of the language overall. This is shown by Gellerstam (1986), which observe different statistical properties in translated language (translationese). This has also been shown to be further exacerbated by post-editing (Toral, 2019) (post-edited), however Daems et al. (2017) have shown that post-editing does not necessarily lead to lower quality translation. The results discussed in Section 3 do suggest that the performance on the GLUE Diagnostic dataset is highly transferable to SweDiagnostics, however, to what extent this is because of post-edited is unknown and could only be determined by future work systematically comparing post-edited to only human translations in the context of NLI.

This dataset together with SweFraCaS represents a first step towards evaluating NLI in Swedish. To get a fairer representation of Swedish and understand the effects of translationese, we encourage future work in creating new resources of NLI sourced from Swedish corpora. Comparing these to SweDiagnostics would not only give more insights into the NLI capabilities of Swedish monolingual and multilingual language models, but also insights into English-Swedish language transfer and language transfer between linguistically close languages more broadly.

5. Ethical considerations

As with any translated resource from a high-resource language to a lower-resource language, there is a risk of cultural biases being unfairly transferred to the target language. More broadly, using translated resources for evaluation could also amplify an anglocentric bias in what counts as the gold standard, which could divert funding from the creation of much needed unique language resources sourced directly from Swedish. For this reason, we encourage SweDiagnostics to be carefully compared with original Swedish resources and we also call for the creation of original NLI resources sourced exclusively from Swedish corpora.

6. Acknowledgments

This work was supported by Nationella språkbanken, which is jointly funded by the Swedish Research Council (2018–2024, grant no. 2017-00626) and 10 partner institutions.

7. Bibliographical References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). *CoRR*, abs/1911.02116.
- Joke Daems, Sonia Vandepitte, Robert J Hart-suiker, and Lieve Macken. 2017. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, 62(2):245–270.
- Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stoltenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. [Gpt-sw3: An autoregressive language model for the nordic languages](#).
- Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.
- Jan Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374.
- Klaus Krippendorff. 2011. Computing Krippendorff’s alpha-reliability. <https://www.asc.upenn.edu/sites/default/files/2021-03/ComputingKrippendorff'sAlpha-Reliability.pdf>.
- Martin Malmsten, Love Börjeson, and Chris Haf-fenden. 2020. [Playing with words at the national library of sweden – making a swedish bert](#).
- Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff’s alpha calculator: A user-friendly tool for computing krippendorff’s alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.
- Felix Morger. 2023. [Are there any limits to English-Swedish language transfer? a fine-grained analysis using natural language inference](#). In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 30–41, Tórshavn, the Faroe Islands. Association for Computational Linguistics.
- Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguistics*, 93:7–16.
- Adam Poliak. 2020. [A survey on recognizing textual entailment as an NLP evaluation](#). In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-1m: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
- Antonio Toral. 2019. [Post-editeese: an exacerbated translationese](#). In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). *Advances in neural information processing systems*, 32.

8. Language Resource References

- Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. [Superlim: A swedish language understanding evaluation benchmark](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page To appear, Sentosa, Singapore. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A

large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Robin Kurtz. 2022. [The KBLab blog: Evaluating Swedish language models](#).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. [Super-glue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Appendix

Coarse-grained	Fine-grained	Size	Neutral	Entailment	Contradiction	
Lexical Semantics	Factivity	68	37	17	14	
	Lexical entailment	140	37	49	54	
	Morphological negation	26	2	14	10	
	Named entities	36	12	18	6	
	Quantifiers	52	18	14	20	
	Redundancy	26	2	24	0	
	Symmetry/Collectivity	28	8	20	0	
Predicate-Argument Structure	Active/Passive	34	17	15	2	
	Anaphora/Coreference	58	22	24	12	
	Coordination scope	40	16	14	10	
	Core args	52	15	27	10	
	Datives	20	4	14	2	
	Ellipsis/Implicits	34	4	16	14	
	Genitives/Partitives	20	2	16	2	
	Intersectivity	46	25	19	2	
	Nominalization	28	4	18	6	
	Prepositional phrases	68	32	34	2	
	Relative clauses	32	16	12	4	
	Restrictivity	26	9	17	0	
	Logic	Conditionals	32	8	18	6
		Conjunction	40	15	15	10
Disjunction		38	17	15	6	
Double negation		28	2	22	4	
Downward monotone		30	17	13	0	
Existential		20	9	7	4	
Intervals/Numbers		38	11	9	18	
Negation		82	22	8	52	
Non-monotone		30	17	7	6	
Temporal		32	11	11	10	
Universal		18	5	7	6	
Upward monotone		34	19	15	0	
Knowledge		Common sense	150	36	56	58
	World knowledge	134	39	63	32	

Table 3: GLUE diagnostics coarse- and fine-grained phenomena of language phenomena.

Multiple Discourse Relations in English TED Talks and Their Translation into Lithuanian, Portuguese, and Turkish

Deniz Zeyrek¹, Giedrė Valūnaitė Oleškevičienė², Amália Mendes³

¹Cognitive Science Dept., Grad. School of Informatics, Middle East Technical University,

²Institute of Humanities, Mykolas Romeris University,

³Center of Linguistics, University of Lisbon (CLUL),

dezeyrek@metu.edu.tr, gvalunaite@mruni.eu, amaliamedes@letras.ulisboa.pt

Abstract

This paper focuses on multiple discourse relations, which refer to more than one sense relation between a pair of discourse segments. It shows how they are realized in English texts and their translations into Lithuanian, Turkish, and Portuguese in TED Multilingual Bank, an annotated corpus of English TED transcripts and translations into multiple languages. The paper overviews the annotation procedure and shows the change and variation of multiple discourse relations in the translations, such as omitting the *and*-component of multiple relations. The cross-linguistically framed analysis reveals that while both senses of a multiple relation can be explicitly conveyed, the salient sense is generally rendered through an overt connective. Even when it is not overtly expressed, it remains inferable and annotated during the annotation stage. By describing the different discourse structures arising from multiple relations and highlighting the implicated components in translation, the research contributes to the understanding of discourse, aims to raise the awareness of translators and translation educators, and bridge the gap between discourse analysis and translation.

Keywords: multilingual corpus, multiple discourse relations, translation, discourse connectives, implicature

1. Introduction

Translation is the process of conveying the messages of the source language into the structure of the target language, while preserving the purpose and essence of the message. Translation studies can benefit from discourse analysis in "discovering patterns and systematicity in the choices made by a translator and for hypothesizing reasons behind these choices on the basis of detailed discourse analytic procedures" (House, 2015, p. 49). Hence, translation researchers can use the knowledge accumulating in discourse analysis to understand how text pieces are structured to maintain coherence.

Discourse relations (also called coherence relations or rhetorical relations) are one of the ways clauses or sentences are structured (Mann & Thompson, 1988; Knott & Sanders, 1998; Marcu, 2000; Asher and Lascarides, 2003, among others). They hold between clauses, groups of clauses, or sentences and are named after the senses they convey – comparison, contrast, contingency, elaboration. They are expressed by a range of linguistic devices, such as conjunctions (*and*, *but*, *so*), adverbials (*however*, *in addition*), or prepositional phrases (*in summary*). These words or word groups are called discourse operators, discourse markers, cue phrases, or discourse connectives (Fraser, 1999), the term we use in the current work. They express a two-place semantic relation relating text spans that have an abstract object interpretation (eventualities, propositions, facts), as depicted by Asher (2012), or are complete clauses, as argued by Pasch et al. (2003).

Discourse relations are among the fundamental notions that enable discourse and pragmatics researchers to understand how texts are organized beyond the sentence level. Lately, the basic blocks, or anchors of discourse relations, i.e., connectives, have been examined extensively, mainly focusing on single (*but*, *so*, *instead*) and complex connectives (*on the contrary*). However, connectives are also known to co-occur with other connectives. In English, an adverb (*otherwise*, *instead*) and a conjunction (*because*, *if*, *so*) or two adverbials (*previously*, *for example*) may co-occur, forming constructions referred to as multiple connectives (Webber et al., 2019).

Examples (1) - (2), both taken from British National Corpus (<http://www.natcorp.ox.ac.uk>), illustrate the phenomenon of multiple connectives. More generally, this situation is referred to as multiple discourse relations, a notion that refers to more than one sense relation that holds between a pair of discourse segments:

- (1) but I'm just not enough of a Facebook user. So instead I'm going to use data from a few kind souls around our company
- (2) Pamela, you are now in my power. But if you comply with my proposals, I will leave you.

In the first example, multiple relations are signaled by a conjunction and an adverb relating the exact text spans. The connective *so* signals the consequence of the fact that I am not using Facebook frequently, *instead* conveys how I'm going to replace my infrequent use of Facebook. In the second example, *but* raises the expectation of a contradiction, and the

contrary expectation is fulfilled immediately in the next segment by the entire conditional sentence. The connectives *but* and *if* relate the interpretations of different spans. The semantic relations that hold between the clauses, i.e. text spans that correspond to the arguments of connectives reflect different discourse structures associated with (1) and (2) as depicted in Figure 1 and 2.

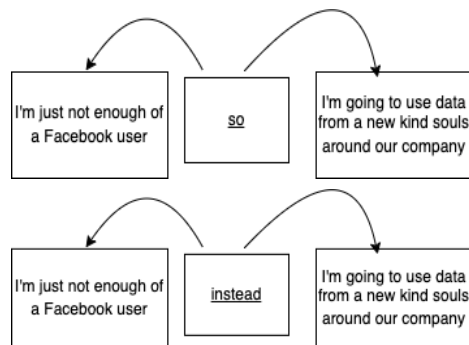


Figure 1: Discourse structure of example (1)

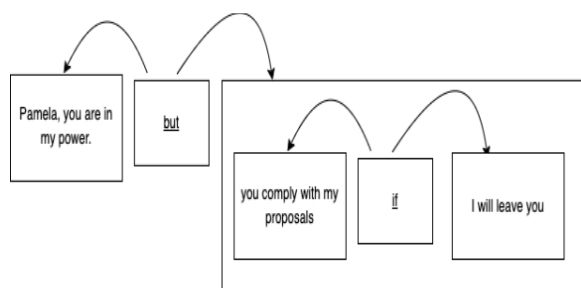


Figure 2: Discourse structure of example (2)

As opposed to example (1), example (2) generates full embedding (Lee et al., 2006), where the relation anchored by the conjunction *if* is fully embedded within one argument of the conjunction *but*. Implicit in this approach is that semantics and pragmatics of discourse are derived compositionally from the structure exposed in the discourse relations between different parts of the text.

In the current work, we are concerned with multiple connectives that link the interpretation of exact text spans, as example (1) shows.¹ Examples like (2) fall out of our scope mainly because they derive discourse structures that need to be analysed separately. What also falls out of scope of multiple relations involves parallel connectives (*not only .. but also*, *on the one hand .. on the other hand*) since, in these cases, one text span presupposes the other, and both parts of the connective act together to relate the text spans, as described in the PDTB 2.0 annotation manual (Prasad et al., 2007).

Multiple connectives are challenging for translators. Word-for-word translations may lead to incorrect or

¹ The discourse structure in Fig. 1 also allows multiple relations that are totally or partly realized by pragmatic

inappropriate results in target texts unless the context in which they occur is correctly interpreted. For example, the English connective sequence *but then* can function as a multiword connective conveying a single, concessive sense (see Fig. 3) though it can also function as a multiple connective conveying the contrastive/concessive sense followed by the temporal sense. This sequence will likely yield inappropriate translations if the human or machine translator misinterprets its meaning in the given context.

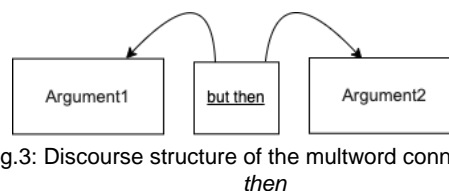


Fig.3: Discourse structure of the multiword connective *but then*

Given the potential benefits of multiple connectives or, more generally, multiple discourse relations to discourse and translation research, we ask: (a) What is the occurrence of multiple discourse relations in the source language, English? (b) What is their variation and change in translated languages? Our data is drawn from TED-Multilingual Discourse Bank or TED-MDB's English, Lithuanian, Turkish, and Portuguese parts (Zeyrek et al., 2020).

The rest of the paper proceeds as follows: Section 2 aims to set the ground and summarizes the related work. It also describes the challenges of automatically extracting multiple connectives from raw texts with no discourse connective annotation. Section 3 presents the data and methodology. It explains the differences between different versions of TED-MDB and describes how multiple relations are spotted, annotated, and then checked for inconsistencies. Section 4 analyzes the annotated data quantitatively, and Section 5 discusses how multiple relations are translated to target texts with examples from the corpus. Finally, in Section 6, the paper is summarized, and some conclusions are drawn.

2. Background

2.1 Discourse Connectives

All languages have discourse connectives, but they differ in various ways, for example, in terms of the inventory and grammatical class of connectives, as shown as early as 1998 by Stede and Umbach (1998). The literature on connectives and discourse relations is rapidly increasing, but due to space constraints, only one source will be referenced in this discussion. Therefore, for further exploration of this topic, readers are encouraged to refer to Zufferey & Degand (2024), who provide an up-to-date account of the theories and various applications in different languages.

connectives (*well anyway*, *well if*) but not analysed in the current paper.

Despite the vast literature on discourse relations and discourse connectives, work on multiple relations and multiple connectives is scarce. In one of the earliest works, Cuenca and Marin (2009) showed the presence and co-occurrence of discourse markers in spoken Catalan and Spanish. Fraser (2013) delineated the combinations of specific and general contrastive discourse markers in English, such as *however in contrast, but yet, but still*.² Zeyrek (2014) dealt with the patterns of co-occurring contrastive/concessive discourse connectives in Turkish. Most recently, Cuenca and Crible (2019) described different degrees of integration of adjacent discourse markers in English.

In a different line of research, Webber (2016) and Rohde et al. (2015, 2018) suggested that multiple connectives are not required to infer multiple senses because speakers can infer multiple relations even without multiple connectives. They argued that factors like the lexical semantics of the adverbials and the properties of the passages that contain them influence the particular relations available in contexts with multiple connectives.

Given the idea that speakers do not need overt connectives to infer discourse senses, how multiple relations of the source text are inferred is a legitimate question. Looking into this issue in an annotated, multilingual corpus will enable us to reach a more complete picture of shallow discourse structure, revealing what the annotators infer from multiple relations in the presence or absence of explicit connectives in the source text and translations.

2.1. The Challenge of Automatically Extracting Multiple Connectives

Exploitation of data without annotations has several challenges if the goal is to discover multiple connectives. Notably, the methods based on collocation are unlikely to produce the desired result. Collocation is a standard method to discover multiword expressions (MWE) (Constant et al., 2017). However, it must deal with many issues, such as ambiguity, which is difficult for all NLP tasks. A quick experiment that retrieves collocating connectives from the raw texts of TED-MDB showed that the ambiguity that impacts the identification of multiple connectives in which we are interested can be derived from two sources: (a) usage ambiguity -- the system cannot decide whether an item participating in a sequence of connectives is serving as a connective or not, (b) multiword ambiguity -- a multiword sequence can be analysed as a single connective as in *but then* or a parallel connective (Webber et al., 2019).³

Usage ambiguity may derive from the lexical ambiguity of words, and it also leads to multiword ambiguity. For instance, the Turkish word *ancak* is ambiguous between 'only' and 'however.' It functions as a connective when it means 'however' in its

context. Thus, the cluster *ancak sonra* 'only then' is a false positive token since *ancak* is not functioning as a connective but as the modifier of the adverbial 'then.'

Example (3) illustrates multiword ambiguity. Here, the Portuguese cluster *não so .. mas* 'not only .. but also,' a parallel connective, is a false positive.

(3) Pt Eles começaram a ver a sustentabilidade não só como uma coisa importante, mas como crucial para o êxito de_o negócio .

'They began to see sustainability not only as an important thing but as crucial to the success of the business.'

Other false positives are derived from cases where two connectives reflect full embedding. An example from Lithuanian *ir jei* 'and if' is provided in (4).

(4) Lt Tai vyksta dėl mūsų pasirinkimo, o galėtų ir nevykti, ir jei pagalvoji, ką darome su šiais duomenimis, tai lyg paimtume teleskopą ir nukreipę į miestą žiūrėtume lyg į mokyklos valgyklą (..)

'It's because of our choice, and it couldn't happen, and if you think about what we're doing with this data, it's like taking a telescope and pointing at the city, we're looking like a school canteen (..)'

For these reasons, manual annotation of multiple connectives and other types of multiple discourse relations is valuable, as it will pave the road toward better automatic systems.

3. The Current Work

The current work primarily focuses on analyzing the occurrence and variations of multiple discourse relations annotated in the English, Lithuanian, Portuguese, and Turkish segments of TED-MDB. Despite the limited scope of our data, our aim is to shed light on how multiple relations are translated and potentially open a new line of research. We hope that our observations will raise awareness, particularly by highlighting the implicated and consistently inferred components in translation.

3.1 TED-MDB

TED-MDB is a multilingual discourse corpus annotated by following the rules and principles of the Penn Discourse Treebank 3.0 (PDTB) (Webber et al., 2019). It is a resource of six TED talks in English and translations into multiple languages, with annotations revealing the shallow discourse structure of texts. With a connective-based approach, it annotates how the underlying discourse relations are realized in texts. Thus, it annotates explicit relations, those conveyed by an overt connective, and nonexplicit ones, where a connective is absent. The labels used for nonexplicit relations involve Implicit, Alternative Lexicalization (AltLex), Entity Relation (EntRel), and

² Fraser's specific discourse markers include *on the other hand, instead, rather* and general ones involve *but, yet, still*.

³ Our thanks to Mustafa Erolcan Er for running a preliminary experiment to automatically retrieve collocating connectives to spot candidate multiple connectives in raw texts.

No Relation (NoRel).⁴ A sense label is assigned to each relation except EntRels and NoRels. The senses are assigned by selecting the most appropriate semantic category from the PDTB 3.0 sense hierarchy based on four first-level senses: Expansion, Temporal, Contingency, and Comparison. Each of these categories is specified further at a second level. The first-level senses have second-level and, in some cases, third-level sense categories encoding directionality. Briefly, Expansion refers to the elaboration relations between two text spans. The category Temporal subsumes time-related eventualities. Contingency relations encompass Cause and Condition relations and their further specifications. Comparison refers to the relations between two eventualities where differences are highlighted.

For implicit relations, the annotators insert a connective that best captures the discourse sense inferred. These are called “implicit connectives” and distinguished from relations cued by overt connectives, known as “explicit connectives” in the PDTB framework.

English, Portuguese, Lithuanian, and Turkish parts of TED-MDB have gone through several updates since the first release of the corpus. Multiple relations are systematically annotated in the most recent version of the PDTB (version 3.0), and the four language sets of TED-MDB have been updated by these new principles. Appendix 1 lists the distribution of discourse relation realization types across languages in TED-MDB.

The main extensions over TED-MDB are described in Özer et al. (2022). In the extended version, intra-sentential implicit relations and multiple relations are annotated in four languages by searching them in the circumstances determined by the PDTB 3.0 (these are listed in Table 1). Although a few multiple relations have already been annotated in the first version, more instances are spotted while annotating intra-sentential relations in the extended version. The extended version also involves the automatic alignment of the discourse relations of three target languages with the English part. A relation-linking approach is developed, where each argument of a discourse relation, its realization type, and its senses are matched. Relation linking through word alignment achieved a reasonable degree of precision, meaning the links it finds are highly likely to be an actual match.

In the current work, we rechecked the aligned dataset by examining it manually for alignment errors or inconsistencies.

⁴ In the PDTB framework, the label AltLex stands for relation types that contain an alternative way of lexicalizing a discourse relation (*for this reason, as a consequence*). Annotators spot them while annotating a relation, where the insertion of an overt connective leads to redundancy. For this reason, they are grouped as nonexplicit relations. The PDTB 3.0 introduced a new relation realization type,

3.2 Annotation of Multiple Relations

In the revised version of TED-MDB, multiple relations are searched in three circumstances introduced in the PDTB 3.0.⁵

Multiple relations are searched in instances where:

- 1 a relation that holds between two discourse segments is conveyed by a multiple explicit connective,
- 2 the explicit connective *and* conveys one relation between a pair of spans, annotators infer (and insert) a separate sense, as well,
- 3 there is an implicit relation between two spans, and annotators also infer (and insert) a separate sense.

Table 1: Circumstances where multiple relations are searched and annotated

3.2.2.1 Circumstance1 (Multiple Explicit Connectives): Each component of multiple explicit connectives is annotated separately with the argument spans they link and their respective senses.

3.2.2.2 Circumstance2 and Circumstance3: Multiple relations in these circumstances are annotated only in connection with the explicit or implicit conjunction *and* anchoring intra-sentential implicit relations.

To correctly identify multiple relations in circumstance2 and circumstance3, annotators are guided by specific questions. For example, multiple relations that fit circumstance2 are spotted by asking the questions shown just below excerpt (5):

(5) Now these initiatives create a more mobile workplace, and they reduce our real estate footprint ... (TED-MDB, Talk no. 1927)

Do you infer an implicit sense conveyed by ‘and’, such as causality or temporality? If so, annotate it separately by inserting an appropriate implicit connective such as ‘so’ or ‘then’.

The questions below example (6) capture multiple relations of circumstance3.

(6) The petals unfurl, they open up, the telescope turns around. (TED-MDB, Talk no. 1976)

What is the implicit discourse relation that holds between adjacent clauses? Annotate each relation separately by inserting an implicit connective. If you infer another sense, annotate it as well with an appropriate implicit connective you will insert.

Hypophora, annotated as AltLex in TED-MDB (Zeyrek et al., 2018).

⁵ The fourth instance, which the PDTB 3.0 annotates, involves cases where an AltLex or AltLexC conveys one relation between a pair of spans, but annotators also infer a different sense. However, these instances are not annotated in TED-MDB.

4. Analysis of the Corpus

4.2 Multiple Explicit Connectives

The analysis of the revised dataset shows that the least frequently occurring type of multiple relations is multiple explicit connectives (the first category in Table 1). There are 7 such tokens in English all involving the use of *and* and a separate discourse adverbial, with a corresponding number of 7 tokens in Portuguese, 6 in Lithuanian and the Turkish set. Examples from English and the matching connectives in translations are listed in Appendix 2. The table shows that the way the relation is conveyed differs as the translators may sometimes omit one of the components of multiple explicit connectives though they often translate both parts verbatim (also see the examples in Section 5).

4.3 Multiple Relations with Inferred Senses

Multiple relations annotated with one or more inferred senses, those spotted in *circumstance2* and *circumstance3* of Table 1, occur more frequently than multiple explicit connectives. In the English section of the corpus, 56 multiple relations with inferred senses are found. Appendix 3 presents these relations categorized by the connectives, where the implicit or explicit connective *and* and the inferred senses are counted separately.

To understand the change and variation in the translation of multiple relations with inferred senses, we checked how many relations of the source language are aligned with target texts. The analysis showed that Portuguese texts have the highest number of corresponding relations (35), while Turkish and Lithuanian translations have lower numbers (32 and 31, respectively).

Secondly, we checked the discourse relation realization types of the matching relations. The results are presented in confusion matrices in Tables 2 – 4. The tables show that translators vary regarding how they render English multiple relations with inferred senses; for example, they employ the well-known translation strategies of explicitation or implicitation.⁶ The tables show that although all target languages resort to implicitation and explicitation, Portuguese ranks the highest in the implicated and explicitated cases. For instance, according to Table 2, Portuguese translators implicated 6 instances of the *and*-component out of the 16 cases aligned with English.

Tables 3 and 4 indicate that Turkish and Lithuanian translations tend to retain the *and*-component more often compared to Portuguese, suggesting a lower frequency of its implicitation.

⁶ In this work, implicitation refers to omitting a discourse connective present in the source text. In Tables 2-4, implicitation is assessed by the number of times an explicit relation of the source text is translated as an implicit

	Pt			Total
	En	AltLex	Explicit	
Explicit	0	10	6	16
Implicit	1	9	9	19
Total	1	19	15	35

Table 2: How explicit-*ands* and accompanying implicit components, as annotated in English, are realized in Portuguese

	Tr					Total
	En	AltLex	EntRel	Explicit	Implicit	
Explicit	0	0	11	2	0	13
Implicit	2	1	5	10	1	19
Total	2	1	16	12	1	32

Table 3: How explicit-*ands* and accompanying implicit components, as annotated in English, are realized in Turkish

	Lt			Total
	En	Explicit	Implicit	
Explicit	10	2	12	
Implicit	2	17	19	
Total	12	19	31	

Table 4: How explicit-*ands* and accompanying implicit components, as annotated in English, are realized in Lithuanian

Thirdly, we investigated how translators addressed the inferred components that accompany *and*-relations annotated in English. Within our dataset, these inferred components encompass various senses, including Cause (expressed by 'as a result', 'so', 'consequently'), Purpose ('in order'), Temporality ('then'), and Level of detail ('in other words') (refer to Tables 5 – 7).

	Tr					Total
	En	AltLex	EntRel	Explicit	Implicit	
as a result	1	1	1	4	1	8
consequently	0	0	1	0	0	1
in order	1	0	0	0	0	1
in other words	0	0	1	0	0	1
so	0	0	0	1	0	1
then	0	0	1	3	0	4
therefore	0	0	1	0	0	1
Total	2	1	5	8	1	17

Table 5: How inferred components, as annotated in English, are realized in Turkish

The tables indicate that target languages can make the implicit component explicit. For example, Turkish renders 5 out of 17 inferred senses explicitly, and Portuguese renders 8 out of 16 (see Tables 5 and 6). Lithuanian translations demonstrate a higher degree of faithfulness to the original texts, explicitating few of

relation. Explicitation is the reverse process, where a connective is used in translation, although it is absent in the source text.

the senses inferred in the English annotation process (refer to Table 7).

En	Pt			Total
	AltLex	Explicit	Implicit	
as a result	1	2	2	5
consequently	0	3	0	3
in order	0	1	0	1
in other words	0	0	0	0
so	0	0	1	1
then	0	2	3	5
therefore	0	0	1	1
Total	1	8	7	16

Table 6: How inferred components, as annotated in English, are realized in Portuguese

En	Lt			Total
	Explicit	Implicit	Total	
as a result	2	3	5	
consequently	0	1	1	
in order	0	0	0	
in other words	0	0	0	
so	0	0	0	
then	0	5	5	
therefore	0	0	0	
Total	2	9	11	

Table 7: How inferred components, as annotated in English, are realized in Lithuanian

Tables 5 – 7 demonstrate that the implicit sense associated with *and*-relations in English annotations remains discernible in translations and is annotated accordingly during the annotation stage. This pattern is consistently observed across all translations, indicating that the senses inferred during the English annotation stage are also identified during their respective annotation stages. Further research on more significant amounts of data is needed but these initial observations imply that the more salient sense is either translated overtly or if not, it remains discernible and labeled with an appropriate implicit connective during the annotation stage.

5. Discussion

This section zooms into specific examples from the corpus to assess the change and variation in the translation of multiple relations.

5.1 Translating Multiple Explicit Connectives

Since the annotations mainly capture the connective *and* in multiple relations, this section focuses on its usage in the source text and possible implicitation in translated texts.

It is known that the connective *and* is highly prone to implicitation (Zufferey, 2016), and researchers have suggested that this is due to its being a weak conjunction (Asr and Demberg, 2012) or an underspecified discourse marker (Crible et al., 2019).

Whether the *and*-component of multiple connectives is kept in the translation or undergoes implicitation is

interesting, as it could contribute to the current understanding of the implicitation of *and*. However, it may also be the case that each connective of a multiple relation is kept or omitted independently of the other, and it is worth looking into the data with this perspective.

In the rest of this section, the examples are presented in each of the four languages if a target text is aligned with the source text. The annotated explicit connective is underlined, and the discourse realization type and the sense(s) are shown in parentheses.

Example (7) concerns *and so*: with *and*, the speaker signals a continuation; with *so*, a consequence. The consequence (or Cause:Result) sense is added to the discourse after signaling the continuation. So, readers interpret the text as follows: The fact that many of the author's early memories involved intricate daydreams is a result of the deep restlessness, a primal fear that they would fall prey to a life of routine and boredom.

(7) En There was a deep restlessness in me, a primal fear that I would fall prey to a life of routine and boredom. And so many of my early memories involved intricate daydreams where I would walk across borders, forage for berries, and meet all kinds of strange people living unconventional lives on the road. (Explicit; Expansion:Conjunction; Contingency:Cause)

Tr İçimde derin bir rahatsızlık var hayatın tek düzeliğine ve sıklıkla kurban düşeceğime dair ilkel bir korku. Ve bu yüzden çocukluk dönemi hatıralarımın çoğu sınırlarda yürüyüp, çilek peşinde koştuğum ve farklı farklı insanlarla karşılaştığım yollarda sıradışı bir hayat sürdürdüğüm karma karışık hayallerdi. (Explicit; Expansion:Conjunction; Contingency:Cause)

Lt Nenustygau vietoje, bijojau, kad tapsiu rutinos ir nuobodulio grobiu. Todėl daugumoje mano vaikystės prisiminimų įvairūs užsisvajojimai, kuriuose aš kertu sienas, ieškau uogų, sutinku visokiausius keistuolius - nesuvaržytus, gyvenančius kelyje. (Explicit; Contingency:Cause)

Pt Sentia uma profunda inquietação, um medo primordial de que seria vítima de uma vida de rotina e aborrecimento. Por isso muitas das minhas primeiras memórias envolviam sonhar acordada e de forma elaborada onde passaria fronteiras, a recolher bagas e a conhecer todo o tipo de pessoas estranhas, com vidas fora do convencional, pela estrada fora. (Explicit; Contingency:Cause)

Analysis of the translations of (7) reveals diverse strategies employed by translators. Some translators opt to directly translate the multiple explicit connective, while others choose to convey only the salient sense, such as Cause:Result, using an explicit connective. For instance, the Turkish translation maintains both the Expansion and Cause senses through equivalent multiple explicit connectives. However, the approaches differ in Lithuanian and Portuguese translations. In these languages, *and* is

implicated, and the Cause sense is conveyed using a single explicit connective. Consequently, during the annotation stage, the Expansion sense is not inferred, yet annotators from both languages consistently infer the Cause sense. Despite these variations in translation, there is convergence among translators across different languages, as the more salient sense is always inferred. Example (8) is an instance of *and then*. Again, the connective signals a continuation of the discourse; then, the temporal relation is added. The interpretation is that we can see those planets after the star shade flies 50,000 kilometers from the telescope and is held right in its shadow. In this instance, the Portuguese and Lithuanian translations perfectly match the source text, capturing both discourse relations of the original text with equivalent multiple explicit connectives. However, in the Turkish translation, the relation is conveyed by a different connective type: a modified AltLex *ancak bu şekilde* 'only in this way', conveying a Manner sense. The annotator infers a Manner and an implicit Conjunction relation during the annotation stage. Some more cases like this exist in the corpus and reveal that the salient relation of the source text may be interpreted differently by the annotators of different languages due to a mismatching connective used in the translation.

(8) En (..) it [the star shade] has to fly 50,000 kilometers away from the telescope that has to be held right in its shadow, and then we can see those planets. (Multiple Explicit, Expansion:Conjunction; Temporal:Asynchronous)

Pt Esta sombra estelar tem cerca de metade do tamanho de um campo de futebol e tem que se distanciar 50 000 quilômetros do telescópio que tem que ser mantido na sua sombra, e então poderemos ver os planetas (Multiple Explicit, Expansion:Conjunction; Temporal:Asynchronous)

Lt (..) jis turi atsidurti tikslioje vietoje ir tada pamatysime tas planetas. (Multiple Explicit, Expansion:Conjunction; Temporal:Asynchronous)

Tr Bu yıldız gölgeleyici yaklaşık yarım futbol sahası büyüklüğünde ve gölgesi içinde tutulması gereken teleskoptan 50.000 kilometre uzakta uçması gerekiyor. (implicit = ve 'and') Ancak bu şekilde gezegenleri görebiliriz. (AltLex, Expansion:Manner)

In summary, in this section, our corpus analysis provides insights into the translation of multiple explicit connectives. We observe that the more salient meanings, such as Cause or Temporality, are generally preserved in the target text through explicit connectives, while the less salient sense, such as Expansion:Conjunction conveyed by *and* tends to be implicated. That is, the *and*-component of a multiple explicit relation may be implicit in translation, but the more salient sense remains discernible via overt connectives. These observations suggest that multiple relations exhibit varying degrees of saliency. However, our analysis also identifies translation mismatches, which are inherent to the translation process. TED translators, often non-professionals,

may encounter challenges in conveying the multiple senses of the source text, possibly due to linguistic and contextual issues, leading to occasional discrepancies in the outcome.

5.2 Translating Multiple Relations with Inferred Senses

Having investigated how the original multiple explicit connectives (those spotted in circumstance1 of Table 1) are translated, this section focuses on how English multiple relations of the second and third circumstances are rendered in translation.

Example (9) involves an implicit intra-sentential relation; the English annotator infers multiple senses regardless (Conjunction, Temporal). The sentence is translated into Portuguese and Lithuanian verbatim. The annotators of these languages infer different senses: The Lithuanian annotator infers an Expansion and a Temporal relation that holds the clauses together. In the annotation stage, the connective *ir* is inserted to anchor the Expansion relation, the connective *tada* to anchor the Result sense, as required by the annotation guidelines. In Portuguese, however, only the Temporal sense is inferred - the relation is labeled with a single implicit connective (*depois* 'later'). Turkish translation differs from the others because the relation is translated with an overt cue (*ve* 'and'), and in the annotation stage, it is labeled with the sense of Asynchronous. Like the example (7) discussed above, this example indicates the salience of the Asynchronous sense because, in all the target texts, the relation is assigned the Asynchronous sense at the annotation stage, among other inferred senses, if any. The revised dataset has many examples where the target relation is not captured as a multiple relation. Nevertheless, annotators consistently identify the more prominent meaning annotated in the English multiple relation in our data.

(9) En (..) they open up (Imp1 = and, Imp2 = then) the telescope turns round (Implicit: Expansion:Conjunction; Temporal:Asynchronous)

Lt Žiedlapiai skleidžiasi, atsiveria (Imp1 = ir 'and' Imp2 = tada 'then') teleskopas apsisuka (Implicit: Expansion:Conjunction; Temporal:Asynchronous)

Pt (..) abrem -se (Imp = depois 'later') o telescópio vira -se (Implicit, Temporal:Asynchronous)

Tr Yapraklar açılıp genişliyor ve teleskop yön değiştiriyor (Explicit; Temporal:Asynchronous)

Finally, the intra-sentential relation in (10) is expressed through an explicit *and*-relation in English, and the annotator infers an Expansion:Conjunction sense and a separate Cause:Result sense as well. In translating this text, the explicit connective is kept in Portuguese and Lithuanian. In the annotation stage, the translations are labeled the same way as English. In the Turkish translation, the explicit connective is omitted. In the absence of *and*, only the implicit Result sense is inferred at the annotation stage. Once more, we interpret the varied annotations of this example not

as divergence but as convergence. This is because annotators from different languages consistently infer the more prominent Causal sense, even when it is not explicitly expressed in the target text.

(10) En It's a terrible shadow, and (Imp = as a result) we can't see planets. (Explicit; Expansion:Conjunction; Contingency:Cause)

Pt (..) uma sombra terrível E (Imp = por conseguinte) não conseguimos ver planetas (Explicit; Expansion:Conjunction; Contingency:Cause)

Lt Šešėlis didžiulis. Ir (Imp = todėl) planetų mes nematome (Explicit; Expansion:Conjunction; Contingency:Cause)

Tr Bu kötü bir gölge. (Imp = böylece 'thus') Gezegenleri göremeyiz. (Implicit; Contingency:Cause).

6. Summary and Conclusion

In summary, the analysis of multiple relations in the source text and their translation to multiple languages revealed the following:

- The analysis of multiple explicit relations that involve *and* revealed that the source text is often translated by keeping both components and if not, the more salient sense is inferrable by the annotator. These findings underscore the presence of varying degrees of saliency in multiple relations. Generally, it is the more salient relation that is explicitly conveyed, while the less salient one, that is, the expansion sense of *and*, can be implicitated.
- To further understand the issues surrounding multiple relations, we investigated *and*-relations where additional senses are inferred. In these cases, whether the explicit *and*-component is kept in translation or undergoes implicitation, our observation holds: In both of these instances, annotators of different languages often converge in inferring the salient sense in the target text corresponding to that annotated in English.

- Finally, inappropriate translations or human error in the annotation stage cannot be totally overridden. These create noise in the data and should be analyzed with caution.

The present study shows the use of a parallel, aligned dataset in investigating a specific discourse phenomenon. It sheds light on how multiple relations are treated in translation and invites translation researchers to consider these constructions in different languages. For NLP researchers, it emphasizes the challenges of automatically extracting multiple connectives, and given manual annotation costs, it highlights the need to develop discourse parsers that handle them as well as other connective types. Finally, the research also has implications for pedagogy since it increases the awareness of translators and translation teachers.

Despite these conclusions, the work is not without its limitations. The annotated multiple relations are limited to those where the connectives link the same spans and only the multiple relations associated with the conjunction *and*. The results are limited by the overall corpus size and four language sets. Multiple relations need to be readdressed in future investigations by drawing upon different genres and more amounts of data in different target languages.

In future research, we aim to experiment with the automatic extraction of multiple explicit connectives in both English and translated languages using parallel corpora. This endeavor will enhance our understanding of shallow discourse structure from the view of multiple relations cross-linguistically and contribute to the development of more robust computational tools for discourse and translation.

7. Appendices

Appendix 1: The distribution of discourse realization types across languages in TED-MDB (Özer et al., 2022)

Language	Explicit	Implicit	Alex	EntRel	NoRel	Total
English	289 (40%)	254 (36%)	46 (6%)	78 (11%)	49 (7%)	716
German	240 (43%)	214 (38%)	17 (3%)	59 (11%)	30 (5%)	560
Lithuanian	377 (46%)	315 (38%)	18 (2%)	79 (10%)	32 (4%)	821
Polish	218 (37,5%)	195 (33,5%)	11 (2%)	104 (18%)	52 (9%)	580
Portuguese	269 (40%)	311 (46%)	29 (4%)	38 (6%)	33 (5%)	680
Russian	237 (42%)	221 (39%)	20 (4%)	57 (10%)	30 (5%)	565
Turkish	315 (41%)	264 (35%)	60 (8%)	70 (9%)	51 (7%)	760
Total	1945	1774	201	485	277	4682

Appendix 2: Multiple explicit connectives in English and their correspondences in target texts. Connectives that are implicated are enclosed within braces.

En DRID	En	Pt	Tr	Lt
DR169	and at the same time	e ao mesmo tempo	ve aynı zamanda	ir tuo pačiu
		'and at the same time'	'and at the same time'	'and at the same time'
DR25	and so	então	böylece	taigi
		'then / so'	'thus'	'thus'
DR30	and then	depois	sonra da	ir tik tada
		'after'	'then'	'and only then'
DR112	and then	e então	ancak bu şekilde	ir tada
		'and then'	'only in this way'	'and then'
DR80	and so	e por isso	ve (sonuçta)	dél to
		'and due to this'	'and (consequently)'	'therefore'
DR120	and so	(consequentemente)	Non-aligned	Non-aligned
		'consequently'	--	--
DR42	and so	por isso	ve bu yüzden	todél
		'due to this'	'and due to this'	'therefore'

Appendix 3: English Multiple relations in Circumstance2 and Circumstance3 with their explicit and implicit components

Type	Discourse Connective									Total
	And	and	as a result	consequently	in order	in other words	so	then	therefore	
Explicit	2	18	0	0	0	0	0	0	0	20
Implicit	0	8	10	4	1	1	1	10	1	36
Total	2	26	10	4	1	1	1	10	1	56

8. Bibliographical References

- Asher, N. (2012). *Reference to abstract objects in discourse*. Vol. 50. Springer Science & Business Media.
- Asr, F. T., & Demberg, V. (2012). Implicitness of discourse relations. In Proceedings of COLING 2012 (pp. 2669-2684).
- Crible, L., Abuczki, Á., Burkšaitienė, N., Furkó, P., Nedoluzhko, A., Rackevičienė, S., Valunaite Oleskeviciene, G., & Zikánová, Š. (2019). Functions and translations of discourse markers in TED Talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142, 139-155.
- Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4), 837-892.
- Cuenca, M. J., & Crible, L. (2019). Co-occurrence of discourse markers in English: From juxtaposition to composition. *Journal of Pragmatics*, 140, 171-184.
- Cuenca, M. J., Marin, M. J. (2009). Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics*, 41(5):899-914.
- Zufferey, S., & Degand, L. (2024). *Connectives and Discourse Relations*. Cambridge University Press.
- Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics* 31, no. 7 (1999): 931-952.
- Fraser, B. (2013). Combinations of contrastive discourse markers in English. *International Review of Pragmatics*, 5(2), 318-340.
- House, J. (2015). Global English, discourse, and translation. Linking constructions in English and German popular science texts. *Target. International Journal of Translation Studies* 27(3): 370-386.
- Knott, A., Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics* 30.2 (1998): 135-175.
- Lee, A., Prasad, R., Joshi, A., Dinesh, N., & Webber, B. (2006, December). Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax?. In *5th International Workshop on Treebanks and Linguistic Theories*.
- Mann, W. C., Thompson, S. A. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse* 8, no. 3 (1988): 243-281.

- Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT Press.
- Özer, Sibel, Murathan Kurfalı, Deniz Zeyrek, Amália Mendes, and Giedrė V. Oleškevičienė. 2022. Linking discourse-level information and induction of bilingual discourse connective lexicons. *Semantic Web*, Vol. Pre-press, pp. 1-22.
- Pasch, R., Brauße, U., Breindl, E., & Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren: linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfers* (Konjunktionen, Satzadverbien und Partikeln) (Vol. 2). Walter de Gruyter.
- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A.K., Robaldo, L., & Webber, B.L. (2007). The Penn Discourse Treebank 2.0 Annotation Manual.
- Rohde, H., Dickinson, A., Clark, C., Louis, A., Webber, B., (2015). Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics* (pp. 22–31).
- Rohde, H., Johnson, A., Schneider, N., & Webber, B. (2018). Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2257-2267).
- Stede M, Umbach C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2* (pp. 1238-1242). Association for Computational Linguistics.
- Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistics Annotation Workshop* (pp. 22–31). Berlin: Association for Computational Linguistics.
- Webber, B., Knott, A., & Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. *Computing Meaning: Volume 2*, pp. 229–245.
- Webber, B., Prasad, R., & Lee, A. (2019, May). Ambiguity in explicit discourse connectives. In *Proceedings of the 13th International Conference on Computational semantics-Long papers* (pp. 134–141).
- Zeyrek, D. (2014). *On the distribution of contrastive-concessive discourse connectives ama (but/yet) and fakat (but) in written Turkish*. In P. Suihkonen and L.J. Whaley, editors, *On Diversity and Complexity of Languages Spoken in Europe and North and Central Asia*.
- Zeyrek, D., Mendes, A., & Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the 11th Language Resources and Evaluation Conference-LREC'2018* (pp. 1913-1919). European Language Resources Association.
- Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., & Ogrodniczuk, M. (2020). TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54, 587-613.
- Zufferey S. (2016). Discourse connectives across languages. Factors influencing their explicit or implicit translation. *Languages in Contrast* Vol. 16(2): 264–279.

mini-CIEP+ : A Shareable Parallel Corpus of Prose

Annemarie Verkerk and Luigi Talamo

Saarland University / Saarbrücken, Germany
{annemarie.verkerk, luigi.talamo}@uni-saarland.de

Abstract

In this paper we present mini-CIEP+, a sharable parallel corpus of prose. mini-CIEP+ consists of the first part of ten different works of prose across many different languages, allowing for the cross-linguistic investigation of larger discourse units. Subcorpora typically contain 5750 sentences and almost 125K tokens. Subcorpora have dependency grammar annotation based on the Universal Dependencies standard (de Marneffe et al., 2021). mini-CIEP+ version 1.0 is available in 35 languages, with the aim of increasing the sample to 50 languages. It is shareable due to recent developments in German law, which allow researchers to share up to 15% of copy-righted material with a select group of people for their own research. Hence, mini-CIEP+ is not publically available, but is rather shareable in a modular fashion with select researchers. We additionally describe future plans for further annotation of mini-CIEP+ as well as its limitations.

Keywords: parallel corpus, linguistic typology, copyright

1. Introduction

Linguistic typology, the systematic comparison of language structure across large samples of languages, has traditionally relied on discrete classifications, created by human specialists. Increasingly, however, typologists are using multilingual corpora instead: a collection of utterances (a corpus) is investigated directly using frequency-based or information theoretic measures, yielding continuous measures of language structure that are considerate of variation and sometimes, diachronic change. This approach is sometimes called token-based typology (Levshina, 2016) or corpus-based typology (Levshina, 2022; Schnell and Schiborr, 2022).

This line of work inevitably relies on the availability of cross-linguistic corpora. While many of these have emerged in the last 25 or so years (Tiedemann, 2012; Moran et al., 2022; Rosen et al., 2022; the TenTen corpora) there are distinct biases towards legalese and religious texts; and material gathered outside of those two genres often constitute (web-crawled) text fragments or collections of sentences (such as Tatoeba or the Leipzig Corpora Collection). Register is an important consideration for corpus-based typology, as we know from the study of well-described languages like English that register differences can be immense (Biber, 2012). Doing corpus-based typology solely on legal texts, web-crawled news and the Bible is at best unrepresentative of linguistic diversity.

Here we present mini-CIEP+, a sharable parallel corpus consisting of the first part of ten different works of prose. mini-CIEP+ contains subcorpora in 35 languages in version 1.0 (we aim to include 50 languages until 2028) and is annotated in the Universal Dependencies (UD) standard (de Marneffe et al., 2021). Note that while this is ongoing work, mini-CIEP+ is the first of its kind: 1) it allows for the linguistic investigation of larger discourse units (in contrast to many other web-crawled corpora); 2) the parallel nature of mini-CIEP+ has the advantage that direct comparison of subcorpora is straightforward

and that annotation projection is possible (see Section 7); 3) there are no other prose corpora with this scale or size; and 4) since it contains published prose, there are no issues with variable or poor quality of the material. Given that the works of prose have copyright, we cannot make mini-CIEP+ publicly available; however, recent changes in German law allow us to share it with other researchers. In this paper, we describe the shareable corpus as well as design and implementation choices. Corpus composition and annotation are described in Sections 3 and 4, after the overview on previous work.

2. Previous work

Since the early 2000s, several (large) parallel corpora have emerged: *EuroParl* (Koehn, 2005), *ParaSol* (Slavic prose and beyond, Waldenfels, 2006), the *Parallel Bible Corpus* (Mayer and Cysouw, 2014), *OpenSubtitles* (Lison and Tiedemann, 2016), *ParTy* (movie subtitles, Levshina, 2017), *MULTEXT-East* (Erjavec, 2017), *JW300* (Jehova Witness magazines, Agić and Vulić, 2019) and *ParlaMint* (parliamentary proceedings, Erjavec et al., 2023). Several of these have been compiled in *OPUS* (Tiedemann, 2012).

While these corpora contain texts from a variety of genres, most importantly legal and religious, there is a distinct lack of prose corpora, for the obvious reason that widely translated prose is typically protected under copyright law and cannot be publicly shared. Hence, the corpora used by Stolz and colleagues (for example, Stolz and Gugeler, 2000) are not publicly available and *ParaSol* (Waldenfels, 2006) can be used online but cannot be downloaded; the only exception here is *MULTEXT-East*, a parallel and morpho-syntactically annotated corpus of Orwell's *1984* in 16 languages, which is fully downloadable from the CLARIN repository¹.

Given recent changes to German and EU copyright law, some solutions for this problem have emerged. Schöch et al. (2020) propose preparing derived texts, similar in a way to datasets such as the Google Ngram Viewer² or the HathiTrust Research Center Extracted Features Dataset.³ However, such datasets where

¹ <http://hdl.handle.net/11356/1043>

² <https://books.google.com/ngrams/>

³ <https://analytics.hathitrust.org/datasets/>

only frequency information or information regarding lemmas is available, but not their sequence, are not sufficient for answering many linguistic questions. Gärtner et al. (2021) propose an automated sampling approach, where users have access to 15% of individual copy-righted works (see Section 5). The downside of this approach is that samples are taken from the entirety of the text, so discourse units beyond sentences are not preserved and cannot be investigated. Bański et al. (2017) propose to make use of the long scientific quotation clause in German copyright law, arguing that a compilation of long text segments with newly created annotation constitutes a new, original work. In this case, the corpus creator enters a legal gray zone: how much annotation needs to be added in order for the corpus to be conceived as a new work?

We created mini-CIEP+ to overcome several of these problems; the legal solution is explained in Section 5. First, we describe the corpus in greater detail.

3. Corpus sample and composition

mini-CIEP+ contains a subset of the material of the Corpus of Indo-European Prose Plus (CIEP+, /ki:p plʌs/, see Talamo and Verkerk, 2022). This work-in-progress corpus will contain up to 18 literary works in 50 languages, with a bias towards Indo-European. mini-CIEP+ contains about 14% of 10 of these literary works (see Section 5) in the same languages:

1. IE, Albanian: Standard Albanian
2. IE, Armenian: Eastern Armenian
3. IE, Baltic: Latvian, Lithuanian
4. IE, Celtic: Breton, Irish, Welsh
5. IE, Germanic: Afrikaans, Danish, Dutch, English, German, Swedish
6. IE, Hellenic: Modern Greek
7. IE, Indo-Aryan: Assamese, Bengali, Hindi, Marathi, Nepali, Punjabi, Sinhala, Urdu
8. IE, Iranian: Kurdish, Persian
9. IE, Romance: French, Latin, Italian, Portuguese, Romanian, Spanish
10. IE, Slavic: Bulgarian, Czech, Polish, Russian, Serbo-Croatian, Ukrainian
11. Austronesian: Hawaiian, Indonesian, Maori
12. Bantu: Swahili
13. Basque
14. Dravidian: Tamil
15. Japonic: Japanese
16. Kartvelian: Georgian
17. Koreanic: Korean
18. Semitic: Arabic
19. Sinitic: Mandarin Chinese
20. Turkic: Turkish
21. Uralic: Finnish, Hungarian

⁴ https://en.wikipedia.org/wiki/List_of_literary_works_by_number_of_translations

⁵ Another concern might be how modern the corpus is, given that AAiW and TtLG are from the late nineteenth century, and we have several books from the 1940s and 1980s. However, all of these are considered modern classics and many translations we have obtained are far more recent than these first dates of publication betray.

Given that the translation of prose is driven by monetary impetuses, the mini-CIEP+ language sample is biased towards European and other well-described languages (see Wälchli, 2007). The prose works chosen have been selected first for their popularity, i.e. because they have been widely translated, and second, for being originally written in different languages, so as to avoid English as the sole source language. We are aware that the original texts are written exclusively in Indo-European languages, more specifically, in French, Italian, Spanish and Portuguese (Romance), Dutch, German and English (Germanic) and Modern Greek (Hellenic). Sadly, this bias cannot be avoided; out of the titles listed under the Wikipedia entry 'List of literary works by number of translations'⁴, there are about 80 books that can be loosely classified as 'prose', namely, novels, diaries and plays; however, the majority are originally written in languages from the three above-mentioned branches, especially English. Other works of prose that could be considered come with certain difficulties. Children's stories such as *Pinocchio* often suffer from abridged translations. Books not originally written in one of the languages mentioned above are few; those that exist, such as *The Upright Revolution: Or Why Humans Walk Upright* (by Ngūgĩ wa Thiong'o), are either too short, not modern (*The tragedy of Man*, by Imre Madách), or very hard to obtain (such as Ismail Kadare's *The General of the Dead Army*).⁵

Given these considerations, mini-CIEP+ contains the first part of the following ten texts.⁶ A list of authors, titles, and date of first publication is provided here for brevity; an overview of mini-CIEP+ is available in Table 1. Acronyms refer to columns in that Table.

1. **AA** – Carroll's *Alice's Adventures in Wonderland* [English, 1865]
2. **LG** – Carroll's *Through the Looking-Glass and What Alice Found There* [English, 1871]
3. **AI** – Coelho's *O Alquimista* [The Alchemist, Portuguese, 1989]
4. **Za** – Coelho's *O Zahir* [The Zahir, Portuguese, 2005]
5. **Ro** – Eco's *Il nome della rosa* [The Name of the Rose, Italian, 1980]
6. **Di** – Anne Frank's *Het Achterhuis* [Diary of a Young Girl, Dutch, 1947]⁷
7. **100Y** – García Márquez's *Cien Años de Soledad* [One Hundred Years of Solitude, Spanish, 1967]
8. **Zo** – Kazantzakis' *Βίος και Πολιτεία του Αλέξη Ζορμπά* [Zorba the Greek, Modern Greek, 1946]
9. **Pr** – de Saint-Exupéry's *Le Petit Prince* [The Little Prince, French, 1943]
10. **Pa** – Süskind's *Das Parfum. Die Geschichte eines Mörders* [Perfume: The Story of a Murderer, German, 1985]

⁶ All originals are included. When selecting the translations, we aim for the most recent one or one which has been translated directly from the original (non-mediated).

⁷ Of course, Anne Frank's *Het Achterhuis* is not a work of fiction. We include it because it is the most widely translated Dutch original text, and because in terms of its register, it is not far from the other included texts. Diary entries can be considered stories told from a first-person perspective.

Family, genus	Language	100Y	AA	Di	Al	Ro	Pa	Pr	LG	Za	Zo	T	UD
IE Celtic	Welsh	-	1	1	-	-	-	1	1	-	-	4	p
IE Celtic	Irish	-	1	-	1	-	-	1	1	-	-	4	p
IE Indo-Aryan	Urdu	-	-	-	1	-	-	1	-	1	1	4	p
IE Romance	Latin	-	1	-	-	-	1	1	1	-	-	4	p
IE Germanic	Afrikaans	-	1	1	1	-	-	1	-	-	1	5	p
Dravidian	Tamil	1	1	1	1	-	-	1	-	1	-	6	p
IE Indo-Aryan	Marathi	1	1	1	1	-	-	1	-	1	-	6	p
Basque	Basque	-	1	1	1	-	1	1	1	-	1	7	p
IE Armenian	Armenian	1	1	1	1	1	1	1	1	-	1	9	p
IE Indo-Aryan	Hindi	1	1	1	1	1	-	1	1	1	1	9	p
Austronesian	Indonesian	1	1	1	1	1	1	1	1	1	-	9	p
IE Hellenic	Modern Greek	1	1	1	1	1	1	1	1	1	1	10	p
IE Baltic	Latvian	1	1	1	1	1	1	1	1	1	1	10	p
IE Baltic	Lithuanian	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	Swedish	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	Danish	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	Dutch	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	English	1	1	1	1	1	1	1	1	1	1	10	p
IE Germanic	German	1	1	1	1	1	1	1	1	1	1	10	p
IE Iranian	Persian	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	Portuguese	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	French	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	Italian	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	Romanian	1	1	1	1	1	1	1	1	1	1	10	p
IE Romance	Spanish	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Bulgarian	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Serbo-Croatian	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Czech	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Polish	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Russian	1	1	1	1	1	1	1	1	1	1	10	p
IE Slavic	Ukrainian	1	1	1	1	1	1	1	1	1	1	10	p
Uralic	Finnish	1	1	1	1	1	1	1	1	1	1	10	p
Uralic	Hungarian	1	1	1	1	1	1	1	1	1	1	10	p
Japonic	Japanese	1	1	1	1	1	1	1	1	1	1	10	p
Semitic	Arabic	1	1	1	1	1	1	1	1	1	1	10	p
Sinitic	Man. Chinese	1	1	1	1	1	1	1	1	1	1	10	p
Turkic	Turkish	1	1	1	1	1	1	1	1	1	1	10	p
Koreanic	Korean	1	1	1	1	1	1	1	1	1	1	10	p
IE Celtic	Breton	-	1	-	-	-	-	1	-	-	-	2	t
IE Indo-Aryan	Assamese	-	1	-	1	-	-	-	1	-	-	3	n
IE Indo-Aryan	Nepali	-	1	-	1	-	-	1	-	-	-	3	n
Austronesian	Maori	-	1	1	1	-	-	-	-	-	-	3	n
Austronesian	Hawaiian	-	1	-	-	-	-	1	1	-	-	3	n
Bantu	Swahili	-	1	-	1	-	-	1	-	-	-	3	n
IE Iranian	Kurdish	1	1	-	1	-	-	1	-	-	-	4	t
IE Indo-Aryan	Sinhala	-	-	1	1	-	1	1	-	1	-	5	t
IE Indo-Aryan	Bengali	1	1	1	1	-	-	1	-	1	-	6	n
IE Indo-Aryan	Punjabi	1	-	1	1	-	-	1	-	1	1	6	n
IE Albanian	Albanian	1	1	1	1	1	1	1	1	1	1	10	t
Kartvelian	Georgian	1	1	1	1	1	1	1	1	1	1	10	n

Table 1: Overview of literary works available per language in mini-CIEP+. The last column, "UD", specifies relevant information regarding UD (Universal Dependencies) version 2.13: p (pre-trained model available in Stanza (Qi et al., 2021)), t (treebank available without pre-trained model) and n (no UD treebank available). The languages printed in bold are included in mini-CIEP+ version 1.0.

If all ten books are available, the size of the subcorpus for a single language is approximately 121,000 tokens, or 5750 sentences. The size of each subcorpus is provided in Table 2, in terms of both tokens and sentences – two statistics on key UD dependency labels are also given. However, note that not all ten books are available in all fifty languages (see Table 1). Most or all works of prose are available in most languages, but for some languages only four or fewer are available. In order to have approximately equal subcorpora sizes, we add more prose works to a subcorpus such as that of Irish, which only contains four out of the ten prose works listed above.⁸ Hence, with the addition of two translated works and four native Irish works, the Irish subcorpus has become a comparable rather than a parallel subcorpus – in the sense that the added texts are translated and original prose. In these cases, we aim to obtain at least the English translations or originals, so the paired subcorpora can be used for contrastive analyses.

4. Corpus processing and annotation

The CIEP+ corpus exists both physically and digitally. The first step to obtain the relevant textual material for each subcorpus is to obtain a physical copy of each book (see Section 5) and create or buy in addition a digital version. In most cases, the digital version is created by scanning the book and applying Optical Character Recognition (OCR) to retrieve the contents in plain text format. The result has to be checked and corrected by human annotators, as automated OCR usually generates a lot of mistakes.

Then, the texts that are included in each subcorpus are annotated with metadata for the following information: original author, original title, original publishing date, original language, translator, translation language, translation title, translation date and translation publishing house. The physical books are cataloged in the university library (SULB).

It is not feasible to provide morphosyntactic annotation of such a large and diverse data set by hand. Hence, the first layers of annotation are added automatically. We have chosen to do this within the Universal Dependencies (UD) framework (de Marneffe et al., 2021), for several reasons. Firstly, UD's aim of providing consistent annotation of morphosyntax (including parts of speech, morphology, and syntactic dependencies) across different languages aligns with our own: we need consistent morphosyntactic annotation in order to use the data to ultimately answer typological research questions. The Universal Dependencies project is emerging as the go-to set of treebanks for typologists, given its wide sample of parsed language data, which we (and others) use not only for doing typology on,

but also for training tools that can automatically parse new language data. Secondly, dependency grammar is central to our goals in the larger project, given that we are interested in dependency length optimization and other functional metrics of language-in-use (see Dyer, 2023). Thirdly, given the status of UD as emerging standard of the field implies that there exist a lot of (also future) resources that allow us to parse additional languages (see below), but that also allow prospective users of mini-CIEP+ to convert it to formats of their choice.

The tool chosen to process corrected texts and create automated annotation in the UD standard is the Stanford Stanza natural language analysis package⁹ (Qi et al., 2021). Among the 24 systems participating in the CoNLL 2018 Shared Task (Zeman et al., 2018), Stanford Stanza ranked eight in the labeled attachment score (LAS), second in the Morphology-Aware Labeled Attachment Score (MLAS) and fifth in the Bilexical Dependency Score (BLEX); to the best of our knowledge, only two systems that performed slightly better than Stanza are currently available to the community, UDPipe Future¹⁰ (Straka, 2018; now UDPipe 2) and Turku NLP¹¹ (Kanerva et al., 2018). However, in the CoNLL 2018 Shared Task systems were evaluated on Universal Dependencies treebanks, which widely differ from mini-CIEP+ data in terms of register. We leave for future work a shared task performed on mini-CIEP+ data, comparing Stanford Stanza to other available systems.

At the time of writing, Stanza comes with 138 models, which are pretrained on Universal Dependencies version 2.13 treebanks and cover 38 languages of the sample. These models are used to parse the corrected texts, processing and annotating them in several steps, including sentence splitting, tokenization, lemmatization, parts-of-speech and syntactic dependencies tagging, and, where available, multi-word token expansion and named entity recognition.

This leaves twelve languages without pre-trained Stanza models (see also Table 1). As for some of these low resource languages, we have used small existing Universal Dependencies treebanks to train parsers for three languages, namely for Breton, Kurdish, and Sinhala (results are not included in mini-CIEP+ v. 1.0, but will be in later versions). While we have not formally evaluated these so far, results very much depend on the size (and register) of the UD treebank.

This leaves nine languages in our sample with no or highly limited Universal Dependencies resources (see Table 1). We ourselves started projects to provide resources for two of the low resource languages –

⁸ For Irish, we have added six texts in order to try to approach a similar token size as the other subcorpora:

1. *An Béal Bocht* (The Poor Mouth), Flann O'Brien
2. *An Hobad, nó Anonn agus Ar Ais Arís* (The Hobbit, or There and Back Again), J. R. R. Tolkien
3. *An Leon, an Bandraoi agus an Prios Éadaigh* (The Lion, the Witch and the Wardrobe), C. S. Lewis

4. *Buille Marfach* (A Fatal Blow), Anna Heussaff
5. *Cré na Cille* (Graveyard Clay), Máirtín Ó Cadhain
6. *Rún an Bhonnáin* (The secret of the Bonnán), Proinsias Mac a' Bhaird

⁹ <https://stanfordnlp.github.io/stanza/>

¹⁰ <https://github.com/ufal/udpipe/releases/tag/v2.1.0>

¹¹ <https://turkunlp.org/Turku-neural-parser-pipeline/>

Language	Bks	Token	Sent.	nsubj	obj	Language	Bks	Token	Sent.	nsubj	obj
Albanian	10	135158	6401	8493	10659	Latin	3	9003	670	718	610
Arabic	10	123994	NA	8649	5689	Latvian	10	105635	6234	10023	7506
Armenian	6	68696	3503	4785	4030	Lithuanian	10	105226	6800	7964	4287
Basque	4	19870	1244	1013	1461	Man. Chinese	10	136777	6064	13038	9824
Bulgarian	10	118040	6369	6997	9742	Marathi	8	105197	5990	9208	7629
Czech	10	114149	6263	6868	5555	Persian	10	131749	6039	7058	5316
Danish	10	133082	6250	13478	8772	Polish	10	116429	6228	6011	8820
Dutch	10	133933	6243	12584	6710	Portuguese	10	135648	6281	6903	8400
English	10	138386	6472	12802	6794	Romanian	10	131484	5668	7051	7862
Finnish	5	43335	3278	4067	2529	Russian	10	117115	6245	9868	5868
French	10	144199	6365	11904	9267	Serb.-Croatian	10	115582	5888	7270	7475
German	10	130730	6139	12232	7308	Spanish	10	130947	5731	6113	7633
Mod. Greek	10	125972	5393	6601	6132	Swedish	8	97054	4468	10299	5670
Hindi	8	95667	4536	9147	5532	Turkish	10	94958	5633	6647	7214
Hungarian	10	110675	5812	7378	6804	Ukrainian	10	109248	5757	9085	6721
Indonesian	9	104799	5089	9694	6581	Urdu	4	38217	2368	3456	2243
Irish	10	56920	3165	4896	1981	Welsh	4	33611	1494	2333	1267
Italian	10	137672	6168	6485	7379						

Table 2: Overview of descriptive statistics of mini-CIEP+ version 1.0. Bks = Books; Token = Tokens; Sent = Sentences. nsubj and obj refer to the number of constituents with these labels in each parsed subcorpus. Some subcorpora still lack some texts that have to be processed (see Section 4), which will be part of mini-CIEP+ version 1.1. Further languages listed in Section 3 and Table 1 will be included in future versions.

these projects take the form of manually annotated UD treebanks covering literary works originally written in Albanian (Talamo, in prep.) and Bengali (Dyer, in prep. b). These treebanks are used to train good quality parsers, specifically aimed to the genre featured in our parallel corpus, and allow for automated parsing of the Albanian and Bengali subcorpora. Although others are similarly spearheading solutions for the lack of resources in several languages, this will remain problematic in years to come. This means that seven languages of our sample (Assamese, Georgian, Hawaiian, Maori, Nepali, Punjabi, Swahili) do not currently have any existing UD treebanks; for these languages, we wait for relevant UD treebanks to become available, or find alternative solutions. Such solutions will include zero-shot analysis alongside corrections, for example using UDify (Kondratyuk and Straka, 2019), and converting existing treebanks to the UD standard.

UD's native CoNLL-U format allows for additional annotation in the last column, and the newer CoNLL-U Plus format allows for even more columns. We aim to release versions of mini-CIEP+ with surprisal and information status annotation (see Section 7). For users of mini-CIEP+, these columns can be used for

other types of annotation. The modular nature of the corpus also allows for re-parsing with better models and human correction of automated annotation.

5. Sharing the corpus

Given its size and its cost in terms of resources, we did not wish to create CIEP+ (the Corpus of Indo-European Prose Plus) only for project internal purposes (see also Hartmann's 2023 proposal on Open Corpus Linguistics). German copyright law has changed in 2018 regarding two important aspects: collecting copyrighted material for research and sharing it with a select group of people. The relevant articles are Urheberrecht § 60c and 60d.¹² Under German law, we are allowed to store digital copies of copy-righted works and use these for research if we own the physical books. Then, most relevant for mini-CIEP+ is the following sentence; original German in the footnote below:

*"For the purpose of non-commercial scientific research, up to 15 percent of a work may be reproduced, distributed and made publicly accessible [...] to a defined circle of people for their own scientific research"*¹³

¹² https://www.gesetze-im-internet.de/urhg/_60c.html
https://www.gesetze-im-internet.de/urhg/_60d.html

¹³ "Zum Zweck der nicht kommerziellen wissenschaftlichen Forschung dürfen bis zu 15 Prozent eines Werkes vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden

Hence, we created mini-CIEP+ to legally share 15% of CIEP+ with a specifically designated group of people in order to benefit their research. We believe that there is indeed a pluricentric group of people that would benefit from mini-CIEP+: corpus-based typologists, but also contrastive linguists and language specialists. As we include several low-resource languages, it is our hope that parts of mini-CIEP+ can be used for furthering research into those languages.

To make this possible we have created a data usage agreement (see Appendix A) that specifies the conditions under which mini-CIEP+ can be provided to potential data users. This data usage agreement also asks which subcorpora are needed by the researcher, so that the corpus is really only shared to the extent required. Version management takes place via the author's home page,¹⁴ so that prospective data users know what is available for sharing.

6. CIEP+-based works so far

CIEP+ (the Corpus of Indo-European Prose Plus) is being built in the context of a large research program,¹⁵ in which our team have authored half a dozen of papers in the last four years. In this section, we give an overview of these papers in order to showcase what type of linguistic research can be done on such a resource. As primarily a resource for typologists, CIEP+ was first exploited for addressing one of the oldest topics in linguistic typology, namely, word order variation. Talamo and Verkerk (2022) investigated the order of constituents in five nominal constructions (the order of article, demonstrative, adjective, adposition and relative clause with respect to the noun) in a sample of 11 Indo-European languages, using Shannon's entropy as a metric for word order variability. The results show the high unpredictability of the position of adjectives in Romance and Slavic languages, while the entropy of constructions like determiners and adpositions is generally low. The latter confirms the traditional view of categorical studies; however, there are in fact outliers, as we retrieve phenomena that create variability in the position of prepositions in Dutch and find a certain degree of freedom for demonstratives in Greek, Polish and Welsh.

Talamo (2023) has further expanded the research regarding word order variability within the noun phrase by looking at neglected and hard-to-catch categories such as quantifiers, determiners and numerals; in a sample of 17 languages, Talamo (2023) finds that the variability of demonstratives is found in another Balkan language, Romanian, and reports on the high variability of quantifiers in Irish.

In the field of historical linguistics, Talamo et al. (2024) used CIEP+ to challenge the traditional view which states that subordinate clauses tend to preserve more conservative features than main clauses. Focusing on

adverbial clauses and using frequency data on null subject pronouns and order of subject, object and verb in a sample of 30 Indo-European languages, they show that there are actually very few asymmetries between adverbial and main clauses, both in the synchronic data and during language change, which is modelled using phylogenetic methods.

The prose genre of CIEP+, which is characterized by several dialogues mimicking the spoken language, allows for research into linguistic devices used for reference. In an ongoing study (Steuer et al. in prep.), we are exploring the relations between personal pronouns and their referents, trying to understand how the former encode the information status of the latter. We model the probability in context (surprisal) of personal pronouns in a sample of 15 languages from eight different families using mGPT (Shiliazhko et al. 2022). We expect that these models reflect varying surprisal of personal pronouns based on their frequency and usage patterns, showing that first and second personal pronouns encode less information than third personal pronouns.

Several of these studies, including Talamo et al. (2024) and Levshina et al. (2023), contain comparisons between CIEP+ and UD treebanks. We can confirm that automatically parsed data from CIEP+ behaves similarly (i.e. is correlated with) data from Universal Dependencies treebanks on several measures, including word order variation and pronoun usage. However, there are notable differences between the two data sources, especially concerning individual languages on certain measures. We leave for future work a systematic comparison of CIEP+ and mini-CIEP+ with UD treebanks, with the specific aim of investigating if such differences are rooted in register differences, problems with automated parsing, or inconsistencies in UD annotation across languages.

7. Future plans

Currently, mini-CIEP+ is automatically annotated using the UD framework (de Marneffe et al., 2021, see above) in the same way as CIEP+. However, as mentioned above, we aim to add several types of annotation to mini-CIEP+, which can be shared in future versions. One type of annotation that we aim to add to CIEP+ and mini-CIEP+ is sentence and word alignment. This is obviously a great asset for a parallel corpus, however, performance on automated alignment will vary radically from language pair to language pair. While the pivot language will be English, we will carry out experiments to see if automated sentence alignment can be improved by employing different or even multiple pivots. Alignment is necessary in order to be able to project different types of annotation across the subcorpora. We will focus on information status annotation. Ongoing work (Dyer in prep. a) is preparing information status

[...] für einen bestimmt abgegrenzten Kreis von Personen für deren eigene wissenschaftliche Forschung"

¹⁴ <https://www.uni-saarland.de/lehrstuhl/verkerk.html>

¹⁵ <https://sfb1102.uni-saarland.de>

annotation using human annotators for English, modern Greek, Indonesian, Turkish, and Ukrainian. This version of mini-CIEP+ can also be shared with researchers interested in such annotation.

If data users require us to do so, it is possible to add more languages to the sample, especially for *Alice's Adventures in Wonderland* and *Le Petit Prince*, as these are the corpus' most widely translated books.

8. Conclusion and limitations

We have presented mini-CIEP+, a sharable parallel corpus of prose. We have described its compilation, composition, size, annotation, and plans on how to share it with relevant researchers. This is the first version and more versions are planned for the future.

We conceive of mini-CIEP+ as a modular resource for corpus-based typologists, contrastive linguists and language specialists. Individual subcorpora may not be large (~121,000 tokens), but they are large enough to research a plethora of linguistic phenomena, including semantic and pragmatic features that emerge only in the analysis of bigger discourse units. We hope that mini-CIEP+ will be used and expanded, if so, we will do our best to expand it further in a way that benefits the scientific community. Including other books for individual subcorpora would be possible.

One limitation we cannot fix is the inherent bias in the sample of languages. mini-CIEP+ is a derivative of CIEP+ (the Corpus of Indo-European Prose Plus); the inclusion of mostly Indo-European languages is intentional but at the same time, a regrettable continuation from similar biases in other corpora. Aside from *Le Petit Prince* and, to a lesser extent, *Alice's Adventures in Wonderland*, the corpus' set of prose texts (indeed, published prose in general) tends to be translated in only a very small subset of the world's languages. A positive outlook on this is offered by the larger amount of variety included in the Universal Dependencies treebanks, and in other projects such as TeDDi (Moran et al., 2022). A worthwhile solution is for corpus-based typologists to find ways to be able to analyze heterogeneous data sources, possibly with the help of NLP tools. These will not always have the same register, annotation, size, or even script, but combining (still scarce) resources on the languages of the world will be essential in future ventures in quantitative typology.¹⁶

9. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

10. Bibliographical References

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics, pages 3204–3210. Florence. Association for Computational Linguistics.

Bański, P., Kamocki, P., and Trawiński, S. (2017). Legal canvas for a patchwork of multilingual quotations: the case of CoMPaRS. Presented at the Corpus Linguistics International Conference 2017, Birmingham.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1): 9–37.

Dyer, A. (2023). Revisiting dependency length and intervener complexity minimisation on a parallel corpus in 35 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 110–119.

Dyer, A. (in preparation a). Cieplnf: A multilingual parallel corpus for coreference resolution and information status in the literary domain. [working title]

Dyer, A. (in preparation b). The Shobdokosh dependency corpus of Bengali prose. [working title]

Erjavec, T. (2017). MULTEXT-East. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation*. Dordrecht: Springer, pp. 441-462.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Calzada Pérez, M., de Macedo, L.D. Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Dargis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation* 57(1): 415-448.

Gärtner, M., Kleinkopf, F., Andresen, M., and Kupietz, M. (2021). Corpus reusability and copyright – challenges and opportunities. In H. Lungen, M. Kupietz, P. Bański, A. Barbaresi, S. Clematide, & I. Pisetta (Eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021*, pages 10–19. Leibniz-Institut für Deutsche Sprache.

Hartmann, S. (2023). Open corpus linguistics – or how to overcome common problems in dealing with corpus data by adopting open research practices. Preprint. PsyArXiv.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In D. Zeman & J. Hajič (Eds.) *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133-142. Association for Computational Linguistics.

¹⁶ Author contributions. AV: conceptualization; validation; data collecting; writing (original draft) sections: 1, 2, 3, 4, 6, 7, 8, Appendix; writing (review & editing); LT: validation;

data collecting; data parsing; writing (original draft) sections: 3, 4, 5; writing (review & editing).

- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*, pages 79-86. Phuket, Thailand.
- Kondratyuk, D., and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally'. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779-2795. Hong Kong. Association for Computational Linguistics.
- Levshina, N. (2016). Why we need a token-based typology: a case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica* 50(2): 507–542.
- Levshina, N. (2017). A multivariate study of T/V forms in European languages based on a parallel corpus of film subtitles. *Research in Language* 15(2): 153–172.
- Levshina, N. (2022). Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(2): 129-160.
- Levshina, N., Namboodiripad, S., Allasonnière-Tang, M., Kramer, M.A., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gupton, T., Kidd, E., Liu, Z., Naccarato, C., Nordlinger, R., Panova, A., and Stoyanova, N. (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4): 825-883.
- Lison, P., and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA).
- de Marneffe, M.-C., Manning, C.D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics* 47(2): 255–308.
- Mayer, T., and Cysouw, M. (2014). Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik. European Language Resources Association (ELRA).
- Moran, S., Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and Samardzic, T. (2022). 'TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1150–1158. Marseille. European Language Resources Association (ELRA).
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C.D. (2020). 'Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Online. Association for Computational Linguistics.
- Rosen, A., Vavřín, M., and Zaslina, A.J. (2022) *InterCorp*, Release 15 of 11 November 2022. Institute of the Czech National Corpus, Charles University. Available from: <http://www.korpus.cz>.
- Schnell, S. and Schiborr, N.N. (2022). Crosslinguistic corpus studies in linguistic typology'. *Annual Review of Linguistics* 8(1): 171–191.
- Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., and Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*.
- Shliakhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., and Shavrina, T. (2022). mGPT: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.
- Steuer, J., Talamo, L., and Verkerk, A. (in preparation). Measuring accessibility through surprisal: a cross-linguistic study of personal pronouns. [working title]
- Stolz, T., and Giugeler, T. (2000). Comitative typology – nothing about the ape, but something about king-size samples, the European community and the little prince. *STUF - Sprachtypologie und Universalienforschung* 53(1): 53–61.
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197-207. Brussels. Association for Computational Linguistics.
- Talamo, L. (2023). Using a parallel corpus to study patterns of word order variation: Determiners and quantifiers within the noun phrase in European languages. *Linguistic Typology at the Crossroads* 3(2): 100-131.
- Talamo, L. (in preparation). STAF: The Saarbrücken Treebank of Albanian Fiction. [working title]
- Talamo, L. and Verkerk, A. (2022). A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics*, 34(2), 171–226.
- Talamo, L., Verkerk, A., and Salaberry, I. (2024). A quantitative approach to clause type and syntactic change in two Indo-European corpora. *Italian Journal of Linguistics*, 36.
- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214-2218. Istanbul. European Language Resources Association (ELRA).

von Waldenfels, Ruprecht. (2006). Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment'. In B. Brehmer, V. Zdanova, and R. Zimny (Eds.) *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9. München: Otto Sagner, pp. 123–138

Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *STUF - Sprachtypologie und Universalienforschung* 60(2): 118–134.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018) CoNLL 2018 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, pages 1-21. Brussels. Association for Computational Linguistics.

Appendix A. Draft of the DATA USAGE AGREEMENT FOR THE SHAREABLE PORTION OF THE CORPUS OF INDO-EUROPEAN PROSE PLUS (mini-CIEP+)

mini-CIEP+ is provided by the Corpus Provider, see below, to the Data User, as signed below, under the following terms:

1. mini-CIEP+ may only be used for non-commercial linguistic research or education.
2. Usage of mini-CIEP+ is granted to individual Data Users only. All prospective Data Users of mini-CIEP+ must fill out this data usage agreement individually.
3. The Data User agrees that they will not attempt to use mini-CIEP+ to infringe on the rights of the original copyright holders; i.e. the authors/publishers of the literary works that are part of mini-CIEP+.
4. The Data User certifies that their copy of mini-CIEP+ is stored only in a single copy on computers under administration of the Data User. Data User certifies that they will take proper action for protecting this copy from being accessed, read or copied by any non-authorized person.
5. The Data User agrees to delete mini-CIEP+ after twelve months signing this agreement. The Corpus Provider must be informed that deletion of the corpus by the Data User has been done. An extension of data usage is possible by signing this agreement again.
6. mini-CIEP+ is provided free of charge.
7. mini-CIEP+ comes with absolutely no warranties including (but not limited to) the correctness of the information provided in the text corpus itself.
8. The Data User will not disclose, disseminate, or otherwise share mini-CIEP+ to or with any other person or entity, for any purpose. The Data User has no right to copy, redistribute, transmit, publish or otherwise use mini-CIEP+ for any other purpose.
9. mini-CIEP+ must not be transmitted electronically to other services not under administration of the Data User, such as online translation services.
10. The Data User may include limited excerpts from mini-CIEP+ in articles, reports and other documents describing the results of the Data User's non-commercial linguistic education or research.
11. In no event shall the Corpus Provider be liable to the Data User for direct, indirect, special, incidental, punitive or consequential damages of any kind arising in any way out of this agreement, rights granted herein or by the use of mini-CIEP+.
12. mini-CIEP+, in all forms, shall be and remain the responsibility of the Corpus Provider.
13. The Data User will provide the Corpus Provider with a short summary (less than 100 words, see below) describing the purpose of their research based on mini-CIEP+ and the language sample they require. The Data User agrees that all their actual research activities with mini-CIEP+ will adhere to this description. Using mini-CIEP+ for a different kind of research requires signing a new data usage agreement with a new description.
14. The Data User agrees that their name, contact information, and the research summary are stored in electronic form by the Corpus Provider. This information will be used to (a) inform Data Users when updates of mini-CIEP+ are available and to (b) create anonymized corpus distribution statistics. Additionally, the information might be used to track violations of this agreement. It will be deleted once this Data Usage Agreement is expired or cancelled by the Data User or by the Corpus Provider.
15. Contributions which are based on mini-CIEP+ must cite the following publication: <xxx>
16. Contributions which are based on mini-CIEP+ must correctly cite its version as well as the original works compiled in mini-CIEP+, which can be retrieved from mini-CIEP+'s metadata.
17. The Data User shall email an electronic version of the signed agreement to the Corpus Provider.

Author Index

- Amin, Mohammad Ruhul, 69
Avramidis, Eleftherios, 51
- Banski, Piotr, 94
Branco, António, 24
- Calvary, Gaelle, 85
Cesur, Neslihan, 104
Chen, Jian, 36
Chowdhury, Mufassir Ahmad, 69
- Dangovski, Rumen, 1
Degaetano-Ortlieb, Stefania, 12
Diewald, Nils, 94
- Fischer, Stefan, 12
Fu, Xianghua, 36
- Gaussier, Eric, 85
Gomes, Luís, 24
Grouin, Cyril, 59
- Hannani, Mohamed, 51
Hou, Shilong, 36
- Kar, Sudipta, 69
Klakow, Dietrich, 12
Kose, Mehmet, 104
Krielke, Marie-Pauline, 12
Kupietz, Marc, 94
Kuzgun, Asli, 104
- Laskina, Anna, 85
Leite, Bernardo, 24
Lerner, Paul, 59
Long, Zi, 36
Lopes Cardoso, Henrique, 24
Lyu, Jinze, 36
- Masciolini, Arianna, 111
Mendes, Amalia, 125
Mobin, Md. Shakirul Hasan Khan, 69
Morger, Felix, 118
Mosbach, Marius, 12
- Nakov, Preslav, 1
- Nobin, Zeshan Ahmed, 69
- Osório, Tomás Freitas, 24
- Ramírez, Guillem, 1
Rodrigues, João, 24
- Saha, Sourav, 69
Santos, Rodrigo, 24
Soljagic, Marin, 1
Soudi, Abdelhadi, 51
Steuer, Julius, 12
- Talamo, Luigi, 135
Tang, ZhenHao, 36
Trawinski, Beata, 94
- Valūnaitė Oleškevičienė, Giedrė, 125
Van Laerhoven, Kristof, 51
Verkerk, Annemarie, 135
- Witt, Andreas, 94
- Yıldız, Olcay Taner, 104
Yvon, François, 35
- Zeyrek, Deniz, 125