A Representation Learning Based Approach to the Study of Translationese

Koel Dutta Chowdhury

A dissertation submitted towards the degree
PhD in Computational Linguistics
from the Department of Language Science and Technology
Saarland University

Saarbrücken, 2024

Koel Dutta Chowdhury: *A Representation Learning Based Approach to the Study of Translationese* , a dissertation submitted in fulfillment of the requirements of the degree Ph.D. in Computational Linguistics from the Department of Language Science and Technology at Saarland University.

DEAN OF THE FACULTY: Prof. Dr. Stefanie Haberzettl
Chair – Prof. Dr. Ingo Reich, Saarland University

Advisor, Reviewer – Prof. Dr. Josef van Genabith, Saarland University
Reviewer – Prof. Dr. Antonio Toral, University of Groningen
COMMITTEE MEMBERS:
Prof. Dr. Elke Teich, Saarland University
Dr. Cristina España-Bonet, Saarland University
DAY OF COLLOQUIUM: June 6, 2024
LOCATION: Saarbrücken, Germany

In loving memory of my father, *Hirak*

1951 – 2020

Abstract

Translated texts exhibit systematic linguistic differences compared to original texts in the same language. These differences are referred to as translationese, and can be categorised as either source language dependent or universal. Basic research on translationese aims to understand and characterise the language-specific and language-independent aspects of translationese. Additionally, translationese has practical implications in the context of natural language processing tasks that involve translation. Translationese effects can cause biased results in a variety of cross-lingual tasks. Therefore, understanding, analysing and mitigating translationese is crucial for improving the accuracy and effectiveness of cross-lingual natural language processing. Focusing on representation learning, this dissertation addresses both foundational as well as practical aspects of translationese.

Our first task is to *investigate the effectiveness of representation learning-based methods in mono- and multilingual translationese classification.* Traditional manual feature-engineering based methods for translationese classification may result in potentially partial, non-exhaustive linguistic features and often require linguistic annotation tools. In contrast, our approach involves developing a suite of representation-learning methods based on word embeddings, eliminating the need for manual feature engineering. Our experiments demonstrate superior performance, outperforming previous traditional hand-crafted linguistically inspired feature-selection methods for translationese classification on a wide range of tasks.

Translationese artifacts have been found to exert a substantial influence on diverse downstream tasks involving translated data. Therefore, to mitigate the impact of translationese on downstream tasks, we propose a new approach: *translationese debiasing*. Our research is the first to adapt the Iterative Null Space Projection (INLP) algorithm, originally designed to mitigate gender attributes, to translationese-induced bias in both word and sentence embedding spaces. Additionally, we develop two techniques for debiasing translationese at the word level representations. We confirm the effectiveness of our debiasing approach by comparing the classification performance before and after debiasing on the translationese classification task. Additionally, we demonstrate the practical utility of our debiasing method by applying it to a natural language inference task involving translated data, where we observed improved accuracy as a result of reduced translation-induced bias.

Next, we address the foundational question of *whether translationese signals can be observed in semantic word embedding spaces* and, if so, what practical implications this observation may have. To this end, we propose a novel approach for unsupervised tracking of translationese in semantic spaces, which does not rely on explicit linguistic labels. Our method is based on graph-isomorphism approaches that examine departures from isomorphism between embedding spaces built from original language data and translations into this language. By comparing the normalised distances between these spaces, we are able to identify systematic evidence of translationese. Specifically, we find that as isomorphism weakens, the linguistic distance between etymologically

distant language families increases, providing evidence that translationese signals are linked to source language interference. Following this, we show that the proposed methods are robust under a variety of training conditions, encompassing data size, type, and choice of word embedding models. Additionally, our findings indicate our methods are language-independent, in the sense that they can be applied to multiple languages and are not limited to a specific language or language family.

We extend the work on unsupervised tracking of translationese in semantic spaces to *evaluate the impact of domain in translationese data.* Translationese signals are subtle and may compete with other signals in the data, particularly those related to domain. To address this, we mask domain information by applying our graph-isomorphism methods to different delexicalized representations using various views, including words, parts of speech, semantic tags, and synsets. Our results demonstrate that while source-language interference is most pronounced in lexicalised embeddings (word), it is also present in delexicalised views. This indicates that our lexicalised findings are not only the result of possible topic differences between original and translated texts. Additionally, we show that, regardless of the level of linguistic representation, language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from the degree of departures from isomorphism, using all combinations of original target language and translations into this target language from different source languages.

Finally, we explore whether the graph-based divergence from isomorphism in embeddings can serve as a viable proxy for surprisal at the the level of surface texts. To do this, we explicitly compute the correlation between (a) differences in surface string entropy of original vs. translated data computed by language models trained on originally authored data and (b) divergence from isomorphism between embedding spaces computed on the same text data. Our results show a positive correlation between these two measures with a higher departure from isomorphism between embedding spaces corresponding to a greater difference in surface entropy. Additionally, similar to the findings of graph-based divergence from isomorphism between embedding spaces where higher divergence from isomorphism implicitly indicates higher linguistic distance in terms of language families and, supposedly, surface structural linguistic distance between languages — our entropy-based findings demonstrate that the observed differences in surface string representations between the original and translated datasets also correspond, at a higher level, with the surface structural linguistic distance between the source languages themselves. These results establish an explicit link between our two measures: divergence from isomorphism between original and translated embedding spaces and entropy differences of the surface strings of the same text data.

Zusammenfassung

Übersetzte Texte weisen im Vergleich zu Originaltexten in derselben Sprache systematische sprachliche Unterschiede auf, die zu einer einzigartigen Teilsprache mit eigenen Unterscheidungsmerkmalen führen. Diese Unterschiede werden als *Translationese* (Gellerstam, 1986) bezeichnet und können entweder als ausgangssprachabhängig oder universell kategorisiert werden. Es ist wichtig zu beachten, dass Translationese keine eigenständige Sprache ist, sondern vielmehr eine Reihe von sprachlichen Merkmalen (Baker et al., 1993; Toury, 1980), die übersetzte Texte von denen unterscheiden, die ursprünglich in der Zielsprache geschrieben wurden.

Verschiedene Faktoren tragen zu den Unterschieden zwischen den Originaltexten und übersetzten Texten bei, von denen viele unter die Kategorie der universellen Merkmale der Übersetzung fallen. Zu diesen universellen Merkmalen gehören die Vereinfachung, d. h. die Vereinfachung komplexer Ausgangsstrukturen in der Zielsprache, die Standardisierung, d. h. die Neigung, sich eng an die Normen der Zielsprache zu halten, und die Explizierung, d. h. die Verdeutlichung impliziter Ausgangsstrukturen in der Zielsprache. Im Gegensatz zu diesen übersetzerischen Universalien spiegelt die Interferenz den Einfluss der Ausgangssprache auf das Übersetzungsprodukt wider. Interferenz ist von Natur aus ein sprachpaarspezifisches Phänomen, bei dem isomorphe Strukturen, die Ausgangs- und Zielsprache gemeinsam haben, einander leicht ersetzen können. Dies verdeutlicht den zugrunde liegenden sprachübergreifenden Einfluss der Ausgangssprache auf das Übersetzungsergebnis.

Nach dieser Definition wird das Übersetzen als eine besondere Form der sprachübergreifenden Sprachvarietät betrachtet, die einen Sonderfall innerhalb der breiteren Landschaft des Sprachkontakts darstellt. Trotz des beträchtlichen Umfangs der Forschung zu verschiedenen sprachübergreifenden Sprachvarietäten gibt es eine auffällige Lücke, wenn es um die spezifische Untersuchung des Translationese geht.

Bisherige Studien haben sich vor allem auf zwei Aspekte konzentriert, nämlich auf die Identifizierung charakteristischer Merkmale von Übersetzungen und auf die Klassifizierung von Translationese, wobei sorgfältige korpusbasierte Studien (Baker et al., 1993) verwendet wurden, die Unterschiede in der Verteilung linguistischer Merkmale zwischen Originalen und übersetzten Texten aufzeigen, oder klassifikationsbasierte Ansätze bei denen ein Klassifikator trainiert wird und dann Merkmalsbedeutungsmaße verwendet werden, um darauf zu schließen, welche Merkmale bei der Klassifizierung von Übersetzungen besonders wichtig sind (Rabinovich and Wintner, 2015; Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). Die praktischen Implikationen von Translationswissenschaft gehen jedoch über die Identifizierung linguistischer Merkmale und die Klassifizierung hinaus.

Die ausgeprägten sprachlichen Unterschiede, die zwischen Originaltexten und übersetzten Texten in derselben Sprache zu beobachten sind, lassen sich im Großen und Ganzen entweder als quellsprachenabhängig oder als universell klassifizieren. Die Durchführung von Grundlagenforschung zum Translationese ist wichtig, da sie wertvolle Einblicke sowohl in sprachspezifische als auch in sprachunabhängige

Aspekte dieses sprachlichen Phänomens liefert. Die Grundlagenforschung ermöglicht ein tieferes Verständnis der zugrundeliegenden Muster und Strukturen, die das Translationese definieren, sowie der Variationen, die in verschiedenen Sprachen und Sprachfamilien beobachtet werden.

Es ist jedoch auch wichtig, eine bestehende Lücke in der Erforschung des Translationese zu schließen. Translationese hat spürbare Auswirkungen auf verschiedene sprachübergreifende Aufgaben der natürlichen Sprachverarbeitung (NLP), was zu verzerrten Ergebnissen und verminderter oder künstlich aufgeblähter Leistung führen kann. Daher ist die Untersuchung und Abschwächung von Translationese für die Verbesserung der Genauigkeit und Effektivität dieser Aufgaben von entscheidender Bedeutung. Die Abschwächung von Translationese ist bisher ein wenig erforschtes Gebiet. Ein wichtiger Teil der in dieser Arbeit vorgestellten Forschung besteht darin, diese Lücke zu schließen.

In dieser Arbeit konzentrieren wir uns auf das Repräsentationslernen als einen umfassenden Ansatz, um sowohl die grundlegenden als auch die praktischen Aspekte von Translationese zu behandeln. Mit dieser Forschungsarbeit wollen wir einen Beitrag zum breiteren Feld der sprachübergreifenden Sprachstudien leisten und eine bestehende Lücke im Verständnis und in der Behebung von Translationese schließen.

Zunächst befassen wir uns mit den praktischen Aspekten von Translationese. Frühere Studien haben gezeigt dass die Verwendung verschiedener manuell erstellter Merkmale für überwachtes Lernen für die Klassifizierung von Translationese effektiv sein kann (Avner, Ordan, and Wintner, 2016; Baroni and Bernardini, 2005; Rabinovich, Ordan, and Wintner, 2017; Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). Dies hat sich als nützlich für eine Vielzahl von Aufgaben erwiesen, wie z. B. die Analyse von Übersetzungsstrategien (Lapshinova-Koltunski, 2015), die Untersuchung der Merkmale (Koppel and Ordan, 2011; Rubino, Lapshinova-Koltunski, and Genabith, 2016) von Translationese oder die Bewertung von maschinellen Übersetzungssystemen (Graham, Haddow, and Koehn, 2019; Zhang and Toral, 2019). In Anlehnung an diese Forschungslinie in der Erforschung des Translationese, konzentrieren wir uns auf zwei praktische Aufgaben: (i) mehrsprachige Klassifizierung von Translationese, und (ii) wir schlagen eine neue Aufgabe vor, nämlich die Milderung von Translationese-Artefakten.

Im Rahmen von (i) entwerfen, entwickeln und evaluieren wir verschiedene Repräsentationslernansätze zur Klassifikation mehrsprachigen Translationese und vergleichende sie mit klassischen manuellen, auf Feature-Engineering basierenden Ansätzen für dieselben Daten. In (ii) führen wir die neue Aufgabe ein, Translationese Artefakte aus latenten Repräsentationsräumen zu entfernen. Dies bezieht sich auf den Prozess der Entfernung oder die Reduzierung des Einflusses von Artefakten auf die gelernten Repräsentationen von Texten, um die Eigenschaften der Originaltexte besser wiederzugeben.

Translationese wurde als eine Reihe von spezifischen linguistischen Merkmalen und Konventionen charakterisiert, die Übersetzungen von Originaltexten unterscheiden (Baker et al., 1993; Teich, 2003; Toury, 1979). Frühere bahnbrechende Forschungen zur automatischen Klassifizierung von Übersetzungen (Baroni and Bernardini, 2005; Koppel and Ordan, 2011) verwendeten traditionelle, handgefertigte, linguistisch inspirierte, auf Features basierende, überwachte maschinelle Lernansätze, um Klassi-

fikatoren zu trainieren. Häufig wurden Feature-Ranking-Methoden verwendet, um herauszufinden, welche der Features wichtige Indikatoren für Translationese sind (Avner, Ordan, and Wintner, 2016; Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). Manuelle, linguistisch inspirierte, auf Feature-Engineering basierende Ansätze haben den Vorteil, dass die verwendeten Merkmale (und ihre Rangfolge) für menschliche Experten leicht zu interpretieren sind. Es gibt jedoch einige Gründe, warum es problematisch sein kann, sich bei der überwachten Klassifizierung von Übersetzungen auf manuell erstellte linguistische Merkmale zu verlassen. Eines der Hauptprobleme bei diesem Ansatz ist, dass die manuell erstellten Merkmale möglicherweise nicht vollständig sind und nicht alle wichtigen Unterscheidungsmerkmale der Eingabedaten während des Trainings erfassen. Dies liegt daran, dass die Merkmale auf linguistischen Intuitionen beruhen und möglicherweise nicht alle möglichen Variationen in den Eingabedaten berücksichtigen. Darüber hinaus erfordert die Annotation linguistischer Daten in großem Umfang (automatische) Annotationswerkzeuge, und die Beschaffung linguistischer Annotationswerkzeuge, wie Tokenisierer, Tagger, morphologische Analysatoren, NERs, Chunkers, Parser usw., kann für viele Sprachen eine Herausforderung darstellen (insbesondere für Sprachen mit geringen Ressourcen), was den Nutzen dieses Ansatzes einschränkt. Darüber hinaus ist die automatische Annotation im großen Maßstab immer verrauscht, und die Merkmale können sprach- oder linguistiktheoriespezifisch sein.

Teilweise als Reaktion auf die Beschränkungen der von der linguistischen Theorie inspirierten Merkmale wurden in früheren Arbeiten auch einfache lexikalisierte Merkmale wie Wort-Token und Zeichen-N-Gramme (Avner, Ordan, and Wintner, 2016) oder Merkmale auf der Grundlage von Zählmodellen, Informationsdichte, Überraschung und Komplexität für die Textklassifizierung, insbesondere bei der Bewertung der Übersetzungsqualität, untersucht. Diese Merkmale dienen als Indikatoren für Translationese sowohl von ursprünglich verfassten als auch von manuell übersetzten Texten (Rubino, Lapshinova-Koltunski, and Genabith, 2016). Diese Forschung stützt sich jedoch auf diskrete zählbasierte Ansätze, die Wörter als diskrete Einheiten behandeln, was zu begrenzten Möglichkeiten der Kontextmodellierung führt. In dieser Arbeit erforschen wir einen alternativen Ansatz zur Klassifizierung von mehrsprachigem Translationese, ohne uns auf manuelles Feature Engineering zu verlassen, was uns dazu motiviert, unsere erste Forschungsfrage zu formulieren.

RQ1: *Inwieweit können Techniken des Repräsentationslernens, wie z. B. Einbettungen, übersetzte und nicht übersetzte Texte ohne vorherige linguistische Annahmen unterscheiden?*

Der erste Beitrag dieser Dissertation in Kapitel 3 besteht darin, eine Reihe von auf Repräsentationslernen basierenden Methoden zu entwerfen, zu implementieren und zu evaluieren, so dass die manuelle Erstellung von Merkmalen überflüssig wird. Auf Merkmalen und Repräsentationen basierende Lernmethoden werden in erster Linie von Faktoren wie den Daten, der Aufgabe und dem Lerner beeinflusst, ohne sich auf vorherige linguistische Annahmen oder Vorurteile zu stützen. Dies steht im Gegensatz zu linguistisch inspirierten, handgefertigten Feature-Engineering-Ansätzen, die keine Garantie dafür bieten, dass die Features und Repräsentationen vollständig sind. Daher wollen wir die Ergebnisse, die mit auf Repräsentationslernen basierenden

Ansätzen für die Klassifikation mehrsprachiger Übersetzungen erzielt werden, mit denen unserer früheren klassischen, auf manuellem Feature-Engineering basierenden Ansätze vergleichen, die linguistisch informierte Methoden und automatische Annotationswerkzeuge für dieselben Daten verwenden, um *RQ1* zu behandeln.

In Kapitel 3 zeigen wir, dass bereits statisch eingebettete, auf Repräsentationslernen basierende Ansätze handgefertigte, linguistisch inspirierte Feature-Selection-Methoden für die Übersetzungsklassifikation bei einer Vielzahl von Aufgaben übertreffen. Darüber hinaus führen wir Experimente mit einer Reihe von Ausgangs-/Zielspachenkombinationen in ein- und mehrsprachigen Umgebungen durch, die belegen, dass auf Repräsentationslernen basierende Methoden bei der Generalisierung auf verschiedene mehrsprachige Aufgaben effektiver sind. Darüber hinaus vergleichen wir unsere Ansätze mit sprachübergreifende neuronalen Ansätzen auf denselben Daten und heben hervor, dass die Klassifizierung von Übersetzungen tiefe neuronale Modelle mit starker Kontextmodellierung erfordert, um optimale Ergebnisse zu erzielen.

Übersetzungsartefakte üben einen erheblichen Einfluss auf verschiedene nachgelagerte Aufgaben aus, die mit Übersetzung zu tun haben. In jüngster Zeit wurde die Aufmerksamkeit auf den Einfluss von Translationese auf Bewertungsmetriken gelenkt (Graham, Haddow, and Koehn, 2019; Toral, 2019; Zhang and Toral, 2019). Edunov et al. (2020) and Freitag, Caswell, and Roy (2019) identifizierten Translationese als eine signifikante Quelle für die Diskrepanz zwischen BLEU-Scores (Papineni et al., 2002) und menschlichen Bewertungen. Die Auswirkung der Übersetzungssprache in den Testsätzen ist mit der Auswirkung in den Trainingsdaten (Bogoychev and Sennrich, 2019; Kurokawa, Goutte, and Isabelle, 2009; Lembersky, Ordan, and Wintner, 2012; Riley et al., 2020a) verbunden, unterscheidet sich aber von dieser. Daher ist es für die Verbesserung der Genauigkeit und Effektivität von sprachübergreifendem NLP von entscheidender Bedeutung, Translationese in der Übersetzungsausgabe zu verstehen, zu analysieren und vor allem abzuschwächen. Dies führt uns zur Formulierung unserer nächsten Forschungsfrage.

RQ2:*Ist es möglich, Übersetzungsartefakte effektiv abzuschwächen?*

Bis heute ist diese wichtige Forschungsfrage unteruntersucht. Die *RQ2* zu adressieren, ist der **zweite Beitrag**, den die Dissertation liefert. Wir präsentieren einen Ansatz zur Minderung der negativen Auswirkungen von Translationese auf sprachübergreifende Aufgaben.

Um dies zu erreichen, schlagen wir einen neuen Ansatz vor: Translationese-Debiasing. Wir entwerfen, implementieren und evaluieren diesen Ansatz, der auf latenten Repräsentationen arbeitet. Durch die Reduzierung der Effekte von Translationese auf nachgelagerte Aufgaben zielt dieser Ansatz darauf ab, die Genauigkeit und Wirksamkeit der überlingualen natürlichen Sprachverarbeitung zu verbessern.

Um dieses Ziel zu erreichen, passen wir den Iterativen Nullraumprojektionsalgorithmus (Ravfogel et al., 2020), der ursprünglich zur Reduzierung von Geschlechtsattributen in neuronalen Repräsentationen entwickelt wurde, an übersetzungsbedingte Verzerrungen in Wort- und Satzeinbettungsräumen an. Zusätzlich entwickeln wir zwei Techniken zum Debiasing von Übersetzungsfehlern auf der Wortebene. Wir evaluieren unseren Ansatz, indem wir die Klassifizierungsleistung von Übersetzungsfehlern vor

und nach dem Debiasing vergleichen, und stellen erwartungsgemäß eine geringere Genauigkeit als Folge fest. Darüber hinaus evaluieren wir die Auswirkungen des Debiasing von Übersetzungsartefakten auf die extrinsische Aufgabe der Natural Language Inference (NLI) in zwei verschiedenen Datenumgebungen. Unsere Ergebnisse zeigen, dass das entzerrte Modell in der Lage ist, die Beziehungen zwischen den Sätzen in der Inferenzaufgabe besser zu erhalten und genauere Inferenzen zu produzieren.

Im zweiten Teil befasst sich die Dissertation mit grundlegenden Fragen zum Translationese, einschließlich der Frage, ob die Signale des Translationese in semantischen Worteinbettungsräumen beobachtet werden können und welche praktischen Auswirkungen dies hat. Übersetzte Texte weisen häufig Muster von Interferenzen in aus Ausgangssprache auf, wobei Merkmale des Ausgangstextes auf den Zieltext übertragen werden Teich, 2003; Toury, 1980. Während frühere Studien mit Hilfe von überwachter Klassifikation und Feature-Engineering (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Rabinovich, Ordan, and Wintner, 2017) systematische Belege für Translationese in übersetzten Texten aufzeigen konnten, sind die Auswirkungen von Translationese auf semantische Räume noch weitgehend unerforscht.

Wir entwerfen, implementieren und evaluieren einen strukturierten und nicht überwachten Ansatz zur Erkennung von Translationese in semantischen Räumen ohne die Notwendigkeit expliziter linguistischer Annotationen. Unser Ansatz verfolgt drei Ziele: erstens die Identifizierung von Translationese-Effekten in semantischen Repräsentationen von Texten; zweitens die Entwicklung einer unüberwachten Methode zur Erkennung dieser Effekte ohne menschliche Annotation; und drittens die Bewertung, ob mögliche Domänenunterschiede für einige unserer Ergebnisse verantwortlich sein könnten. Um diese Ziele zu erreichen, konzentrieren wir uns auf zwei primäre Aufgaben: (i) das Aufspüren von Translationese in semantischen Räumen (ii) die Untersuchung des Einflusses der Domäne auf diese Aufgabe.

Die charakteristischen Merkmale übersetzter Texte werden traditionell in zwei Hauptkategorien eingeteilt: Eigenschaften, die sich aus der Interferenz der Ausgangssprache ergeben, und universelle Merkmale, die sich aus der Übersetzung als kommunikativem Prozess selbst ergeben. Frühere Studien verwenden eine Kombination aus lexikalischen und syntaktischen Merkmalen, um zu zeigen, dass Spuren der Ausgangssprache oder *shining-through* (Teich, 2003) in Übersetzungen sichtbar bleiben. Dies ist darauf zurückzuführen, dass lexikalische und syntaktische Merkmale Hinweise auf die Ausgangssprache eines übersetzten Textes geben können (z. B. Wortstellung, grammalische Strukturen).

Während lexikalische und syntaktische Merkmale für die Identifizierung bestimmter Merkmale von Translationese nützlich sein können, ist es wichtig, Translationese zu identifizieren, ohne dass eine explizite Kennzeichnung oder Überwachung erforderlich ist. So können beispielsweise bestimmte Muster oder Strukturen im semantischen Raum eines übersetzten Textes auf Translationese hinweisen, auch wenn sie nicht direkt mit spezifischen lexikalischen oder syntaktischen Merkmalen übereinstimmen. Dies führt uns zu der nächsten Forschungsfrage.

RQ3:*Ist es möglich, Translationese in semantischen Räumen in einer unüberwachten Weise auf zu spüren?*

Um diese Frage zu beantworten, führt der dritte Beitrag dieser Dissertation eine neue Forschungsrichtung ein: das Aufspüren von Translationese-Signalen in einem semantischen Raum, ohne dass eine explizite Kennzeichnung oder Überwachung erforderlich ist. Im Gegensatz zu früheren Arbeiten, die sich auf überwachte Klassifikation und Feature-Engineering stützten, um Translationese zu identifizieren, ist unser Ansatz völlig unbeaufsichtigt und basiert auf einem Schlüsselkonzept: dem Begriff des *Isomorphismus*.

Das Isomorphie-Prinzip besagt, dass Sprachen ein hohes Maß an Übereinstimmung zwischen Bedeutung und Form auf einer Eins-zu-Eins-Basis aufweisen (Barone, 2016). Im Zusammenhang mit der Erkennung von Translationese in semantischen Räumen würde Isomorphie bedeuten, dass der semantische Raum, der aus den Originaldaten der Zielsprache erstellt wurde, derselbe sein sollte wie der Raum, der aus den Übersetzungen in diese Sprache erstellt wurde, und zwar in Bezug auf die Art und Weise, wie die Wörter innerhalb des Einbettungsraums miteinander verbunden sind. Unser Ziel ist es, Translationese auf der Grundlage von Abweichungen von der Graphen-Isomorphie nachzuvollziehen, wobei die ursprüngliche Zielsprache und die Übersetzungen in diese Zielsprache als Graphenstrukturen in den semantischen Räumen dargestellt werden. Die Abweichungen von der Isomorphie zwischen diesen Graphen deuten auf systematische Anzeichen von Translationese hin. Insbesondere stellen wir fest, dass mit abnehmender Isomorphie der linguistische Abstand zwischen etymologisch weit entfernten Sprachfamilien zunimmt, was den Nachweis erbringt, dass die durch die Abweichung von der Isomorphie erkannten Translationese Signale mit der Interferenz der Ausgangssprache verbunden sind. Unsere Ergebnisse sind vergleichbar mit früheren Ansätzen, die auf oberflächlichen Merkmalen wie Wörtern, n-Grammen oder Parser-Ausgaben basieren.

Anschließend zeigen wir, dass die vorgeschlagenen Methoden unter einer Vielzahl von Trainingsbedingungen robust sind, die die Datengröße, den Datentyp und die Wahl der Worteinbettungsmodelle umfassen. Außerdem zeigen unsere Ergebnisse, dass unsere Methoden sprachunabhängig sind, in dem in dem Sinne, dass sie auf mehrere Sprachen angewendet werden können und nicht auf eine bestimmte Sprache oder Sprachfamilie beschränkt sind.

Schließlich setzen wir das Aufspüren von Translationese in semantischen Räumen fort und reduzieren dabei die Auswirkungen möglicher unterschiedlicher Domänen in übersetzten und ursprünglichen Daten, indem wir verschiedene Sichten auf die Daten verwenden (Wörter, PoS, Synsets und semantische Tags). Im vorangegangenen Kapitel haben wir gezeigt, dass Translationese-Signale in semantischen Worteinbettungsräumen, die aus übersetzten und Originaldaten erstellt wurden, erkannt werden können, aber es bleibt unklar, ob die Signale wirklich auf Translationese hinweisen oder ob sie von anderen Faktoren beeinflusst werden, wie z.B. von möglichen thematischen oder Domänen Unterschieden zwischen dem Original und übersetzten Texten. Translationese Signale sind subtil und können mit anderen Signalen in den Daten konkurrieren, insbesondere mit denen, die Spezifika von Domänen zusammenhängen. Dies veranlasst uns zu unserer letzten Forschungsfrage.

RQ4: *Inwieweit lassen sich die in der Antwort auf RQ3 beobachteten Ergebnisse auf Domänenunterschiede zwischen Original und übersetztem Text zurückführen, im Gegensatz*

zu echten Translationese Signalen?

Unser vierter Beitrag in Kapitel 6 untersucht das Zusammenspiel verschiedener linguistischer Repräsentationen (lexikalisch, morphologisch und syntaktisch) und die Frage, ob die Maskierung lexikalischer Informationen und die dadurch bedingte Verringerung potenzieller Domänensignale die Aufgabe der unüberwachten Rückverfolgung von Übersetzungen in semantischen Räumen beeinflussen, um *RQ4* anzugehen. Bei der Analyse von übersetzten Daten, die Texte enthalten, die aus mehreren Quellen (z. B. aus entfernten Sprachen wie Deutsch und Bulgarisch) in dieselbe Zielsprache (z. B. Englisch) übersetzt wurden, können die Ergebnisse unserer Analyse des semantischen Raums durch Domänenunterschiede in den Daten verzerrt werden und nicht durch eigentliche Tranlationese-Signale wie die Ausgangssprachen der Übersetzungen bedingt sein.

Um dies zu berücksichtigen, maskieren wir lexikalische Domäneninformationen in den Daten. Bei diesem Ansatz werden entlexikalisierte Darstellungen verwendet, die Wörter durch Wortarten (Parts of Speech, PoS), semantische Tags oder Synsets ersetzen. Durch die Anwendung unserer Graph-Isomorphismus-Methoden auf diese Darstellungen können wir bestimmte linguistische Merkmale (wie morphologische Informationen oder einfache syntaktische Konfigurationen - PoS-Sequenzen) erfassen und den Einfluss domänenspezifischer lexikalischer Merkmale (insbesondere des Vokabulars) auf die Analyse von Übersetzungen minimieren.

Unsere Ergebnisse zeigen, dass delexikalisierte Darstellungen (PoS, Synsets oder semantische Tags) immer noch erhebliche Interferenzen mit der Ausgangssprache aufweisen. Dies deutet darauf hin, dass die lexikalisierten Ergebnisse auf der niedrigsten Abstraktionsebene (d. h. Wörter) nicht nur auf mögliche Unterschiede in der Domäne zwischen Original- und Translationese Text zurückzuführen sind. Insgesamt ist dies ein Beleg dafür, dass morphologische und einfache syntaktische Repräsentationen in den Daten auch Translationese Signale enthalten. Um das unüberwachte Aufspüren von Translationese Signalen in semantischen Räumen zu bewerten, untersuchen wir außerdem, inwieweit es möglich ist, die Sprachphylogenie oder die genetischen Beziehungen zwischen Sprachen mit diesen delexikalisierten Repräsentationen zu clustern. Wir zeigen, dass unabhängig von der Ebene der sprachlichen Repräsentation aus den Isomorphieabständen Familenverbindungen der Sprachen mit ähnlichen Eigenschaften wie linguistisch motivierte phylogenetische Bäume abgeleitet werden können, wobei alle Kombinationen von ursprünglicher Zielsprache und Übersetzungen in diese Zielsprache aus verschiedenen Ausgangssprachen verwendet werden.

Im vorigen Kapitel haben wir einige implizite Hinweise darauf gegeben, dass die Abweichung von der Isomorphie zwischen Einbettungsräumen auf strukturelle Oberflächenunterschiede zwischen Sprachfamilien hinweisen, die wir aus der linguistischen Literatur kennen. Eine größere Abweichung von der Isomorphie zwischen Einbettungsräumen deutet auf eine größere sprachliche Distanz in Bezug auf von Sprachfamilien und damit auf einen vermeintlich oberflächlichen strukturellen linguistischen Abstand (z. B. Morphologie, Syntax) zwischen Sprachen, wie sie in der linguistischen Literatur beschrieben werden. Dies wird in Kapitel 6 indirekt durch

die POS-Experimente (und die anderen verschiedenen Ansichten) gezeigt die von lexikalischen Informationen in der ursprünglichen und übersetzerischen Einbettung abstrahieren Experimente, bei denen maskierte Ansichten (POS usw.) immer noch vernünftige Unterschiede in der Isomorphie zwischen den Einbettungsräumen im Original und in der Übersetzung zeigen, und darauf aufbauend phylogenetische Stammbaumergebnisse. Dies wirft die Frage auf, ob es explizite Beweise für den Zusammenhang zwischen Einbettungen und strukturellen Oberflächenunterschieden gibt, was uns dazu veranlasst, unsere letzte Forschungsfrage zu formulieren.

RQ5: *Inwieweit kann graphbasierte Divergenz von Isomorphie in Einbettungsräumen als Proxy für Surprisal auf der Ebene von Oberflächentexten dienen?*

Wir behandeln *RQ5* als den **fünften Beitrag** dieser Dissertation. Um dieser Frage nachzugehen, berechnen wir die Korrelation zwischen (a) Unterschieden in der Oberflächenstringentropie von Original- und übersetzten Daten, die von Sprachmodellen berechnet werden, die auf Originaldaten trainiert wurden, und (b) der Abweichung von der Isomorphie zwischen Einbettungsräumen, die auf denselben Textdaten berechnet werden.

Unsere Ergebnisse zeigen eine Korrelation zwischen diesen beiden Maßen, d. h. eine größere Abweichung von der Isomorphie zwischen Einbettungsräumen entspricht einem größeren Unterschied in der Oberflächenentropie. Dies stellt eine direkte Verbindung zwischen dem Einbettungsraum und der Darstellung von Oberflächenstrings her, wobei letztere durch die Entropie des Sprachmodells gemessen wird. Darüber hinaus zeigen unsere Ergebnisse, dass Übersetzungen in dieselbe Zielsprache aus strukturell unterschiedlichen Ausgangssprachen im Allgemeinen höhere Entropieunterschiede aufweisen, während solche aus strukturell ähnlichen Ausgangssprachen geringere Unterschiede aufweisen. Diese Ergebnisse spiegeln die Muster wider, die bei der Divergenz von Isomorphie zwischen Einbettungsräumen beobachtet wurden, wo Übersetzungen aus strukturell stärker divergierenden Sprachen zu einer größeren Divergenz bei der Isomorphie führen. Diese Ergebnisse stellen explizit eine Verbindung zwischen unseren beiden Messgrößen her: der Isomorphiedivergenz zwischen Original- und übersetzten Einbettungsräumen und den Entropieunterschieden in den Oberflächenstrings derselben Textdaten.

Publications

The following publications form the basis of this dissertation and doctoral studies. In case a paper is part of a chapter in this dissertation, it is cited along with the names of (co-)authors, conference details, abstracts and author contributions. The publications are listed chronologically and divided into two groups, i.e., major and side publications, depending on whether they are included in the dissertation or not. Equal first-author contribution is denoted with an asterisk symbol (*).

Major Publications

- **Chapter 6**: Understanding Translationese in Multi-view Embedding Spaces. Koel Dutta Chowdhury, Cristina España-Bonet & Josef van Genabith (2020). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)* (pp. 6056–6062). Barcelona, Spain (Online). International Committee on Computational Linguistics.

- **Chapter 5**: Tracing Source Language Interference in Translation with Graph-Isomorphism Measures. Koel Dutta Chowdhury, Cristina España-Bonet & Josef van Genabith (2021). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 375–385). Online.

- **Chapter 3**: Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification. Daria Pylypenko*, Kwabena Amponsah-Kaakyire*, Koel Dutta Chowdhury*, Josef van Genabith & Cristina España-Bonet (2021). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8596–8611). Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- **Chapter 4**: Towards Debiasing Translation Artifacts. Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, & Josef van Genabith (2022). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)* (pp. 3983–3991). Seattle, United States. Association for Computational Linguistics.

Side Publications

- Understanding the Effect of Textual Adversaries in Multimodal Machine Translation. Koel Dutta Chowdhury, Desmond Elliott (2019). In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)* (pp. 35–40). Hong Kong, China. Association for Computational Linguistics.

- How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet,

Koel Dutta Chowdhury, Josef van Genabith, Elke Teich (2020). In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)* (pp. 280–290). Online. Association for Computational Linguistics.

- EdinSaar@WMT21: North-Germanic Low-Resource Multilingual NMT. Svetlana Tchistiakova, Jesujoba Alabi, Koel Dutta Chowdhury, Sourav Dutta and Dana Ruiter (2021). In *Proceedings of the Sixth Conference on Machine Translation(WMT)* (pp. 368–375). Online. Association for Computational Linguistics.

- Translating away Translationese without Parallel Data. Rricha Jalota, Koel Dutta Chowdhury, Cristina España-Bonet, Josef van Genabith (2023). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7086–7100). Singapore. Association for Computational Linguistics.

- Mitigating Translationese with GPT-4: Strategies and Performance. Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet , Josef Genabith (2024). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (EAMT)* (pp. 411–430). Sheffield, UK. European Association for Machine Translation.

Acknowledgments

First and foremost, I express my sincere gratitude to my supervisor, Josef van Genabith, for giving me the freedom and inspiration to explore different research directions and the guidance to shape them into scientific contributions. I am greatly indebted to my co-supervisor, Cristina España-Bonet, for her steady presence and encouragement, which were a most welcome antidote to my doubts and skepticism. I am deeply grateful to my mentor, Elke Teich, for always providing guidance and support.

I want to thank Annemarie, Katja, Jörg, Katrin, Luigi, and Stefania for sharing their valuable knowledge and guidance throughout this process. I also appreciate all of my academic collaborators for their insights and contributions throughout this journey.

Along the way, I have been fortunate to have many wonderful people by my side. My heartfelt thanks go to Heike, Badr, Pauline, Maria, Dana, Michael, Daria, Miaoran, Cristina, and Uliana for the laughter and solidarity we shared throughout different stages of this journey. Saarbrücken would have been far less happy and warm without each of you. A big thank you to Stalin for being incredibly dependable, and to my little brother Saad for being a solid support every step of the way.

I also want to express my gratitude to all my colleagues in the SFB Centre for the exceptional support system I have had since the first day I arrived in Germany. The research presented in this dissertation was made possible with the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1102 Information Density and Linguistic Encoding. I extend my thanks to the German Research Center for Artificial Intelligence (DFKI) for allowing me to use their computational resources.

My family has been my support system through thick and thin, and I owe all that I am to them—Maa, Payel, Sarthak, Soumya, Syon, and Aria. They have cheered me on from the other side of the world and have always supported my choices and aspirations. I am deeply thankful to my mother, my sanctuary, for her constant care and the unyielding strength she has shown me throughout.

Finally, this thesis is dedicated to the memory of my role model, my father, for no one would be as proud of me as he would be if he were still here. His passion for learning inspired my pursuit of doctoral studies, and his quiet sacrifices have been constant guiding lights along the way. For that, and for everything else, I will be eternally grateful.

Contents

1 Introduction 1

1.1 Overview . . . 2
1.2 Part I: Practical Aspects of Translationese . . . 4
1.2.1 Chapter 3: Multilingual Translationese Classification . . . 4
1.2.2 Chapter 4: Debiasing of Translation Artifacts . . . 6
1.3 Part II: Foundational Aspects of Translationese . . . 8
1.3.1 Chapter 5: Unsupervised Tracing of Translationese in Semantic Spaces . . . 8
1.3.2 Chapter 6: Influence of Domain on Translationese using Multi-View Semantic Spaces . . . 10
1.3.3 Chapter 7: Divergence from Isomorphism and Surprisal in Translationese . . . 11
1.4 Contributions . . . 13

2 Literature Survey and Related Work 17

2.1 Translationese . . . 17
2.1.1 Translational Language Hypotheses . . . 19
2.2 Representation Learning . . . 20
2.2.1 Language Representation Methods . . . 20
2.2.2 Evaluation . . . 26

3 Multilingual Translationese Classification 29

3.1 Introduction . . . 29
3.2 Related Work . . . 31
3.3 Multilingual Data Analysis . . . 32
3.3.1 Datasets . . . 33
3.3.2 Measures . . . 34
3.3.3 Results . . . 37
3.4 Multilingual Translationese Classification . . . 39
3.5 Experimental Setup . . . 39
3.5.1 Model Specifications . . . 39
3.6 Evaluation and Analysis . . . 43
3.7 Conclusion . . . 46

4 Debiasing of Translation Artifacts 49

4.1 Introduction . . . 49
4.2 Related Work . . . 51
4.3 Translationese Debiasing Strategies . . . 53
4.3.1 Iterative Nullspace Projection . . . 53

4.3.2 Translationese Directions . . . 55
4.4 Experimental Setup . . . 56
4.4.1 Datasets . . . 56
4.4.2 Model Specifications . . . 57
4.5 Evaluation and Analysis . . . 58
4.5.1 Translationese Classification Performance . . . 58
4.5.2 Semantic evaluation . . . 59
4.5.3 Clustering . . . 60
4.5.4 Translationese Word Lists . . . 60
4.5.5 Stability of Stepwise Aligned Space Wordlist . . . 61
4.5.6 Extrinsic Evaluation . . . 61
4.6 Conclusion . . . 64

5 Unsupervised tracing of Translationese in Semantic Spaces 67

5.1 Introduction . . . 67
5.2 Related Work . . . 69
5.3 Isomorphism . . . 71
5.3.1 Gromov-Hausdorff (GH) Distance . . . 72
5.3.2 Eigenvector Similarity (EV) . . . 72
5.3.3 Spectral Graph-based Matching (SGM) . . . 73
5.4 Experimental Setup . . . 74
5.4.1 Datasets . . . 75
5.4.2 Vector Spaces . . . 75
5.4.3 Typological Benchmarks . . . 75
5.5 Evaluation and Analysis . . . 76
5.5.1 Language Distance Measures . . . 77
5.5.2 Correlation with Typology, Geography and Phylogeny Benchmarks 78
5.5.3 Comparison with Previous Approaches . . . 79
5.5.4 Robustness Analysis . . . 80
5.5.5 Results on Non-European Languages . . . 82
5.6 Conclusion . . . 83

6 Influence of Domain on Translationese using Multi-View Semantic Spaces 85

6.1 Introduction . . . 85
6.2 Related Work . . . 87
6.3 Experimental Setup . . . 89
6.3.1 Datasets . . . 89
6.3.2 Views . . . 89
6.3.3 Multi-view Embedding Representations . . . 90
6.3.4 Isomorphism . . . 90
6.4 Evaluation and Analysis . . . 91
6.4.1 Classification Analysis . . . 91
6.4.2 Hierarchical Clustering . . . 91
6.4.3 Typological Characterisation . . . 94
6.5 Conclusion . . . 99

7 Divergence from Isomorphism and Surprisal in Translationese 101
7.1 Introduction . . . 101
7.2 Related Work . . . 103
7.3 Experimental Setup . . . 104
7.3.1 Datasets . . . 104
7.3.2 Divergence from isomorphism as a measure for distance . . . 105
7.3.3 Surprisal as a measure for information density . . . 105
7.4 Evaluation and Analysis . . . 106
7.5 Conclusion . . . 108

8 Conclusion & Future Work 111
8.1 Summary of Dissertation . . . 111
8.2 Future Directions . . . 114

Bibliography 117

List of Figures

Figure 1 Overview of the dissertation structure. . . . 3
Figure 2 In Chapter 3, the task involves identifying whether a given text is original (O) or translated (T) in multilingual settings. The output of the model is a label (O or T) for each input during inference. . . . 5
Figure 3 Chapter. 4 aims to debias translationese signals from latent representation spaces. The input for the model is labeled data, with O representing original texts and T representing translated texts. The output is labeled as O', which is assumed to be the generated surface form of the translated input (T) after removing translationese artifacts through debiasing. The resulting output is expected to be more similar to original texts, hence labeled as *O': Original Like*. . . . 6
Figure 4 Chapter 5 studies the task of detecting interference in semantic spaces using graph-based departures from isomorphism. The input for the model is labeled data with O representing original texts and T representing translated texts. The model's output shows that the topology of the semantic space for original texts (O) is more structurally similar to the source S than that of the translated texts (T). . . . 8
Figure 5 Chapter 6 analyses the influence of domain on translationese tracing in semantic spaces at different levels of data abstraction. The input for the model is a sequence of tags. The distances between clusters in the semantic space, represented by Δ, reflect the degree to which a target translated text exhibits the characteristics of the source language. . . . 9
Figure 6 Chapter 7 explores the relationship between graph-based divergence from isomorphism at the level of embedding spaces and surprisal at the the level of surface texts. The input includes isomorphic divergence measures (GH, SGM, EV) and entropy differences, while the output is the correlation between these measures. . . . 12
Figure 7 CBOW and Skip-gram model architectures. . . . 22
Figure 8 Extraction of the translationese wordlists based on the intersection of their top-k nearest neighbours in Original and Translationese spaces. . . . 59
Figure 9 Clustering on Direct Joint Space, before (up-side) and after (down-side) debiasing. . . . 60
Figure 10 Sentence level NLI debiasing in Symmetric settings (*left*), Asymmetric settings (*right*) . . . 63

Figure 11 Pruned gold tree of Serva and Petroni (2008), SP08 in text from Rabinovich, Ordan, and Wintner (2017). . . . 71
Figure 12 Normalised distances between embedding spaces for original en and translations into en from 16 languages given by our three distance measures using 2 different numbers of data points. . . . 77
Figure 13 Normalised distances between embedding spaces for original en and translations into en from 16 languages given by our three distance measures using 2 different numbers of data points. . . . 77
Figure 14 Normalised distances between original texts and their translationese counterparts in English (en). Embedding spaces are constructed using only function words from the Europarl corpus (left) and the UN parallel corpus (Tolochinsky et al., 2018) (right). . . . 81
Figure 15 GH-RAW . . . 94
Figure 16 GH-PoS . . . 94
Figure 17 GH-SemTag . . . 94
Figure 18 GH-Synset . . . 94
Figure 19 EV-RAW . . . 95
Figure 20 EV-PoS . . . 95
Figure 21 EV-SemTag . . . 95
Figure 22 EV-Synset . . . 95
Figure 23 SGM-RAW . . . 96
Figure 24 SGM-PoS . . . 96
Figure 25 SGM-SemTag . . . 96
Figure 26 SGM-Synset . . . 96
Figure 27 Correlation between Graph-Based Divergence Measures and Entropy . . . 106
Figure 28 Pair Plot between Isomorphic Divergence and Entropic Differences (H) in Translated Texts . . . 107

List of Tables

Table 1 Datasets for evaluation of word similarity/relatedness 27
Table 2 The total number of paragraphs in each dataset. On average there are about 80 tokens each paragraph. 33
Table 3 Stylistic **absolute differences** in text statistics measures between original and translated halves. White cells indicate an increase in the original split, while teal cells indicate a decrease. White cells denote no significant difference. 34
Table 4 Stylistic **absolute differences** in text richness measures between original and translated halves. White cells indicate an increase in the original split, while teal cells indicate a decrease. White cells denote no significant difference. 35
Table 5 Translationese classification average accuracy on the mono- and multilingual test sets (average and standard deviation over 5 runs). 42
Table 6 BERT translationese classification accuracy of all TRG–SRC and TRG–ALL models on TRG–SRC and TRG–ALL test sets (average and standard deviation over 5 runs). Columns: training set; rows: test set. 43
Table 7 Translationese classification accuracy of the ALL–ALL[3] model on all the other test sets (average and standard deviations over 5 runs). The difference from the actual trained model performance is indicated in parentheses 46
Table 8 Translationese classification accuracy of the ALL–ALL[8] model on all test sets (average and standard deviations over 5 runs). The difference from actual trained model performance is indicated in parentheses. 46
Table 9 Sentence Embedding Classification accuracy on original versus translationese using different models. After INLP debiasing, translationese classification reduces to random 50% accuracy in all cases. 57
Table 10 Classification accuracy on original versus translationese with word embeddings using our two approaches, before and after debiasing with INLP. 57
Table 11 Size of translationese word lists created with the usage change algorithm (Gonen et al., 2020). 59
Table 12 Stability of Joint Space Wordlist 61
Table 13 Test set accuracies (5 runs average) for models trained and tested on original SNLI, back-translated with German as pivot and back-translated debiased at word-level 62

Table 14 Mean Kendall correlations with SP08 and average URIEL for various number of datapoints and the function words experiment (FW). 79
Table 15 Mean distance between SP08 and reconstructed phylogenetic trees as compared to previous literature using words, function words (FW), parts of speech (POS), phrase structures (PS) and dependency relations (DepRel) as features. 79
Table 16 Examples of the level of abstraction 90
Table 17 Classification accuracy (%) of English O vs. English T on different levels of abstractions 91
Table 18 Average frequency distributions of Parts-Of-Speech tags for translations into English across language families. Cell colors indicate higher frequencies. 97
Table 19 Averaged frequency distributions for translations into English across language families using Semantic Tags from Bjerva, Plank, and Bos (2016). Cell colors highlight higher frequencies. 98

1

Introduction

Language contact is described as a cover term for the ways in which speakers of different languages or language varieties interact with each other and influence each other's language use. This may entail speaker-internal restructurings of language constructs, such as the adoption of new vocabulary or grammatical structures, as well as the influence of other languages or language varieties on the bi- or multilingual repertoire (Kerswill, 1996) of a speaker. Language contact can occur in a variety of contexts, such as migration, trade, colonisation, or general multilingual settings where exposure to different languages is common.

One specific instance of language contact arises when a particular language system is influenced by another. Translated texts exhibit systematic linguistic differences compared to original texts in the same language, resulting in a unique sub-language with its own distinctive features. These differences are referred to as *translationese* (Gellerstam, 1986), and can be categorised as either source language dependent or universal. It is important to note that translationese is not a distinct language in and of itself, but rather a set of linguistic characteristics (Baker et al., 1993; Toury, 1980) that differentiate translated texts from those originally written in the target language.

Various factors contribute to the distinctions observed between original texts and their translated counterparts, many of which fall under the category of universal features of translation. These universals include simplification, which involves simplifying complex source structures or content in the target language; standardisation, which refers to an inclination to adhere closely to the standards of the target language; and explicitation, which entails making implicit source structures or content more explicit in the target language. In contrast to these translation universals, interference reflects the influence of the source language on the translation product. Interference, inherently, is a language-pair specific phenomenon; where isomorphic structures shared between the source and result of the translation can readily replace each other (Toury, 2012). This highlights the underlying cross-linguistic influence of the source language on the translation outcome. Under this definition, translationese is viewed as a distinct form of cross-lingual language variety, representing a unique case within the broader landscape of language contact.

Despite the considerable amount of research conducted on various cross-lingual language varieties, there is a noticeable gap when it comes to the specific investigation of translationese. Previous studies have predominantly focused on two aspects, namely identifying distinctive features of translationses and classifying translationese, using careful corpus-based studies (Baker et al., 1993; Chesterman, 2004; Gellerstam, 1986) showing differences in the distribution of linguistic features between original and translated texts or classification-based approaches where a classifier is trained and then feature importance measures are used to reason back to which features are specifically important in translationese classification (Baroni and Bernardini, 2005; Ilisei et al., 2010; Koppel and Ordan, 2011; Rabinovich and Wintner, 2015; Rubino, Lapshinova-Koltunski, and Genabith, 2016). However, the practical implications of translationese extend beyond identifying linguistic features of and classifying translationese.

The distinctive linguistic variations observed between original texts and their translated counterparts within the same language can be broadly classified as either source language dependent or universal in nature. Conducting basic research on translationese is important as it provides valuable insights into both language-specific and language-independent aspects of this linguistic phenomenon. Basic research enables a deeper understanding of the underlying patterns and structures that define translationese, as well as the variations observed across diverse languages and language families.

However, it is also important to address an existing gap in translationese research. Translationese has tangible effects on various cross-lingual natural language processing (NLP) tasks, potentially leading to biased results and decreased or artificially inflated performance (Graham, Haddow, and Koehn, 2019; Toral et al., 2018; Zhang and Toral, 2019). Therefore, investigating and mitigating translationese becomes essential for improving the accuracy and effectiveness of these tasks. Mitigating translationese is to date an under-researched area. An important part of the research presented in the thesis is to address this gap.

In this thesis, we focus on representation learning as a comprehensive approach to address both the foundational and practical aspects of translationese. By undertaking this research, we aim to contribute to the broader field of cross-lingual language studies and address an existing gap in understanding and mitigating translationese.

In the next section, we provide a conceptual overview of the work, including the motivation, problem, research questions, and contributions of each chapter.

1.1 Overview

The dissertation is divided into two thematic parts, each examining different aspects of translationese. The first part concentrates on the practical aspects of translationese, while the second part delves into its foundational aspects. The practical section of the

dissertation aims to develop highly performant methods for classifying translationese and, importantly, mitigating the impact of translationese on cross-lingual NLP tasks, whereas the foundational part intends to better trace and understand the linguistic patterns and structures underlying translationese in embedding spaces. To achieve these goals, we employ a representation learning-based approach, which allows us to investigate both the practical and foundational aspects of translationese in a comprehensive manner.

These two parts are based on and extend four peer-reviewed publications presented at major NLP conferences where the author of this thesis is the first author. The specific contributions to the papers by the author of the thesis are detailed in Section 1.4 below. The order of the chapters follows the logical progression of this dissertation. At the same time, the structure of this dissertation is intentionally not kept chronological[1] but rather reflects the evolution of the ideas and arguments supporting the dissertation. Each chapter is self-sufficient: it poses its own research questions, presents related work, proposes a methodology, and describes the experimental results. However, there is also a clear thread that connects all the chapters.

In Figure.1, we outline the interconnections between the various chapters, with higher-level chapters in the hierarchy as foundational chapters pointed to below.

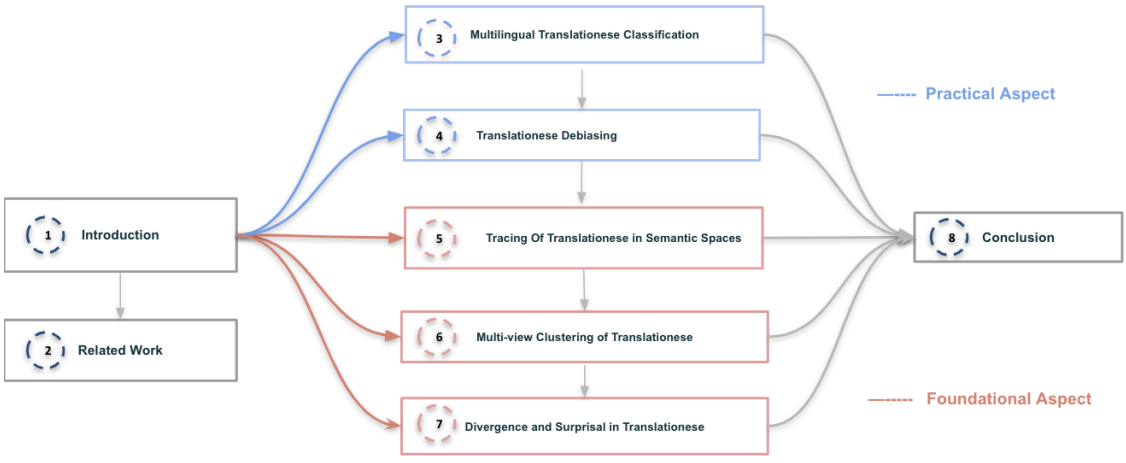

Figure 1: Overview of the dissertation structure.

1 Publications are presented in the thesis chapters in an order that does not necessarily correspond to the order in which they were published.

1.2 Part I: Practical Aspects of Translationese

First, we address practical aspects of translationese. Earlier studies have shown that using various hand-crafted features for supervised learning can be effective for translationese classification (Avner, Ordan, and Wintner, 2016; Baroni and Bernardini, 2005; Rabinovich, Ordan, and Wintner, 2017; Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). This has been shown to be useful for a variety of tasks, such as analysing translation strategies (Lapshinova-Koltunski, 2015), studying the characteristics (Koppel and Ordan, 2011; Rubino, Lapshinova-Koltunski, and Genabith, 2016) of translationese or evaluating machine translation systems (Graham, Haddow, and Koehn, 2019; Zhang and Toral, 2019). Following this line of research in translationese studies, we focus on two practical tasks: (i) multilingual translationese classification, and (ii) we propose a novel task, namely debiasing of translationese artifacts.

In the context of (i), we design, develop and evaluate various representation learning approaches to multilingual translationese classification and compare them to classical manual feature-engineering-based approaches on the same data. In (ii), we introduce a new task of debiasing translationese artifacts from latent representation spaces. This refers to the process of removing or reducing the influence of translationese on the learned representations of texts, in order to better reflect characteristics of original texts.

1.2.1 Chapter 3: Multilingual Translationese Classification

Translationese has been characterised as the set of specific linguistic features and conventions that distinguish translations from originally authored text (Baker et al., 1993; Teich, 2003; Toury, 1979). Earlier seminal research on automatic translationese classification (Baroni and Bernardini, 2005; Ilisei et al., 2010; Koppel and Ordan, 2011) used traditional hand-crafted linguistically inspired feature-engineering-based supervised machine learning approaches to train classifiers. Often feature-ranking methods were used to reason back to which of the features would be important indicators of translationese (Avner, Ordan, and Wintner, 2016; Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). Manual linguistically inspired feature engineering-based approaches have the advantage that the features used (and their rankings) are readily interpretable to human experts. However, there are a few reasons why relying on hand-crafted linguistic features for supervised translationese classification can be problematic. One of the main concerns with this approach is that manually designed hand-crafted features may not be exhaustive and may not capture all the important discriminative characteristics of the input data during training. This is because the features are based on linguistic intuitions and may not account for all

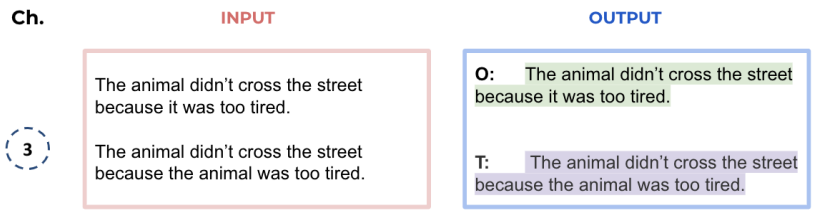

Figure 2: In Chapter 3, the task involves identifying whether a given text is original (O) or translated (T) in multilingual settings. The output of the model is a label (O or T) for each input during inference.

possible variations in the input data. Furthermore, annotating linguistic data at scale with linguistic information requires (automatic) annotation tools, and the availability of linguistic annotation tools, such as tokenisers, taggers, morphological analysers, NERs, chunkers, parsers etc., for many languages can be challenging (particularly for low resource languages), which limits the usefulness of this approach. Furthermore, automatic annotation at scale is always noisy, and features may be language or linguistic theory-specific. Partially as a response to the limitations of linguistic theory-inspired features, some earlier seminal work also explored simple lexicalised features including word tokens and character n-grams (Avner, Ordan, and Wintner, 2016), or features based on count models, information density, surprisal, and complexity for text classification, especially in translation quality estimation (Rubino, Lapshinova-Koltunski, and Genabith, 2016). These features serve as indicators of translationese for distinguishing between originally authored and manually translated texts. However, this research relied on discrete count-based approaches that treat words as discrete units, leading to limited context modelling capabilities.

In this thesis, we explore an alternative approach to multilingual translationese classification without reliance on hand-crafted feature engineering, which motivates us to formulate our first research question.

RQ1: To what extent can representation learning techniques, such as embeddings, discern translated and non-translated texts, without prior linguistic assumptions?

As the **first contribution** of this dissertation in Chapter 3, we design, implement and evaluate an array of representation-learning-based methods based on neural networks, eliminating the need for manual feature engineering. Figure 2 presents a brief chapter overview.

Feature and representation learning-based methods are primarily influenced by factors such as the data, the task, and the learner, without relying on prior linguistic assumptions or preconceptions. This stands in contrast to linguistically inspired hand-crafted feature engineering approaches, which provide no guarantees that

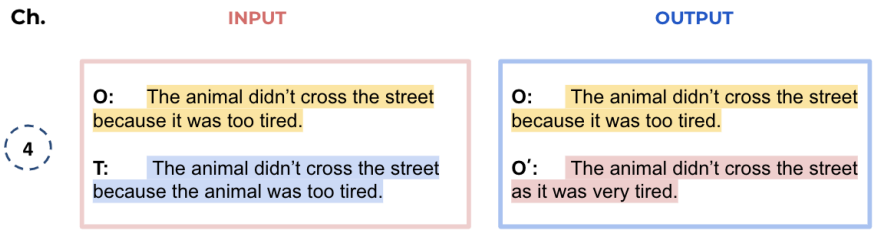

Figure 3: Chapter. 4 aims to debias translationese signals from latent representation spaces. The input for the model is labeled data, with O representing original texts and T representing translated texts. The output is labeled as O', which is assumed to be the generated surface form of the translated input (T) after removing translationese artifacts through debiasing. The resulting output is expected to be more similar to original texts, hence labeled as *O': Original Like*.

the features and representations are in any sense comprehensive. Therefore, we compare the results obtained from representation learning-based approaches for multilingual translationese classification with those of our previous classical manual feature-engineering-based approaches, using linguistically informed methods and automatic annotation tools on the same data, to address *RQ1*.

In Chapter 3, we show that already static embedding-based representation-learning-based approaches outperform hand-crafted linguistically inspired feature-selection methods for translationese classification on a wide range of tasks. In addition, we perform experiments across a range of source-target language combinations in mono- and multilingual settings, providing evidence that representation learning-based methods are more effective in generalising to different multilingual tasks. Further, we compare our approaches to more sophisticated neural approaches on the same data and highlight that translationese classification requires deep neural models with strong context modelling for optimum results.
The content presented in Chapter 3 is based on:

> Daria Pylypenko*, Kwabena Amponsah-Kaakyire*, **Koel Dutta Chowdhury***, Josef van Genabith & Cristina España-Bonet (2021). "Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification". In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
> URL: https://aclanthology.org/2021.emnlp-main.676/

1.2.2 Chapter 4: Debiasing of Translation Artifacts

Translationese artifacts may exert a substantial influence on diverse downstream tasks that involve translation. Recently, attention has been directed towards the influence of translationese on Machine Translation evaluation metrics and evaluation (Graham, Haddow, and Koehn, 2019; Toral, 2019; Zhang and Toral, 2019). Edunov et al. (2020)

and Freitag, Caswell, and Roy (2019) identify translationese as a significant source of misalignment between BLEU scores (Papineni et al., 2002) and human evaluations. The impact of translationese in test sets is related to but distinct from its impact in the training data (Bogoychev and Sennrich, 2019; Kurokawa, Goutte, and Isabelle, 2009; Lembersky, Ordan, and Wintner, 2012; Riley et al., 2020a). Therefore, understanding, analysing and and, most importantly mitigating translationese in translation output is crucial for improving the accuracy and effectiveness of cross-lingual NLP. This leads us to formulate our next research question.

RQ2: Is it possible to effectively attenuate translation artifacts from latent representation spaces?

To date, this important research question remains understudied. Addressing *RQ2* is the **second contribution** provided by the dissertation. Figure 3 presents a brief chapter overview. We present an approach to mitigate the negative impact of translationese on cross-lingual tasks. To accomplish this, we propose a new approach: translationese debiasing. We design, implement and evaluate this approach operating on latent representations. By reducing the effects of translationese on downstream tasks, this approach seeks to improve the accuracy and efficacy of cross-lingual natural language processing.

To achieve this goal, we adapt the Iterative Null Space Projection algorithm (Ravfogel et al., 2020), which was originally designed to reduce gender attributes in neural representations, to translationese-induced bias in both word and sentence embedding spaces. Additionally, we develop two techniques for debiasing translationese at the level of word representations. We evaluate our approach by comparing translationese classification performance before and after debiasing and as expected, observe reduced accuracy as a result of debiased translationese artifacts. Further, we evaluate the effects of debiasing translation artifacts on the extrinsic task of Natural Language Inference (NLI) in two different data settings. Our findings show that the debiased model is able to better preserve the relationships between the sentences in the inference task and produce more accurate inferences.

The content presented in Chapter 4 is based on:

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, & Josef van Genabith (2022). "Towards Debiasing Translation Artifacts". In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. Seattle, United States. Association for Computational Linguistics. URL: https://aclanthology.org/2022.naacl-main.292/

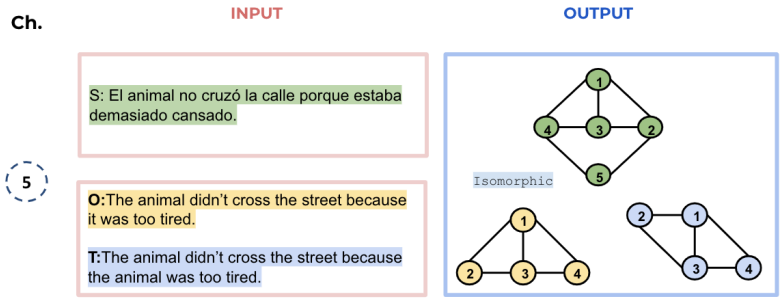

Figure 4: Chapter 5 studies the task of detecting interference in semantic spaces using graph-based departures from isomorphism. The input for the model is labeled data with O representing original texts and T representing translated texts. The model's output shows that the topology of the semantic space for original texts (O) is more structurally similar to the source S than that of the translated texts (T).

1.3 Part II: Foundational Aspects of Translationese

In the second part, the dissertation addresses foundational questions about translationese, including whether translationese signals can be observed in semantic word embedding spaces and the practical implications of this. Translated texts often exhibit patterns of source language interference, where characteristics of the source text transfer to the target text (Teich, 2003; Toury, 1980). While previous studies have used supervised classification and feature engineering (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Rabinovich, Ordan, and Wintner, 2017) to demonstrate systematic evidence of translationese in translated texts, the impact of translationese on semantic spaces remains largely unexplored.

We design, implement and evaluate a structured and unsupervised approach to detect translationese in semantic spaces without the need for explicit linguistic labels. Our approach has three objectives: first, to identify translationese effects in semantic representations of text; second, to develop an unsupervised method to detect these effects without human annotation; and third, to assess whether possible domain differences may account for some of our results. To achieve these objectives, we focus on two primary tasks: (i) tracing of translationese in semantic spaces (ii) investigating the impact of domain on this task.

1.3.1 Chapter 5: Unsupervised Tracing of Translationese in Semantic Spaces

The distinguishing characteristics of translated texts have been traditionally classified into two main categories: properties that stem from *interference* of the source language (Toury, 1979), and universal traits resulting from translation as a communicative process itself. Previous studies use a combination of lexical and syntactic features to show that footprints of the source language or *shining-through* (Teich, 2003) remain visible in translations. This is due to the fact that lexical and syntactic features can

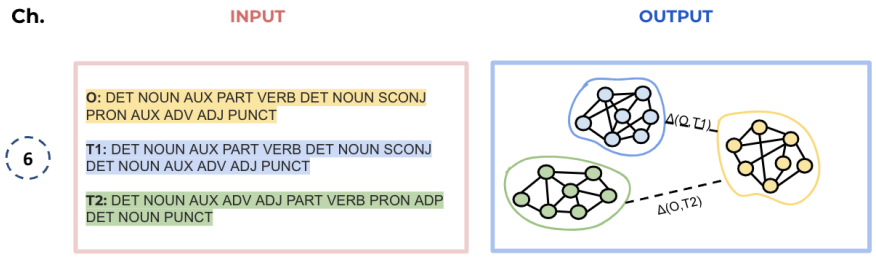

Figure 5: Chapter 6 analyses the influence of domain on translationese tracing in semantic spaces at different levels of data abstraction. The input for the model is a sequence of tags. The distances between clusters in the semantic space, represented by Δ, reflect the degree to which a target translated text exhibits the characteristics of the source language.

provide clues (e.g., word order choices, grammatical structures) about the source language of a translated text. While lexical and syntactic features can be useful for identifying certain characteristics of translationese, it is important to identify translationese without the need for explicit labeling or supervision. For example, certain patterns or structures in the semantic space of a translated text may be indicative of translationese, even if they do not correspond directly to specific lexical or syntactic labels or features. This prompts us to the next research question.

RQ3: Is it possible to track translationese in semantic spaces in an unsupervised manner?

Thus, to address *RQ3*, the **third contribution** of this dissertation introduces a new line of research: tracking of translationese signals in a semantic space without the need for explicit labelling or supervision. Figure 4 presents a brief chapter overview. Unlike previous work that relied on supervised classification and feature engineering to identify translationese, our approach is fully unsupervised and based on a key concept: the notion of *isomorphism*.

The isomorphic principle maintains that languages have a high level of correspondence between meaning and form on a one-to-one basis (Barone, 2016). In the context of detecting translationese in semantic spaces, strict isomorphism would imply that the semantic space created from original target language data should be the same as the space created from translations into that language in terms of how words are interconnected within the embedding space. We aim to trace translationese based on deviation from graph-isomorphism, where the original target language and translations into this target language are represented as graph structures connecting word embeddings in the semantic spaces. The departures from isomorphism between these graphs indicate systematic evidence of translationese. Specifically, we find that as isomorphism weakens, the linguistic distance between etymologically distant language families increases, providing evidence that the translationese signals detected by departure

from isomorphism are linked to source language interference. Our results perform on par with previous approaches (Bjerva et al., 2019; Rabinovich, Ordan, and Wintner, 2017) based on surface-level features such as words, n-grams, or parser outputs.

Following this, we show that the proposed methods are robust under a variety of training conditions, encompassing data size, type, and choice of word embedding models. Additionally, our findings indicate that our methods are language-independent, in the sense that they can be applied to multiple languages and are not limited to a specific language or language family.
The content presented in Chapter 5 is based on:

> **Koel Dutta Chowdhury**, Cristina España-Bonet, & Josef van Genabith (2021). "Tracing Source Language Interference in Translation with Graph-Isomorphism Measures". In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*. Online. URL: https://aclanthology.org/2021.ranlp-1.43.pdf

1.3.2 Chapter 6: Influence of Domain on Translationese using Multi-View Semantic Spaces

We continue tracking translationese in semantic spaces while reducing the impact of possible different domains in translated and original data using various views of the data (words, PoS, synsets, and semantic tags). In the previous chapter, we show that translationese signals can be detected in semantic word embedding spaces built from translated and original data, but it remains inconclusive if they are truly indicative of translationese, or whether they may be influenced by other factors such as possible domain differences between the original and translated texts reflected in the lexical dimensions of the original and the tranlated texts. Translationese signals are subtle and can compete with other signals in the data, particularly those related to domain.

This prompts us to our next research question.

> *RQ4: To what extent can the outcomes observed in response to RQ3 be attributed to variations in domain between original and translated texts, as opposed to true translationese signals?*

Our **fourth contribution** in Chapter 6 explores the interplay of different linguistic representations (lexical, morphological, and syntactic) and whether masking lexical information and thereby reducing potential domain signals influence the task of unsupervised tracing of translationese in semantic spaces to address *RQ4*. Figure 5 presents a brief chapter overview. When analysing translated data that contains texts translated from multiple sources (such as distant languages like German and Bulgarian) into the same target language (e.g., English), the results of our semantic

space analysis may be skewed towards domain differences in the data rather than proper translationese signals such as the source languages of the translations.

To account for this, we mask lexical domain information in the data. This approach involves using delexicalised representations that replace words with parts of speech (PoS), semantic tags, and synsets. By applying our graph-isomorphism methods to these representations, we can capture distinct linguistic dimensions (such as, morphological information or simple syntactic configurations - PoS sequences) and minimise the influence of domain-specific lexical features (in particular vocabulary) on the analysis of translationese.

Our results show that delexicalised representations (PoS, Synsets, and Semantic Tags) still exhibit significant source language interference. This indicates that the lexicalised results from the lowest level of abstraction (i.e., words) are not solely due to possible differences in the domain between original and translated texts. Overall, this provides evidence that morphological and simple syntactic representations in the data also carry translationese signals. Additionally, to evaluate the unsupervised tracing of translationese signals in semantic spaces, we examine the extent to which it is possible to cluster language phylogeny or the genetic relationships between languages with these delexicalised representations. We show that, regardless of the level of linguistic representation, language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from the isomorphism distances, using all combinations of original target language and translations into this target language from different source languages.
The content presented in Chapter 6 is based on:

> **Koel Dutta Chowdhury**, Cristina España-Bonet, & Josef van Genabith (2020). "Understanding Translationese in Multi-view Embedding Spaces". In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain (Online). International Committee on Computational Linguistics.
> URL: https://aclanthology.org/2020.coling-main.532/

1.3.3 Chapter 7: Divergence from Isomorphism and Surprisal in Translationese

In the previous chapter, we provide some implicit evidence that divergence from isomorphism between embedding spaces indicates structural surface differences between language families we know from the linguistic literature. Higher divergence from isomorphism between embedding spaces indicates higher linguistic distance in terms of language families and with that supposedly surface structural linguistic distance (e.g. morphology, syntax) between languages as from the linguistic literature (Haspelmath et al., 2005). This is shown indirectly in Chapter 6 by the POS (and the other different views) experiments that abstract from lexical information in the original and translationese embedding experiments where masked views (POS etc.)

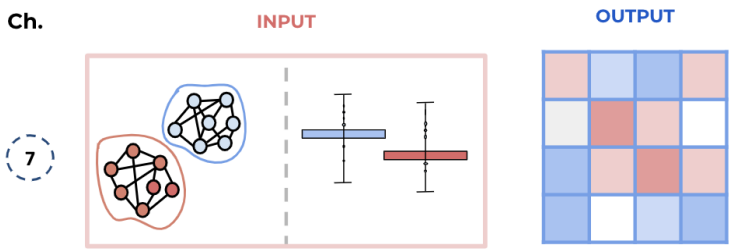

Figure 6: Chapter 7 explores the relationship between graph-based divergence from isomorphism at the level of embedding spaces and surprisal at the the level of surface texts. The input includes isomorphic divergence measures (GH, SGM, EV) and entropy differences, while the output is the correlation between these measures.

still show reasonable original vs. translationese differences in isomorphism between the embedding spaces and based on this, phylogenetic family tree results. This raises the question of whether there is any explicit evidence to support the connection between embeddings and structural surface differences, prompting us to formulate our last research question.

RQ5: To what extent can graph-based divergence from isomorphism in embedding spaces act as a proxy for surprisal at the the level of surface texts?

Figure 6 presents a brief chapter overview. We address *RQ5* as the **fifth contribution** of this dissertation. To demonstrate this, we compute the correlation between (a) differences in surface string entropy of original vs. translated data computed by language models trained on originally authored data and (b) divergence from isomorphism between embedding spaces computed on the same text data. Our results show a correlation between these two measures, showing that a higher departure from isomorphism between embedding spaces corresponds to a greater difference in surface entropy. Additionally, our findings show that translations into the same target language from structurally divergent source languages generally exhibit higher entropy differences, while those from structurally similar source languages show smaller differences. These results mirror the patterns observed in the findings of graph-based divergence from isomorphism between embedding spaces, where translations into the same target language from structurally more divergent source languages lead to increased divergence in isomorphism, implying an implicit connection between embeddings and structural surface differences as generally assumed in the linguistic literature regarding language family relations and surface structural differences (Haspelmath et al., 2005). In contrast, in this chapter, we provide hard evidence for these patterns by linking our two explicit measures: divergence from isomorphism between original and translated embedding spaces and entropy differences of the surface strings of the same text data.

1.4 Contributions

The dissertation is composed of the following peer-reviewed articles, each accompanied by citations that include the names of (co-)authors, conference details, abstracts, and author contributions.

- **Chapter 6: Understanding Translationese in Multi-view Embedding Spaces**
  Koel Dutta Chowdhury, Cristina España-Bonet & Josef van Genabith (2020). In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.
  Abstract: Recent studies use a combination of lexical and syntactic features to show that footprints of the source language remain visible in translations, to the extent that it is possible to predict the original source language from the translation. In this paper we focus on embedding-based semantic spaces, exploiting departures from isomorphism between spaces built from original target language and translations into this target language to predict relations between languages in an unsupervised way. We use different manifestations of the same data, in the form of different features, such as words, Part-of-Speech, Semantic Tags, and Synsets to understand these relations. Our analysis shows that (i) semantic distances between original target language and translations into this target language can be detected using the notion of isomorphism, (ii) language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from the distances and (iii) that, perhaps surprisingly, even delexicalised embeddings exhibit significant source language interference, indicating that the lexicalised results are not "just" due to possible differences in topic between original and translated texts. To the best of our knowledge, it is the first time, departures from isomorphism between embedding-based semantic spaces is used to detect translationese.
  Author Contribution: As a first author, Koel Dutta Chowdhury conceptualized the core research idea, conducted the experiments, performed the analysis and wrote the paper. Cristina España-Bonet provided useful references on unsupervised clustering and was involved in shaping the research questions. Josef van Genabith advised the project, presented the need for delexicalisation, and provided feedback to improve the final version of the paper.

- **Chapter 5: Tracing Source Language Interference in Translation with Graph-Isomorphism Measures**
  Koel Dutta Chowdhury, Cristina España-Bonet & Josef van Genabith (2021). In *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.
  Abstract: Previous research has used linguistic features to show that trans-

lations exhibit traces of source language interference and that phylogenetic trees between languages can be reconstructed from the results of translations into the same language. Recent research has has been shown that instances of translationese (source language interference) can even be detected in embedding spaces, comparing embeddings spaces of original language data with embedding spaces resulting from translations into the same language, using a simple Eigenvector-based divergence from isomorphism measure. To date it remains an open question whether alternative graph-isomorphism measures can produce better results. In this paper, we (i) explore Gromov-Hausdorff distance, (ii) present a novel spectral version of the Eigenvector-based method, and (iii) evaluate all approaches against a broad linguistic typological database (URIEL). We show that language distances resulting from our spectral isomorphism approaches can reproduce genetic trees at par with previous work without requiring any explicit linguistic information and that the results can be extended to non-Indo-European languages. Finally, we show that the methods are robust under a variety of modeling conditions.

Author Contribution: As a first author, Koel Dutta Chowdhury conceptualized the core research idea, conducted the experiments, performed the analysis and wrote the paper. Cristina España-Bonet presented the need for introducing low resource settings, and robustness analysis. Josef van Genabith supervised the project and provided important insights on isomorphism studies and helped with proofreading.

- **Chapter 3: Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification**
  Daria Pylypenko*, Kwabena Amponsah-Kaakyire*, Koel Dutta Chowdhury*, Josef van Genabith & Cristina España-Bonet (2021). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
  Abstract: Traditional hand-crafted linguistically-informed features have often been used for distinguishing between translated and original non-translated texts. By contrast, to date, neural architectures without manual feature engineering have been less explored for this task. In this work, we (i) compare the traditional feature-engineering-based approach to the feature-learning-based one and c analyse the neural architectures in order to investigate how well the hand-crafted features explain the variance in the neural models' predictions. We use pre-trained neural word embeddings, as well as several end-to-end neural architectures in both monolingual and multilingual settings and compare them to feature-engineering-based SVM classifiers. We show that (ii) neural architectures outperform other approaches by more than 20 accuracy points, with the BERT-based model performing the best in both the monolingual and multilingual settings; (iii) while many individual hand-crafted translationese

features correlate with neural model predictions, feature importance analysis shows that the most important features for neural and classical architectures differ; and (*iv*) our multilingual experiments provide empirical evidence for translationese universals across languages.
Authors Contribution: First author*. Koel Dutta Chowdhury jointly conceptualized the research, conducted experiments related to embedding-based classification and performed analysis. Daria Pylypenko and Kwabena Amponsah-Kaakyire implemented the deep neural models and helped with paper writing. Josef van Genabith and Cristina España-Bonet supervised the project and provided important feedback on the final drafts of the manuscript.

- **Chapter 4: Towards Debiasing Translation Artifacts**
  Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, & Josef van Genabith (2022). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.
  Abstract: Cross-lingual natural language processing relies on translation, either by humans or machines, at different levels, from translating training data to translating test sets. However, compared to original texts in the same language, translations possess distinct qualities referred to as translationese. Previous research has shown that these translation artifacts influence the performance of a variety of cross-lingual tasks. In this work, we propose a novel approach to reducing translationese by extending an established bias-removal technique. We use the Iterative Null-space Projection (INLP) algorithm, and show by measuring classification accuracy before and after debiasing, that translationese is reduced at both sentence and word levels. We evaluate the utility of debiasing translationese on a natural language inference (NLI) task, and show that by reducing this bias, NLI accuracy improves. To the best of our knowledge, this is the first study to debias translationese as represented in latent embedding space.
  Author Contribution. As a first author, Koel Dutta Chowdhury conceptualized the research, identified the problem, conducted the experiments, performed the analysis and wrote the paper. Josef van Genabith advised and helped with the conceptualization of translationese subspaces. Rricha Jalota ran the extrinsic evaluation pipeline. Cristina España-Bonet suggested ideas on intrinsic evaluations of the debiased word representations. All co-authors provided feedback.

2

Literature Survey and Related Work

This chapter introduces fundamental concepts related to the two main topics in this dissertation: (1) translationese, and (2) representation learning (i.e., approaches for inducing semantic representations of text).

2.1 Translationese

The study of translation has been a topic of interest for scholars across various disciplines, including linguistics, computational linguistics, literature, translation studies, and cultural studies, among others. Translation, defined as the transfer of meaning from one language to another, is a complex process that involves not only linguistic competence but also cultural knowledge and contextual understanding. The goal of translation is to produce a target text that conveys the meaning and intent of the source text while also taking into account the linguistic and cultural differences between the two languages.

One unique characteristic of translated texts is the presence of what is known as translationese (Gellerstam, 1986). Translationese refers to the presence of unique features in translated texts which are not present (to the same extent) in texts originally authored in the same language. These features may include syntactic constructions, lexical choices, and other linguistic patterns that are influenced by the source language and the process of translation itself. Translationese has been the subject of much discussion and analysis in the field of translation studies, and various hypotheses and theories have been proposed to explain its nature and impact.

One of the earliest works on this topic, Toury (1979) identified translationese as a linguistic system, being a form of *interlanguage*, that "enjoys an intermediate status between the source language and target language", and exhibits an "interference of these two codes [source language, target language] in the performance of the learner". More recently, Toury (2012) has also suggested that translations can be thought of as distinct dialects of the target language, highlighting the ways in which translated text differs from originally authored text in the same language.

Duff (1981) used the term "the third code" or "the third language" to describe this phenomenon, emphasizing the influence of the source language on translated texts. However, Frawley (1984) was the first to identify translated language as an autonomous and distinctive sub-language, introducing the term "the third code". According to Frawley (1984),

> "the translation itself [is] essentially a third code which arises out of the bilateral consideration of the matrix and target codes: it is, in a sense, a sub-code of each of the codes involved."

Gellerstam (1986) adopted the term *translationese* and demonstrated that there were statistically significant differences in the frequencies of loan words and colloquialisms, among other lexical features, between texts in original Swedish and in Swedish texts originating from translation from English. He noted that these differences between the original and translated texts do not necessarily reflect poor translation but rather a statistical phenomenon brought on by the systematic effect of the source language on the target language.

Baker et al. (1993) proposed translation universals, which are features that are considered invariant and apply to all translated texts independently of the source language and translation direction. Baker et al. (1993) points out that:

> "it will be necessary to develop tools that will enable us to identify universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems."

In other words, Baker's theory postulated the existence of universal tendencies that can be found in all translations, irrespective of the source text or target language. Her proposed initial translation universals hypotheses are grouped around three fundamental tendencies: *Simplification*, *Levelling out* (also referred to as conventionalisation) and *Explicitation*.

In addition to the aforementioned translation universals, Toury (2012) proposed two other general properties of translated texts, called *Normalisation* (also referred to as *homogeneisation*) and the law of *Interference* (also referred to as *shining-through* (Teich, 2003) from the source text.

While Toury (2012)'s laws of translation and Baker et al. (1993)'s universal translation hypotheses are significant breakthroughs in the research methods within translation studies, other scholars also have made important contributions regarding understanding the nature of translated texts and different standards, hypotheses or tendencies such as transfer, translation unique items, asymmetry (Kenny, 1998; Mauranen, 2008; Pastor et al., 2008). In the following section, we outline the translational language hypotheses.

2.1.1 Translational Language Hypotheses

There have been several studies that have identified specific linguistic patterns that occur frequently in translated languages. These patterns can be grouped into five main hypotheses: simplification, explicitation, levelling out, interference, and normalization.

1. Simplification: This refers to "the tendency to simplify the language used in translation" (Baker et al., 1993). For example, this can take the form of shorter sentences, the use of simpler words and phrases, avoidance of repetitions, disambiguation, and the removal of punctuation marks. As a result, translated texts are easier to understand than original texts as translators tend to simplify the intricate linguistic features found in the source text into less complicated, simplified features in the target text for the readers. Subsequently, translated texts may exhibit a lower lexical density and variety compared to original texts.

2. Explicitation: This refers to the tendency to "spell things out rather than leave them implicit" in translated texts (Baker et al., 1993), resulting in translated texts that tend to be longer than their original counterparts. This can manifest in various ways, such as increased length of translations, overuse of explanatory vocabulary and conjunctions, more frequent use of cohesive markers (e.g., therefore, thus, hence) or the replacement of pronouns with more explicit noun phrases in translated texts (Koppel and Ordan, 2011).

3. Levelling out or Conventionalisation: This refers to the fact that translational languages tend to "steer a middle course between any two extremes, converging towards the centre"(Baker et al., 1993), i.e. translated texts exhibit relatively more similar language features or higher level of homogeneity to each other than their sources. For example, Baker et al. (1993) notes that translated texts tend to exhibit similar values in measures such as lexical density, type-token ratio, and mean sentence length, whereas non-translated texts exhibit greater variation in these measures.

4. Interference or Shining-through: The prominent characteristics of source-language interference (Toury, 1979) in the translated text are also popularly termed as "shining-through" (Teich, 2003) in the literature. Toury (1979) observed that when a source text is translated, structural elements of the source text tend to be transferred to the target text. This refers to the fact that translated texts often retain the linguistic characteristics of the source language, rather than conforming to the norms of the target language. Volansky, Ordan, and Wintner (2015) notes that this phenomenon can operate on different levels from transcribing source-language words to applying loan translations to exerting structural influence.

5. Normalisation or Conservatism: This refers to the tendency of translations to conform to patterns and practices that are typical of the target language, even to the extent of exaggerating them (Baker et al., 1993). This can manifest in a variety of ways, such as the use of certain language patterns or conventions that are more common in the target language, or the exaggeration of certain language features in order to align more closely with target language norms.

In the second part of the thesis, we specifically focus on tracing and analysing the properties that stem from interference of the source language (Chapters 5 and 6) because it relates to the direct transfer of features from the source language to the target language.

2.2 Representation Learning

Representation learning is a central aspect of modern natural language processing (NLP) and involves identifying principles that allow machines to learn linguistic representations from raw input data. The approaches discussed in this thesis fall within the framework of representation learning, which consists of machine learning techniques that learn helpful representations for various tasks from raw input.

In this section, we briefly introduce the representation learning methods that are used to learn meaningful and useful word embeddings for the study of translationese in this thesis.

2.2.1 Language Representation Methods

For many years, the standard practice in language representation involved human-crafted features. The conventional approach centered around two key methodologies: the bag-of-words (BoW) model, developed by Harris (1954), and n-gram models, introduced by Baker (1975), Jelinek (1976), and Shannon (1948). In the bag-of-words (BoW) model, a text is represented as a list of words and how often they appear (a word frequency list). It assumes documents are similar if they share similar words. However, BoW does not consider grammar or word order. Alternatively, the n-gram model keeps track of word sequences by predicting the probability of the next word based on its previous $n-1$ words.

As with all NLP tasks, a numerical representation of text input is required for translationese detection in order for computational models to process it. This section introduces embeddings, which are dense representations of words or subwords and are used in the research work presented in this thesis. Generally, embeddings can be divided into two broad categories: (i) static word representations and (ii) contextualised word representations. Embeddings reflect distributional properties in the

data: words in similar contexts tend to have similar meanings (Harris, 1954) and will be mapped close to each other in the embedding space. In this thesis, we focus on embeddings that are obtained through neural architectures.

2.2.1.1 Static Representations

Static word embedding methods are a type of method for learning word vectors that map words from large text corpora to a numerical representation, or a real-valued vector representation in a high-dimensional space. The representations are static as the end result of the process is a single dense vector for a word type (or sub-word type). These word vector representations are used to capture the semantic relationships between words based on the context in which they appear, and they can be learned using a variety of approaches, including unsupervised learning with document statistics or self-supervised learning with a neural network model on a particular task.

Static word embedding methods, such as ***word2vec***, ***GloVe***, and ***fastText*** , have been widely used in natural language processing tasks and have been shown to produce high-quality word vectors that can capture the semantic relationships between words. In addition to their use in NLP, these methods have also been applied in other areas, such as information retrieval and recommendation systems, due to their ability to learn meaningful word vectors from large amounts of unannotated data. These methods have proven to be valuable tools for a variety of applications, and they continue to be an active area of research in the field of natural language processing. Below, we discuss some of these methods.

The word2vec approach, developed by Mikolov et al. (2013) in 2013 at Google, is a method for learning word vectors from unannotated text data using a feed-forward neural network-based model. The goal of word2vec is to learn meaningful word vectors that capture the semantic relationships between words. To achieve this, the model is trained to predict the context words (i.e., the words that occur in the vicinity of a target word) given the target word.

There are two different learning models that are commonly used as part of the word2vec approach: Continuous Bag-of-Words (CBOW) and Continuous Skip-gram. CBOW predicts a target word based on a window of context words, while Skip-gram uses a target word to predict a window of context words. Both models are based on the idea of using distributed representations of words, where each word is represented by a dense vector of real numbers. These models are shown in Figure 7.

The (Continuous bag-of-words, CBOW) model is a method for learning word vectors that involves predicting a target word based on a window of context words. Specifically, given a window of context words $W_t = w_{t-n}, \ldots, w_t, \ldots, w_{t+n}$ surrounding a target word w_t, the CBOW model aims to minimize the negative log-likelihood of the training data under the objective:

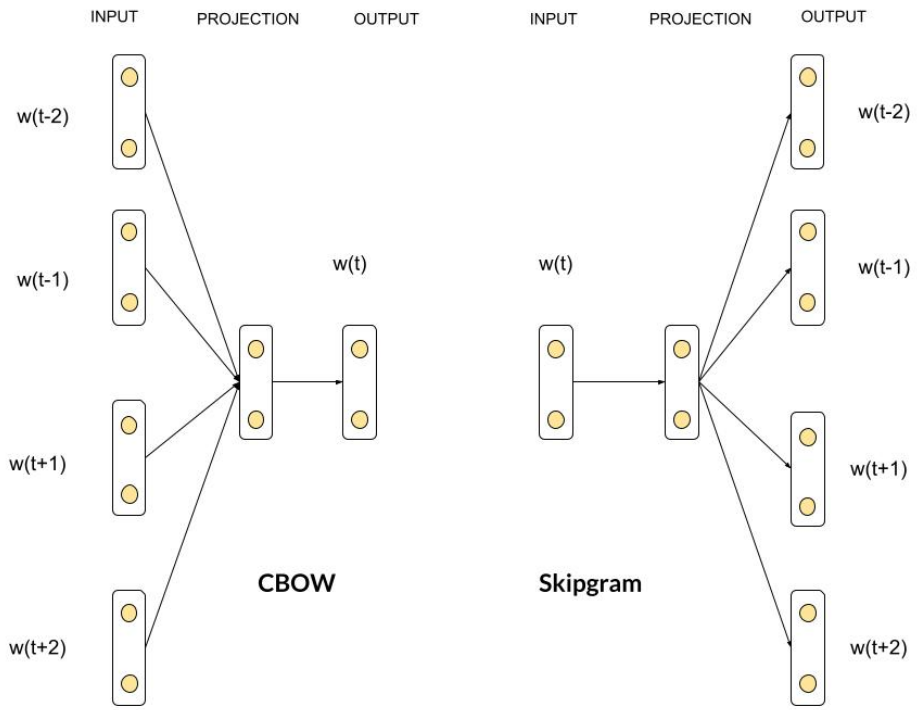

Figure 7: CBOW and Skip-gram model architectures.

$$\mathcal{L}_{CBOW} = -\log P(w_t|W_t)$$

where $P(w_t|W_t)$ is the probability of the target word w_t being the center word, conditioned on the context words W_t.

Intuitively, the Skip-gram with Negative Sampling (SGNS) model is a method for learning word vectors that can be viewed as the opposite of the Continuous Bag-of-Words (CBOW) model. While CBOW predicts a target word based on a window of context words, SGNS uses a target word to predict a window of context words. In SGNS, each word in a text is used as input to a neural network-based classifier to predict words in a certain range (referred to as a window) before and after the given word. However, the traditional Skip-gram model suffers from computational intractability and overfitting issues when dealing with large vocabularies.

Mikolov et al. (2013) introduced negative sampling to address these problems by reformulating the objective function. Instead of trying to maximize the probability of all context words given a target word, SGNS maximises the probability of true context words and minimizes the probability of negative samples. During training, for each target word, the model randomly selects a small number of negative samples from a noise distribution that is defined separately from the data. These negative samples are selected based on their low likelihood of co-occurring with the target word.

The objective of the SGNS model is then to maximize the probability of true context words and minimize the probability of negative samples. This is achieved by training a binary logistic regression classifier that takes as input the target word and a word representation and predicts whether the word is a true context word or a negative sample. The word representations are learned by maximizing the dot product between the word vector and the context vector for true context words and minimising the dot product for negative samples.

Given a sequence of words $w_{t-n}, \ldots, w_t, \ldots, w_{t+n}$ with a center word w_t, the SGNS model predicts a window of context words by minimising the Skip-gram objective:

$$\mathcal{L}_{skipgram} = \frac{1}{T} \sum_{t=1-n \leqslant j \leqslant n, j \neq 0}^{T} \log P(w_{t+j}|w_t)$$

where T represents the total number of terms in the sequence and n represents the size of the context before and after the center word.
In contrast to the shallow window-based method like *word2vec* which only uses local context windows, Pennington, Socher, and Manning (2014) proposed an analytical approach to learn efficient word representations (Global Vectors for Word Representation, GloVe) that can capture global relationships between words. In particular, this method is based on matrix factorization techniques on aggregated global word-to-word co-occurrence statistics from a corpus. Concretely, the semantic relationship of word w_i and word w_j is determined based on the ratio of their co-occurrences probabilities with a third context word w_k, expressed as $\frac{P(w_i, w_k)}{P(w_j, w_k)}$.

Bojanowski et al. (2017) proposed a method for generating word vectors or word embeddings that involves representing each word as the sum of the vectors of its character n-grams. Specifically, let a word w be represented as a sequence of character n-grams $n_1, n_2, \ldots, n_m$, where each character n-gram n_i is associated with a unique vector representation $\mathbf{v}_{n_i}$. Then, the word vector $\mathbf{v}_w$ for word w can be defined as:

$$\mathbf{v}_w = \sum_{i=1}^{m} \mathbf{v}_{n_i}$$

This approach allows the model to generate word vectors for out-of-vocabulary words i.e. words unseen during training by using the vectors of their individual character n-grams.

2.2.1.2 Contextualised Representations

Recent advances in NLP rely on contextual word embeddings. These embeddings use deep neural networks to incorporate the context of each word in a sentence in their

representations. Static word embeddings assign a single numerical representation to each token, which can be limiting in understanding the ambiguity of word meanings in different contexts. Contextualised embedding models, on the other hand, have the ability to capture context-dependent semantics, making them particularly useful for tasks that require an understanding of the contextual usage of words. These models are typically trained on large datasets using a self-supervised learning approach, and the resulting representations can be fine-tuned for specific tasks. One popular architecture for learning contextualised word embeddings is the Transformer architecture (Vaswani et al., 2017) which employs self-attention mechanisms to model contextual dependencies across an entire input sequence. A major breakthrough in contextual embeddings was ELMo (Embeddings from Language Model) (Peters et al., 2018), which generates word representations using a bidirectional LSTM(Hochreiter and Schmidhuber, 1997). ELMo captures hierarchical linguistic features across layers, producing dynamic embeddings that reflect entire sentence structures. A more powerful approach emerged with BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019), a Transformer-based model that learns contextual word representations through bidirectional pretraining. BERT is trained using two key objectives: masked language modeling (MLM) and next sentence prediction (NSP). In MLM, the model randomly masks a portion of the input tokens (typically 15%) and learns to predict the masked words based on surrounding context. This helps BERT capture bidirectional dependencies within a sentence. The NSP task is designed to improve the understanding of sentence relationships by BERT. During pretraining, the model is given two sentences and must determine whether the second sentence naturally follows the first. When processing a pair of sentences, BERT treats them as separate segments. Segment embeddings help distinguish tokens belonging to different sentences, enabling the model to correctly interpret sentence relationships. These segment embeddings are combined with word embeddings and position embeddings to form the final input representation for each token. For a given token, its input representation is constructed by adding the corresponding token, segment, and position embeddings. Additionally, BERT uses two special tokens that have specific meanings: the separator token ([SEP]) which indicates different segments of the input sequence and the sequence start token ([CLS]) which represents the whole input sequence, and accordingly, its final hidden state is used to predict for sequence classification tasks. A multilingual BERT (mBERT) (Devlin et al., 2019) extends this model by pretraining on Wikipedia texts from 104 languages, using a shared word-piece vocabulary. Unlike the original English BERT (EN-BERT), which is trained only on English data, mBERT is designed to support cross-lingual learning. Fine-tuning the BERT model has led to state-of-the-art results for various natural language processing tasks. Following BERT, OpenAI introduced the GPT (Generative Pre-trained Transformer) series (Radford et al., 2018), based on a decoder-only Transformer. Unlike BERT, which learns bidirec-

tional representations, GPT is autoregressive, predicting the next token sequentially. GPT-2 (Radford et al., 2019) achieved state-of-the-art results on benchmarks such as the WebText language modeling dataset and the BLiMP image captioning dataset. To improve efficiency, ALBERT (A Lite BERT) (Lan et al., 2019) introduced parameter reduction techniques, including factorized embedding parameterization, cross-layer parameter sharing, and Sentence Order Prediction (SOP), reducing model size while maintaining performance. For multilingual tasks, XLM (Cross-Lingual Language Model) Lample and Conneau (2019) uses three key components: Causal Language Modeling (CLM), which involves predicting the next token, similar to GPT; Masked Language Modeling (MLM), which is similar to BERT's pretraining objective; and Translation Language Modeling (TLM), a unique component in XLM that learns from aligned parallel texts across languages, improving cross-lingual transfer. This allows XLM to learn more general and transferable representations for words across different languages, making it better suited for cross-lingual tasks. Other optimizations include RoBERTa (Robustly Optimized BERT Approach) (Liu et al., 2019b), which removes NSP, trains with larger batches, and improves token masking strategies. Additionally, BART (Bidirectional and Auto-Regressive Transformer) (Lewis et al., 2020) combines denoising autoencoding with Transformer-based sequence generation, excelling at tasks like summarization and machine translation, and has obtained state-of-the-art results on benchmarks including the WMT machine translation dataset and the CNN/Daily Mail summarization dataset. Currently, NLP is experiencing a paradigm shift with large language models (LLMs), which scale Transformer-based architectures to trillions of tokens (Zhao et al., 2023). GPT-3 (Brown et al., 2020), with 175 billion parameters, introduced in-context learning, where models follow natural language instructions without additional training. FLAN (Wei et al., 2021) demonstrates the benefits of fine-tuning a pre-trained LLM with a 137 billion-parameter model based on a collection of formatted task instances. This approach enables the model to effectively tackle held-out tasks described in terms of text instructions.

Despite their impressive performance, LLMs often misunderstand human instructions, generate factually incorrect (hallucinated) information (Bender et al., 2021; Lin, Hilton, and Evans, 2021) or produce biased content (Ferrara, 2023). In response, the research community explores ways to align LLMs with human expectations (Wang et al., 2023). A noteworthy strategy, reinforcement learning from human feedback (Ouyang et al., 2022) stands out as a prominent approach post the success of ChatGPT[1]. This method includes learning from human preferences through a reward model, which is trained with human-rated outputs.

We employ contextualised representations (BERT, mBERT, XLM) in Chapters 3, 4 and 6.

1 https://chat.openai.com/

2.2.2 Evaluation

Since we employ word representations using different approaches throughout the dissertation, it is crucial to evaluate the quality of these representations. In recent years, a number of approaches to evaluate word representations in a vector space have been put forth. In this section, we outline the intrinsic evaluations for word representations that recur throughout this thesis and are essential to each chapter.

Intrinsic evaluation is a common method for evaluating the quality of word vectors, which are numerical representations of words that capture their meaning or semantics. In intrinsic evaluation, a set of word pairs is collected, and humans assign each pair a score indicating the similarity or relatedness of the words. The scores are then compared to the similarity or relatedness scores produced by the word vectors, allowing researchers to evaluate the ability of the word vectors to capture the meaning of words and the relationships between them.

Similarity refers to how much two words or concepts are alike or share common features in their meanings, making words that are similar in meaning tend to be interchangeable in certain contexts (e.g., car and vehicle). Relatedness, on the other hand, measures the association between words or concepts in meaning or context, where related words may not share similar meanings but are linked or frequently co-occur (e.g., scissors and paper). Some datasets, like RG-65 (Rubenstein and Goodenough, 1965), SimLex-999 (Hill, Reichart, and Korhonen, 2015) and WordSim (Finkelstein et al., 2001), focus on measuring similarity, while others, like MTurk-771 (Halawi et al., 2012), focus on measuring relatedness. In either case, intrinsic evaluation is a useful method for evaluating the quality of word vectors and understanding their ability to capture the meaning of words and the relationships between them. Table 1 provides an overview of these datasets.

When evaluating the similarity of word vectors or word embeddings, it is common to use cosine similarity as the evaluation metric. Cosine similarity is a measure of similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$ in a vector space, and it is calculated as:

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

where θ is the angle between vectors $\mathbf{a}$ and $\mathbf{b}$, and $|\mathbf{a}|$ and $|\mathbf{b}|$ are the magnitudes of vectors $\mathbf{a}$ and $\mathbf{b}$, respectively. A high cosine similarity indicates that the vectors are closely aligned, while a low or negative cosine similarity indicates that the vectors are not closely aligned.

After computing the cosine similarity between the word vectors, it is common to compare the similarity scores produced by the model with human-assigned similarity scores using a correlation coefficient like Spearman's rank correlation coefficient (ρ).

Dataset	Description
RG-65	65 pairs were evaluated for semantic similarity on a scale from 0 to 4.
SimLex-999	999 pairs evaluated on a scale from 0 to 10 for semantic similarity
WordSim-353 ALL	353 pairs were evaluated for semantic similarity on a scale from 0 to 10
WordSim-353 REL	252 pairs, a subset of WordSim-353 containing no pairs of similar concepts
MTurk-287	287 pairs were evaluated for semantic relatedness on a scale from 0 to 5
MTurk-771	771 pairs were evaluated for semantic relatedness on a scale from 0 to 5
WordSim-353	353 pairs were evaluated for semantic similarity on a scale from 0 to 10

Table 1: Datasets for evaluation of word similarity/relatedness

Spearman's rank correlation coefficient is a measure of the strength and direction of the relationship between two variables, and it is calculated as:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\text{cov}(X, Y)$ is the covariance between variables X and Y, and σ_X and σ_Y are the standard deviations of variables X and Y, respectively. A high Spearman's correlation coefficient indicates that the word vectors produced by the model are able to capture the similarity relationships between words as judged by humans, while a low Spearman's correlation coefficient indicates that the word vectors are not able to capture these relationships well. Beyond intrinsic evaluation methods like cosine similarity and correlation coefficients, some additional methods that are commonly used include:

Extrinsic evaluation: This involves evaluating the performance of the word vectors on downstream natural language processing tasks, like language translation, text classification, or question answering. By integrating the word vectors into task-specific models and measuring their impact on performance, researchers can determine their practical utility and generalization ability. Common evaluation metrics include accuracy, F1-score, BLEU (for translation), and perplexity (for language modeling). Strong performance in these tasks indicates that the embeddings effectively capture meaningful linguistic patterns.

Human evaluation: Instead of using automated metrics like cosine similarity or correlation coefficients, researchers can also directly ask human annotators to evaluate

the quality of the word vectors. This process includes rating word pairs for similarity and relatedness, evaluating the contextual accuracy of embeddings—particularly for polysemous words—and performing sentence completion or association tasks to test logical and linguistic coherence. Human evaluation is especially useful in detecting subtleties, biases, and inconsistencies that automated methods might overlook.

Visualization: Another effective approach to evaluating word vectors is through dimensionality reduction techniques that map high-dimensional word embeddings into a lower-dimensional space, making their relationships more interpretable. Methods like t-SNE (t-distributed stochastic neighbor embedding), PCA (principal component analysis), and UMAP (Uniform Manifold Approximation and Projection) help in understanding the structure of embeddings by revealing clusters of semantically similar words. These techniques provide insights into how well embeddings capture linguistic regularities, synonymy, antonymy, and word associations.

The word similarity evaluation recurs throughout the dissertation in all chapters and is integral to it.

3

Multilingual Translationese Classification

This chapter demonstrates how to distinguish between translated and non-translated (original) texts. We start with a stylistic analysis of the two, and we then investigate the effectiveness of representation learning-based methods in mono- and multilingual translationese classification. Traditional methods for translationese classification rely on manually selecting potentially partial, non-exhaustive linguistic features and often require linguistic annotation tools. In contrast, our approach involves developing a suite of representation-learning methods, eliminating the need for manual feature engineering. We perform our experiments across a range of source-target language combinations in mono- and multilingual settings and demonstrate that representation learning-based approaches outperform the traditional feature engineering methods for all tasks.

3.1 Introduction

Translationese classification is a well-studied area in modern computational Translation Studies often drawing on work in computational stylometry. Previous research (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Rabinovich and Wintner, 2015; Volansky, Ordan, and Wintner, 2015) has shown that there are discernible differences between original texts and their translated versions, and identifying these differences using supervised classification has been a major area of investigation. Many of the classifiers presented in previous work make use of features that are based, at least in part, on translational language hypotheses (simplification, explicitation, levelling out, interference, and normalization), as detailed in Section 2.1.1. However, while some of these features are directly tied to a particular hypothesis, others are not, and many features can belong to multiple hypotheses.

The conventional approach to identifying translationese has been to use supervised machine learning algorithms (mostly classification) with hand-crafted features. These features include character n-grams, part-of-speech (POS), function words, discourse markers, etc. and represent potentially important measures of translationese. However,

there are a number of reasons why relying on hand-crafted linguistic features for supervised translationese classification can be problematic:

> Dependence on expert knowledge: Hand-crafted features are frequently created by domain experts who have in-depth knowledge of linguistics and the distinctive characteristics of the translated text. This can be a time-consuming and expensive process, and thus it might be difficult to create features for every possible language and domain.
>
> Non-Exhaustiveness: Manually designed features may be partial and nonexhaustive in a sense that they are based on our linguistic theories and intuitions, and thus may not be guaranteed to capture all important characteristics of the input data with respect to translationese seen during training.
>
> Lack of Scalability and Generalisation: Hand-crafted features are particularly specific to the task and language for which they were extracted, which can restrict the scalability of the classifier. When tested on different languages or domains than the one it was originally trained on, the classifier might not perform well due to the features being too specific. This lack of adaptability makes it challenging to use the classifier with new data or in different languages or domains without substantial modifications.

The last problem mentioned in the previous paragraph is not unique to translationese classification but can occur in other machine learning tasks as well. However, they are also relevant to the task of translationese classification. In view of these limitations, we believe there is clearly a need to classify translationese without the use of hand-crafted linguistic features.

With this, we formulate our first research question:

RQ1: To what extent can representation learning techniques, such as embeddings, discern translated and non-translated texts, without prior linguistic assumptions?

To answer this, we investigate the effectiveness of representation learning-based methods in mono- and multilingual translationese classification. We develop a suite of representation-learning methods based on word embeddings, eliminating the need for manual feature engineering. Some of these representations are static and do not require further training, while others may undergo fine-tuning during the training process. We find that our embedding-based approaches are a superior alternative to the previous hand-crafted linguistically inspired methods as measured by classification accuracies. Further, we compare our approaches to current state-of-the-art neural representation learning models (fine-tuned BERT and LSTM).

The remainder of this chapter is structured as follows. In Section 3.2 we describe related work on translationese classification. We profile our our data set (Amponsah-Kaakyire et al., 2021) in Section 3.3. In Section 3.5 we define our experimental settings and analyse the results in Section 3.6. Finally, we summarise the main findings in Section 3.7.

3.2 Related Work

Previous studies (Baroni and Bernardini, 2005; Ilisei et al., 2010; Koppel and Ordan, 2011; Volansky, Ordan, and Wintner, 2015) show that translated texts differ from their original counterparts to the extent that they can be accurately identified by means of automatic classification. The seminal work of Baroni and Bernardini (2005) marked the beginning of the machine learning (ML) paradigm in translationese research. Baroni and Bernardini (2005) relied on the use of Support Vector Machines (SVM) with n-gram features for the task of classifying Italian texts as translated or non-translated. According to their results, the SVMs were highly dependent on lexical cues, the distribution of function word n-grams, and morpho-syntactic categories in general. As a result, it was possible to automatically distinguish professional translations from non-translated texts using only shallow features. According to their analysis of bi-grams, they conclude that translations tend to reuse topic-dependent sequences and structural patterns, but non-translated texts tend to reuse topic-independent sequences. Another key finding from their study was that an ML algorithm (SVMs) could accurately distinguish between translations and non-translations even when professional translators were unable to do so.

Ilisei et al. (2010) employed supervised machine learning classifiers, trained with 21 simplification universal features such as average sentence length, type-token ratio, etc., to differentiate between original texts and human translated texts. Their dataset focused on medical and technical domains and involved comparisons between translated texts by students and professionals, and originals, demonstrating that the Support Vector Machine (SVM) classifier achieved a notable 97.62% accuracy.

Using function word frequencies to represent chunks of text, Koppel and Ordan (2011) used a Bayesian logistic regression approach on the Europarl corpus (Koehn, 2005) to automatically identify translated and non-translated texts. Their high accuracy results based on the Europarl corpus lend support to the translationese and interference hypotheses. Their findings also indicate that the classification accuracy of translationese in language L1 using a classifier trained on L2 data is higher if languages L1 and L2 are typologically close, such as two languages like French and Spanish coming from the same Romance language family.

Volansky, Ordan, and Wintner (2015) proposed and explored an extensive set of translationese features based on translation hypotheses outlined in Section 2.1.1 to

differentiate translationese from original text. In support of the *Interference* hypothesis, the analysis involved lexical features, including part-of-speech, character uni-grams, bi-grams, and tri-grams. Except for character unigrams, which exhibited an accuracy rate of only 85%, all others yielded results surpassing 90% accuracy in translationese classification. Beyond the lexical features tied to the *Interference* property, cohesive markers, serving as an *Explicitation* feature, achieved 81% accuracy; threshold point-wise mutual information (PMI) as a *Normalisation* feature achieved 66% accuracy, and mean word rank, a *Simplification* feature, achieved 77% accuracy.

Some earlier seminal work also explored simple lexicalised features including word tokens and character n-grams (Avner, Ordan, and Wintner, 2016; Popescu, 2011), part-of-speech tags or lemmas (Baroni and Bernardini, 2005; Halteren, 2008; Kurokawa, Goutte, and Isabelle, 2009), or features based on count models, information density, surprisal, and complexity for text classification, especially in translation quality estimation (Rubino, Lapshinova-Koltunski, and Genabith, 2016). These features serve as indicators of translationese for both originally authored and manually translated texts (Rubino, Lapshinova-Koltunski, and Genabith, 2016). However, this research relied on discrete count-based approaches that treat words as discrete units, leading to limited context modelling capabilities. However, until recently, neural approaches to translationese have received much less attention. To the best of our knowledge, the only work is that of Sominsky and Wintner (2019) who used a BiLSTM-based architecture to automatically detect the translation direction of each bilingual parallel (bitext) training sample, where the network is presented with two parallel sentences simultaneously and has to determine which sentence is original and which is a translation. They report accuracy up to 81.0% on the Europarl corpus.

More recently, Riley et al. (2020b) used a CNN classifier to differentiate translated from original target text, and then used this classifier to tag the training data for an NMT model to produce translations that look like the original text. Human evaluation demonstrated that this produced more accurate and natural translations and showed fewer translationese effects.

3.3 Multilingual Data Analysis

In this section, we employ a portfolio of well-known measures from stylometry, including both surface and information-theoretic features, as well as readability scores to linguistically profile the data we use in our classification experments. Our aim is to uncover observable stylistic differences between our original and translated data, setting the context in which we aproach **RQ1**.

Our stylometric analysis highlights several key points. Firstly, it indicates distinct stylistic aspects between original and translated texts. Secondly, we show that stylistic variation exists between texts translated into the same target language but from

Corpus	Train	Dev	Test
TRG–SRC	30,000	6,000	6,000
TRG–ALL	30,000	6,000	6,000
ALL–ALL[3]	89,000	19,000	19,000
ALL–ALL[8]	67,000	14,000	14,000

Table 2: The total number of paragraphs in each dataset. On average there are about 80 tokens each paragraph.

different source languages. Lastly, this study contributes to extended data analysis by providing insights into the unique stylistic characteristics of the multilingual datasets used to address **RQ1**.
In the following sections, we provide a brief description of the data we used, the measures we employed to analyse the data, and the results of our analysis.

3.3.1 Datasets

We utilise the Multilingual Parallel Direct Europarl (MPDE) corpus, as described by Amponsah-Kaakyire et al. (2021). The MPDE is a parallel corpus sourced from a subset of European Parliament proceedings. This corpus includes annotations indicating which sections are original and which are translated, presented in parallel paragraphs[1]. For more details, refer to the MPDE repository[2]. Annotations identify the source (SRC) and target languages (TRG) with the term "original" or "translationese".

For our task, we use the portion of MPDE that only provides translations of originally authored texts as texts already translated from another language may differ from directly translated texts (i.e. where the source is originally authored) in some ways. We focus on three European languages for the initial experiment: German (DE), English (EN), Spanish (ES). Subsequently, we expand the scope to include Greek (EL), French (FR), Italian (IT), Dutch (NL), and Portuguese (PT) in our multilingual multi-source classification experiments. In the following, we use the "TRG–SRC" notation (with a dash) to identify monolingual translationese corpora (i.e. corpora where half of the data is originals, half translationese): TRG is the language of the corpus, SRC is the source language, from which the translation into the TRG language was done in order to produce the translationese half. The "TRG←SRC" notation (with an arrow) denotes the result of translating a text from SRC into TRG language. It is used to refer to the translationese half of the monolingual corpus.

For our analysis, we create four data sets with statistics summarised in Table 2:

1 This results from the fact that the translations of paragraphs are not aligned sentence-wise. While the original paragraph may have i sentences, one translation may have j sentences, and another k.
2 github.com/UDS-SFB-B6-Datasets/Multilingual-Parallel-Direct-Europarl

Corpus	Tokens	Types	%Digits	%Punct.	%FW	Sent.len	Tok.len	Syl
EN – DE	212140	412	0.08	0.12	1.54	7.37	0.11	484219
EN – ES	48405	6670	0.08	0.13	0.21	3.28	0.06	54003
ES – EN	126091	3820	0.05	0.43	0.36	8.54	0.03	780776
ES – DE	378489	6160	0.01	0.41	0.49	15.47	0.03	780766
DE – EN	187025	21851	0.03	0.18	2.49	12.67	0.17	268801
DE – ES	277916	16211	0.02	0.11	2.14	18.82	0.08	498169
EN – ALL	134262	1329	0.08	0.13	0.84	8.1	0	188846
ES – ALL	252935	7925	0.03	0.41	0.42	17.17	0.03	524599
DE – ALL	229160	12567	0.1	0.16	2.35	15.69	0.13	389825

Table 3: Stylistic **absolute differences** in text statistics measures between original and translated halves. White cells indicate an increase in the original split, while teal cells indicate a decrease. White cells denote no significant difference.

Monolingual single-source data: DE–EN, DE–ES, EN–DE, EN–ES, ES–DE, ES–EN. For each corpus, there is an equal number of translated and original paragraphs in the same language.

Monolingual multi-source data: DE–ALL, EN–ALL, ES–ALL. For DE–ALL, e.g., half of the data is DE original texts, and the other half contains equal proportions of DE←ES and DE←EN.

Multilingual multi-source data: ALL–ALL[3]. There is an equal number of originals: DE, EN and ES, which together make up 50% of the examples. The other 50% which are translated are equal proportions of DE←EN, DE←ES, EN←DE, EN←ES, ES←DE and ES←EN.

DE, EN and ES are relatively close typologically. We conduct additional experiments in order to investigate how well classification can be performed when more and more distant languages are involved:

Multilingual multi-source data large: ALL–ALL[8], balanced in the same way as ALL–ALL[3], but with the addition of Greek (EL), French (FR), Italian (IT), Dutch (NL) and Portuguese (PT).

3.3.2 Measures

This section describes the stylometric features based on text statistics, richness (information) and readability measures used in our study.

Text Statistics Measures.

Corpus	Lexical Density	TTR	Legomena	Dislegomena	Yule-I	GF	WS	FHR
EN – DE	0.01	0.46	0.19	0.07	14.54	1.99	-	-
EN – ES	0.02	0.57	0.17	0.04	16.16	1.15	-	-
ES – EN	0.01	0.14	0.02	0.01	4.25	-	-	5.024
ES – DE	0.01	0.27	0.19	0.06	3.47	-	-	5.032
DE – EN	0.02	0.06	0.07	0.01	2.64	-	4.12	-
DE – ES	0.02	0.12	0.22	0.02	6.08	-	9.00	-
DE – ALL	0.02	0.24	0.07	0.05	7.15	-	0.22	-
ES – ALL	0.01	0.04	0.03	0.02	2.22	-	-	5.007
EN – ALL	0.02	0.29	0.1	0.05	13.19	1.94	-	-

Table 4: Stylistic **absolute differences** in text richness measures between original and translated halves. White cells indicate an increase in the original split, while teal cells indicate a decrease. White cells denote no significant difference.

Tokens: Number of words in text.

Types: Number of different words in text.

Digits (%digits): The percentage of digits in a text.

Punctuations (%punct.): The percentage of punctuation symbols in a text.

Function Words (% FW): The percentage of function words in a text.

Sentence Length (sent.len): This measures the number of tokens in a sentence.

Token Length (tok.len): The average token length in terms of characters.

Syllables (syl.): Number of syllables in a text.

Text Richness Measures.

Lexical Density: Lexical Density is described in Toral (2019) as the proportion of content words (adverbs, adjectives, nouns and verbs) to total number of words. This is also described as *information load* in Ilisei et al. (2010).

Type-Token Ratio (TTR): This measures the ratio of the vocabulary (types) to the sample size (tokens). Higher value of TTR indicates higher degree of lexical variety. For ease of readability and comparison, we multiplied TTR scores by 500.

Hapax Legomenon: The percentage of words with frequency 1 in the text segment.

Hapax DisLegomenon: The percentage of words with frequency 2 in the text segment.

Yule-K (Yule, 2014): Vocabulary richness index that exhibits stability in different text sizes.

$$K = 10^4 \times [\frac{[\sum_{X=1}^{X} f_x X^2] - N}{N^2}] \quad (1)$$

where, X is vector of frequencies of each type; N is the number of tokens and f_x represents frequencies for each x. That is, it indicates how constant the vocabulary is across the text, with higher values suggesting more repetitive vocabulary. In practice, we use the inverse of Yule-K- Yule's I measure, which is considered to be a more robust measure of vocabulary diversity because they are less affected by changes in text length[3].

Text Readability Measures.

Gunning Fog: The Gunning fog index (Gunning et al., 1952) is a measure of text complexity of reading English-language texts that are based on the average number of complex words (those with three or more syllables) and the average number of long sentences in a text. The Gunning Fog index (GF) is calculated using the formula:

$$GF = 0.4 \times \left[(\frac{w}{s}) + 100 \times (\frac{c}{w}) \right]$$

where w is the number of words, s is the number of sentences, and c is the number of complex words (defined as words containing more than 2 syllables).

The higher the GF index, the more difficult the text is to understand. Generally, a score of 8 or lower indicates that the text is easily understandable by an average 11-year-old student, while a score of 9-10 indicates that the text is understandable by an average 13- to 15-year-old student. A score of 11-12 indicates that the text is understandable by an average 16- to 18-year-old student, while a score of 13 or higher indicates that the text is generally only understandable by college graduates.

Wiener Sachtext formula: The Wiener Sachtextformel (Bamberger and Vanacek, 1984) is a readability formula used to measure the difficulty of reading German-language texts. The formula takes into account the average sentence length and average number of syllables per word in the text and produces a score that indicates how easy or difficult the text is to read. The Wiener Sachtext formula is calculated as follows:

$$WS = 180 - (asl \times 1.2) - (asw \times 56)$$

3 Scores are multiplied by 100 for ease of readability.

where asl is the average sentence length in words, and asw is the average number of syllables per word. A higher score indicates easier readability, while a lower score indicates more difficult readability. The maximum score is 100, which indicates the text is very easy to read, and the minimum score is 0, which indicates the text is very difficult to read.

Fernández-Huerta Readability Formula: The Fernández-Huerta (Fernández Huerta et al., 1983) readability formula is a method of measuring the readability of Spanish language texts. It calculates the readability score based on the number of syllables per word and the number of words per sentence.

The Fernández-Huerta Readability Formula is calculated as follows:

$$\text{FHR} = 206.84 - (0.6 \times \frac{sy}{w}) - (1.02 \times \frac{w}{s})$$

where sy is the total number of syllables in the text, w is the total number of words in the text, and s is the total number of sentences in the text. The Fernandez-Huerta index ranges from 0 to approximately 206, with higher scores indicating an easier-to-read text.

3.3.3 Results

The result of the absolute differences in stylometric measures between original and translated halves are shown in Tables 3 and 4. Table 3 shows that the number of tokens and average sentence length varies depending on the target language regardless of the source language. Translated text into Spanish has longer sentences than original Spanish text, indicating that human translation creates longer sentences compared to the originals. In contrast, translations into German have shorter average sentence lengths than German originals, implying that translators conveyed the same message more efficiently in the target language. However, no clear trend is observed in English translations. Sentence splitting is common in translation and is seen as a form of simplification, where longer and more complex statements are divided into shorter, simpler ones. Thus, variations in sentence length in the same target language are observed depending on the specific source language from which the translations were made. Table 3 further shows that there are some differences in the use of punctuation between original and translated texts in our MPDE dataset. Punctuation marks are used to structure information within sentence boundaries, significantly reducing ambiguity. The explicitation hypothesis posits that translated texts are characterized by reduced ambiguity. However, from Table 3, we find that this tendency is only reflected in translations into English.

Another indicator of translationese is the ratio of syllables. Previous studies have shown that translated texts tend to have longer words and sentences compared

to original texts in the target language. Based on our findings, it appears that the number of syllables is consistently greater in translated texts for translations into English and Spanish, regardless of the source language. However, there is no clear trend observed for translations into German. Translationese is also manifested in the inflated frequencies of function words (Volansky, Ordan, and Wintner, 2015). Table 3 shows that the percentage of function words is generally higher in translated texts as compared to original texts for translations into English, while the percentage of function words is higher in the original text for translations into German and Spanish.

In Table 4 we measure lexical density as the percentage of content words. Previous research has demonstrated that translated texts have a lower percentage of content words (adverbs, adjectives, nouns and verbs) compared to original texts, indicating that they are lexically simpler than the source text (Scarpa, 2006). This trend has also been observed in a study by Toral (2019), where they discovered that machine-translated outputs have lower lexical diversity than both human translations and the original text in the same target language. Our results align with these findings, showing that, regardless of monolingual single-source or multi-source data, original texts consistently demonstrate higher lexical density than their translated counterparts.

The type-token ratio has been used in previous studies to measure the lexical variety in translated texts (Riley et al., 2020b; Toral, 2019). Our results indicate that original texts have a greater richness of vocabulary across all language pairs. Another characteristic of translated texts is that they tend to exhibit a substantially lower rate of words that occur only once (hapax legomena) or twice (hapax dislegomena) in a text than original texts (De Clercq et al., 2021). We see the same in our results. Table 4 shows that originals always exhibit more hapax legomena or dislegomena than their translated counterparts, indicating higher lexical diversity in the original texts.

Finally, we also analyse readability scores between originals and translated texts using language-dependent readability measures. For English, we use Gunning-Fog (G.Fog), for German we use Wiener Sachtext (W.Sachtext) and for Spanish, we use the Fernandez-Huerta (F.Huerta) formula. For German texts, we observe that in all cases, originals achieve lowers scores than their translated counterparts, implying that the original texts are easier to read than translated ones. This indicates that translationese features tend to retain source language interferences which might hamper the overall readability of the text. However, for Spanish texts, we see that translated texts show higher scores, indicating easier readability. In case of English, no distinct pattern is identified in this regard.

Overall, this analysis revealed a stylistic disparity between original and human-translated texts. We find that there is significant variation in the languages used within each label, original and translated. In fact, we discover that stylistic results differ substantially for specific language pairs, and even for the same target language with different source languages. This highlights the need for a language-agnostic

approach to translationese classification. Based on the findings of the stylometric analysis, especially those related to semantics (vocabulary richness and diversity), we extend previous studies on translationese classification based on classical feature engineering methods (Ilisei et al., 2010; Koppel and Ordan, 2011; Volansky, Ordan, and Wintner, 2015) by using embedding-based representation learning approaches. We briefly outline the classification task and classical feature-engineering as well as representation learning-based models in the section that follows.

3.4 Multilingual Translationese Classification

For all datasets introduced in Section 3.3.1, we perform translationese classification on tasks in the following settings:

Single-source and Multi-source: We perform binary classification – original vs. translated, where for a given target language, the model is trained to detect translationese from monolingual single-source e.g. TRG–SRC_1 and TRG–SRC_2, or is translations from multi-source e.g. TRG–ALL.

Multilingual multi-source and Cross-language: We evaluate the models trained on one dataset on the other ones, in order to verify how well the model trained to detect translationese in multilingual data performs on monolingual data: ALL–ALL[3] on TRG–ALL, and ALL–ALL[3] on TRG–SRC.

3.5 Experimental Setup

3.5.1 Model Specifications

For the tasks defined above, we employ the following models:

3.5.1.1 Feature-Engineering-Based Classification

We draw inspiration from the feature set presented in (Rubino, Lapshinova-Koltunski, and Genabith, 2016) based on translationese properties and extract our selected features using the *INFODENS* toolkit (Taie, Rubino, and Genabith, 2018) to train and evaluate a classifier. We apply a support vector machine classifier (SVM) with a linear kernel and fit the hyperparameter C on the validation set. The results are labeled as `Handcr.+SVM` throughout this chapter. In particular, we use:

surface features: average word length, syllable ratio, paragraph length.
lexical features: lexical density, type-token ratio.
unigram bag-of-PoS

language modelling features: log probabilities and perplexities with and without considering the end-of-sentence token, according to forward and backward n-gram language models ($n \in [1;5]$) built on tokens and POS tags.

n-gram frequency distribution features: percentages of n-grams in the paragraph occurring in each quartile ($n \in [1;5]$).

Our experiments use the training data to build language models and n-gram frequency distributions. POS-tagging is performed using SpaCy[4]. The n-gram language models are calculated using SRILM (Stolcke, 2002). Features are scaled based on their highest absolute values. Note that all our features are delexicalised.

3.5.1.2 Embedding-Based Classification

Previous research (Lebret and Collobert, 2015; Mitchell and Lapata, 2008) employ distributed representations of words to address a variety of NLP tasks. The straightforward approach simply adds up the word vector representations in a sentence or document to produce sentence or document-level representations which are then used for the downstream applications. However, for translationese classification tasks, these representation-based approaches are yet to be explored. Our primary goal is twofold: (i) firstly, there is no guarantee that hand-crafted feature engineering-based approaches really capture everything relevant to the classification at stake; and (ii) explore whether performance can be enhanced beyond (i).

In the next section, we present several methods based on vector representations for classifying original and translated data that are not biased towards particular linguistic features as the `SVM` model presented in Section 3.5.1.1. All these models use semantic word representations.

Average pre-trained embeddings + SVM (Wiki+SVM). In this approach, we compute an average of all token vectors in the paragraph to be classified and use this mean vector as a feature vector to train a SVM classifier with linear kernel. We work with the publicly available language specific 300-dimensional pre-trained Wiki word vector models trained on Wikipedia using ***fastText***[5] (Joulin et al., 2016).

Gaussian distributions for similarity-based classification (Wiki+Gauss.+SVM). Here, we introduce a method that represent words not as vector points but as continuous densities in latent space. In particular, we explore Gaussian function embeddings (with diagonal covariance), in which both means and variances are learned from data. We follow Das, Zaheer, and Dyer (2015) and Nikolentzos et al. (2017), Gourru, Velcin, and Jacques (2020) and represent a text as

4 https://spacy.io/
5 https://fasttext.cc/docs/en/pretrained-vectors.html

a multivariate Gaussian distribution based on the distributed representations of its words.

The method assumes that each word w is a sample drawn from a Gaussian distribution with mean vector μ and covariance matrix σ^2:

$$w \sim \mathcal{N}(\mu, \sigma^2) \tag{2}$$

A paragraph is then characterized by the average of its words and their covariance. More specifically, let the mean vector of the two texts be μ_i and μ_j and their covariance matrices be σ_i^2 an σ_j^2, then the similarity between these mean vectors can be calculated using cosine similarity between them, i.e.,

$$sim(\mu_i, \mu_j) = \frac{\mu_i . \mu_j}{||\mu_i||||\mu_j||} \tag{3}$$

Similarly, the similarity between the co-variance matrices of the two texts can be computed using the element-wise product between matrices:

$$sim(\sigma_i^2, \sigma_j^2) = \frac{\Sigma \sigma_i^2 \circ \sigma_j^2}{\|\sigma_i^2\|_F \times \|\sigma_j^2\|_F} \tag{4}$$

where $(\cdot \circ \cdot)$ indicates the Hadamard or element-wise product between the matrices with summation over all elements and $\|\cdot\|_F$ is the Frobenius norm for matrices.

Finally, the similarity between texts is represented by the convex combination of the similarities of their mean vectors μ_i and μ_j and their covariances matrices σ_i^2 and σ_j^2:

$$similarity = \alpha(sim(\mu_i, \mu_j)) + (1 - \alpha)(sim(\sigma_i^2, \sigma_j^2)) \tag{5}$$

where $\alpha \in [0,1]$. Finally, a SVM classifier is employed using the kernel matrices of Equation 5 to perform the classification where the kernel represents similarities between pairs of texts. We work with the same pre-trained Wikipedia ***fastText*** embeddings as in `Wiki+SVM` for the words in the model and initialize the ones not contained in the model to random vectors.

fastText* classifier (FT).** As a first neural classifier, we use ***fastText (Joulin et al., 2016) which is an efficient neural network model with a single hidden layer. The ***fastText*** model represents texts as a bag of tokens and bag of n-gram tokens. Embeddings are averaged to form the final feature vector. A linear transformation is applied before a hierarchical softmax function to calculate the class probabilities. Word vectors are trained from scratch on our data (refer to Section 3.3.1).

Corpus	Handcr. +SVM	Wiki +SVM	Wiki +Gauss. +SVM	fastText (FT)	Wiki +FT	LSTM	BERT
DE–EN	71.5±0.0	77.7±0.1	67.6±0.1	88.4±0.0	89.2±0.0	89.5±0.4	**92.4±0.2**
DE–ES	76.2±0.0	79.4±0.3	68.2±0.2	90.9±0.0	91.9±0.0	91.9±0.2	**94.4±0.1**
EN–DE	67.6±0.7	72.5±0.2	64.5±0.2	85.1±0.0	85.9±0.1	86.8±0.5	**90.7±0.1**
EN–ES	70.1±0.2	77.5±0.4	67.1±0.4	87.6±0.0	88.7±0.0	89.1±0.3	**91.9±0.4**
ES–DE	71.0±0.0	75.7±0.4	70.1±0.4	88.4±0.0	89.1±0.0	90.2±0.2	**92.3±0.2**
ES–EN	66.7±0.0	70.1±0.3	67.0±0.7	87.0±0.1	87.9±0.0	88.8±0.4	**91.4±0.3**
DE–ALL	72.6±0.0	64.3±0.0	65.1±0.1	87.4±0.0	88.3±0.0	88.5±0.2	**90.9±0.3**
EN–ALL	65.3±0.0	64.6±0.0	62.5±0.1	82.7±0.0	84.4±0.0	84.2±0.4	**87.9±0.4**
ES–ALL	67.4±0.0	67.3±0.0	66.5±0.2	84.9±0.0	85.9±0.0	87.0±0.3	**89.9±0.1**
ALL–ALL[3]	58.9±0.0	–	–	85.0±0.0	–	84.4±0.3	**89.6±0.2**
ALL–ALL[8]	65.4±0.1	–	–	70.4±0.1	–	77.2±0.3	**84.6±0.2**

Table 5: Translationese classification average accuracy on the mono- and multilingual test sets (average and standard deviation over 5 runs).

Pre-trained embeddings + FT (Wiki+FT). In this model we work with the pre-trained word vectors from Wikipedia to initialize the ***fastText*** classifier. The data setting is directly comparable to `Wiki+SVM`, a non-neural classifier.

3.5.1.3 Deep learning based Classification

We use two other neural models for the classification task: an LSTM and a transformer. While the LSTM is trained from scratch, we fine-tune a pre-trained BERT model (Devlin et al., 2019) on our training set.

LSTM. We train a single-layer uni-directional LSTM with 128 embedding and hidden dimensions. We then take all the hidden states using average pooling, and pass the resulting 128-dimensional representation to a binary linear classifier. We use a batch size of 16, learning rate of $1 \cdot 10^{-3}$, and the Adam optimiser with *Pytorch* defaults.

Bidirectional Encoder Representations from Transformers (BERT). We use the BERT-base multilingual uncased model (12 layers, 768 hidden dimensions, 12 attention heads) (Devlin et al., 2019). Fine-tuning is done with the *simpletransformers*[6] library. For this, the representation of the [CLS] token goes through a pooler, where it is linearly projected, and a `tanh` activation is applied. Afterwards, it undergoes dropout with a probability of 0.1 and is fed into a binary linear classifier. We use a batch size of 32, a learning rate of $4 \cdot 10^{-5}$, and the Adam optimiser with epsilon $1 \cdot 10^{-8}$. Models were fine-tuned on 4 GTX1080Ti GPUs.

6 github.com/ThilinaRajapakse/simpletransformers

	DE–EN	DE–ES	EN–DE	EN–ES	ES–DE	ES–EN	DE–ALL	EN–ALL	ES–ALL
DE–EN	92.4±0.2	76.6±0.7	-	-	-	-	90.5±0.3	-	-
DE–ES	82.6±1.1	94.4±0.1	-	-	-	-	91.8±0.4	-	-
EN–DE	-	-	90.7±0.1	64.7±1.4	-	-	-	87.3±0.4	-
EN–ES	-	-	72.9±0.9	91.9±0.4	-	-	-	88.6±0.4	-
ES–DE	-	-	-	-	92.3±0.2	78.8±0.9	-	-	90.6±0.1
ES–EN	-	-	-	-	78.8±1.6	91.4±0.3	-	-	89.0±0.2
DE–ALL	87.3±0.6	85.3±0.4	-	-	-	-	90.9±0.3	-	-
EN–ALL	-	-	81.7±0.5	78.3±0.7	-	-	-	87.9±0.4	-
ES–ALL	-	-	-	-	85.9±0.9	85.0±0.6	-	-	89.9±0.1

Table 6: BERT translationese classification accuracy of all TRG–SRC and TRG–ALL models on TRG–SRC and TRG–ALL test sets (average and standard deviation over 5 runs). Columns: training set; rows: test set.

3.6 Evaluation and Analysis

Paragraph-level translationese classification results in Single-source and Multi-source settings with mean and standard deviations over 5 runs are reported in Table 5. We observe that all our embedding-based approaches, except for `Wiki+Gauss.+SVM` perform better than `Handcr.+SVM`. This answers **RQ1** that is — can word embedding based feature and representation leaning approaches discern translated and non-translated texts? Among the approaches with the SVM classifier, `Wiki+SVM` performs best in the single-source settings, but shows a lower accuracy than `Handcr.+SVM` in the multi-source (TRG–ALL) settings. `Wiki+Gauss.+SVM` performs worst apart from on ES–EN and ES–ALL.

In the monolingual single-source settings, we observe that overall accuracy is slightly lower when the source language of a translation is typologically closer to the original text language, i.e. it becomes more difficult to detect translationese. Specifically, DE–EN tends to have lower accuracy than DE–ES; EN–DE lower accuracy than EN–ES; and ES–EN lower accuracy than ES–DE. Accuracy generally drops when going from single-source to the multi-source setting, e.g. from DE–EN and DE–ES to DE–ALL. The EN–ALL dataset is the most difficult for most of the models among the TRG–ALL datasets. The ALL–ALL[3] setting exhibits comparable accuracy to the TRG–ALL setting for the neural models, but for the SVM there is a drop of around 9 points. Throughout our discussion, we always report absolute differences between systems. The ALL–ALL[8] data results in reduced accuracy for most architectures, except `Handcr.+SVM`. Using the pre-trained Wiki embeddings helps improve the accuracy of the ***fastText*** method in all cases.

Overall, the `BERT` model outperforms other architectures in all settings, followed closely by the other deep-learning architecture, LSTM. Neural-classifier-based models substantially outperform the other architectures: the SVMs trained with hand-crafted linguistically-inspired features, e.g., trail BERT by ~20 accuracy points.

3.6.0.1 Multilinguality and Cross-Language Performance

Since neural architectures perform better than the non-neural ones, we perform the multilingual and cross-language analysis only with the neural models. We evaluate the models trained on one dataset on the other ones, in order to verify:

- Whether for a given target language, the model trained to detect translationese from one source language, can detect translationese from another source language: TRG–SRC_1 on TRG–SRC_2, and TRG–SRC on TRG–ALL;
- How well the model trained to detect translationese from multiple source languages can detect translationese from a single source language: TRG–ALL on TRG–SRC, and ALL–ALL[3] on TRG–SRC;
- How well the model trained to detect translationese in multilingual data performs on monolingual data: ALL–ALL[3] on TRG–ALL, and ALL–ALL[3] on TRG–SRC.

Table 6 shows the results of cross-data testing for the monolingual models for the best-performing architecture (BERT). For the single-source monolingual models, we observe a relatively smaller drop (up to 13 percentage points) in the worst configuration in performance when testing TRG–SRC on TRG–ALL (as compared to testing TRG–SRC on TRG–SRC), and a larger drop (up to 27 points) when testing TRG–SRC_1 on TRG–SRC_2 (as compared to testing TRG–SRC_1 on TRG–SRC_2). The fact that classification performance stays above 64% confirms the hypothesis that translationese features are source-language-independent.

Another trend is that in cross testing TRG–SRC_1 and TRG–SRC_2, the model where the source language is more distant from the target suffers a larger performance drop when tested on the test set with the closer-related source language, than the other way around. For instance, the DE–ES model tested on the DE–EN data suffers a decrease of 17.8 points, and DE–EN model tested on the DE–ES data suffers a decrease of 9.8 points. This may be due to DE–EN having learned more of the general translationese features, which helps the model to obtain higher accuracy on the data with a different source, while the DE–ES model may have learned to rely more on the language-pair-specific features, and therefore it gives lower accuracy on the data with the different source. A similar observation has been made by Koppel and Ordan (2011).

For the multi-source monolingual models (TRG–ALL), testing on TRG–SRC_1 and TRG–SRC_2 datasets shows a slight increase in performance for a source language that is more distant from the target, and a slight decrease for the more closely-related source language (as compared to testing TRG–ALL on TRG–ALL).

Table 7 displays the results of testing the multilingual (ALL–ALL[3]) models on all test sets for the neural architectures, as well as Handcr.+SVM. We observe that the

largest performance drop (as compared to testing on ALL–ALL[3] test set) happens for the EN–DE test set. For the DE–ES set, the performance actually increases for the neural models, but not for the `Handcr.+SVM`. We extended this experiment in Table 8, testing the ALL–ALL[8] on all test sets to further complement our multilingual analysis with more diverse languages and observe a similar trend, which is in line with the accuracy of the ALL–ALL[3] models on all test sets.

We also compare the performance of ALL–ALL[3] on different test sets to the original performance of the models trained on these datasets (in parentheses). There is a relatively larger drop in accuracy for the TRG–SRC data, than for TRG–ALL data. The largest drop for neural models is 6.7 accuracy points whilst the smallest performance drop for the `Handcr.+SVM` is 12.7. This highlights the ability of the neural models to learn features in a multilingual setting which generalize well to their component languages whereas the `Handcr.+SVM` method does not seem to work well for such a case. However, for ALL–ALL[8] models, Table 8 shows a large performance drop across all architectures as compared to the results from the models specifically trained for the task. The actual models are trained on language-specific features, whereas the ALL–ALL[8] model is trained on more diverse data containing typologically distant languages and thus captures less targeted translationese signals.

In summary, we observe that:

- For a given target language, even though a neural model trained on one source language can decently identify translationese from another source language, the decrease in performance is substantial.
- Neural models trained on multiple sources for a given target language perform reasonably well on single-source languages.
- Neural models trained on multilingual data ALL–ALL[3] perform reasonably well on monolingual data, especially for multi-source monolingual data.
- Using more source and target languages (ALL–ALL[8]) leads to a larger decrease in cross-testing accuracy.

Overall, the neural models perform well in such scenarios, with ***fastText*** performing not far behind the BERT- and LSTM-based deep learning approaches. These results demonstrate the effectiveness of feature and representation-learning approaches for translationese classification, where models are able to learn the features and patterns of translated text without the need for manual feature engineering, answering research question **RQ1**. Furthermore, the improved performance of transfer learning-based models like BERT and LSTM suggests that pre-training on large amounts of text data can significantly enhance the ability of the model to learn relevant features for translationese classification.

Corpus	Handcr. +SVM	fastText (FT)	LSTM	BERT
DE–EN	58.5±0.0 (↓13.0)	85.9±0.0 (↓2.5)	86.6±0.7 (↓2.9)	90.5±0.3 (↓1.9)
DE–ES	57.0±0.0 (↓19.2)	88.3±0.0 (↓2.6)	85.3±0.3 (↓6.9)	91.5±0.2 (↓2.9)
EN–DE	50.0±0.0 (↓17.6)	81.5±0.1 (↓3.6)	80.9±0.3 (↓5.8)	87.2±0.4 (↓3.5)
EN–ES	50.5±0.0 (↓19.6)	84.6±0.0 (↓3.0)	83.8±0.6 (↓5.3)	88.9±0.3 (↓3.0)
ES–DE	50.0±0.0 (↓21.0)	85.6±0.0 (↓2.8)	85.7±0.5 (↓4.6)	90.4±0.4 (↓1.9)
ES–EN	51.3±0.0 (↓15.4)	84.1±0.0 (↓2.9)	82.1±0.4 (↓6.7)	89.0±0.4 (↓2.4)
DE–ALL	59.9±0.0 (↓12.7)	87.2±0.0 (↓0.2)	85.9±0.5 (↓2.6)	90.8±0.1 (↓0.1)
EN–ALL	50.2±0.0 (↓15.1)	82.9±0.0 (↑0.2)	82.2±0.2 (↓2.1)	88.1±0.5 (↑0.2)
ES–ALL	50.0±0.0 (↓17.4)	84.8±0.0 (↓0.1)	85.2±0.5 (↓1.8)	89.8±0.3 (↓0.1)
ALL–ALL[3]	58.9±0.0 (0.0)	85.0±0.0 (0.0)	84.5±0.2 (0.0)	89.6±0.2 (0.0)

Table 7: Translationese classification accuracy of the ALL–ALL[3] model on all the other test sets (average and standard deviations over 5 runs). The difference from the actual trained model performance is indicated in parentheses

Corpus	Handcr. +SVM	fastText (FT)	LSTM	BERT
DE–EN	53.0±0.5 (↓18,5)	71.0±0.3 (↓17.4)	79.9±0.5 (↓9.6)	85.5±0.4 (↓6.9)
DE–ES	51.3±0.3 (↓24.9)	73.2±0.3 (↓17.7)	79.0±0.5 (↓12.9)	87.9±0.3 (↓6.5)
EN–DE	48.3±0.1 (↓19.3)	65.8±0.2 (↓19.3)	72.9±0.4 (↓13.8)	79.0±0.5 (↓11.7)
EN–ES	50.3±0.1 (↓19.8)	68.9±0.3 (↓18.7)	75.6±0.8 (↓13.5)	83.2±0.4 (↓8.7)
ES–DE	50.0±0.0 (↓21.0)	71.1±0.2 (↓17.3)	76.0±0.7 (↓14.2)	83.8±0.3 (↓8.5)
ES–EN	53.2±0.5 (↓13.5)	69.9±0.2 (↓17.1)	75.4±0.7 (↓13.4)	82.8±0.2 (↓8.6)
DE–ALL	53.1±0.5 (↓19.5)	72.1±0.3 (↓15.3)	79.7±0.6 (↓8.8)	86.8±0.2 (↓4.1)
EN–ALL	48.4±0.2 (↓16.9)	67.0±0.2 (↓15.7)	74.4±0.5 (↓9.9)	81.1±0.1 (↓6.8)
ES–ALL	50.8±0.3 (↓16.6)	70.4±0.2 (↓14.5)	75.9±0.6 (↓11.1)	83.2±0.3 (↓6.7)
ALL–ALL[3]	53.2±0.3 (↓5.7)	70.5±0.2 (↓14.5)	76.7±0.5 (↓7.7)	83.7±0.1 (↓5.9)
ALL–ALL[8]	65.4±0.1 (0.0)	70.4±0.1 (0.0)	77.2±0.3 (0.0)	84.6±0.2 (0.0)

Table 8: Translationese classification accuracy of the ALL–ALL[8] model on all test sets (average and standard deviations over 5 runs). The difference from actual trained model performance is indicated in parentheses.

3.7 Conclusion

In this chapter, we explore the ability of feature and representation learning-based methods to discern between translated and non-translated texts, addressing **RQ1**. Traditional approaches to translationese classification involve designing and selecting linguistic features manually, which can be a time-consuming and potentially non-exhaustive process. In contrast, feature and representation learning-based methods are highly performant and eliminate the need for manual feature engineering, making them more efficient and effective.

To provide context for our classification-based analysis, we also conduct a series of stylistic analyses on our monolingual single-source and multi-source data. Our analysis revealed that there is significant variation in the languages used within

each label, original and translated. In fact, we found that stylistic results could differ substantially for specific language pairs, and even for the same target language with different source languages. This highlights the need for a language-agnostic approach to translationese classification.

We propose a suite of word-embedding-based and representation learning approaches that rely on static or contextual embeddings and can be used with or without pre-trained model weights. We experiment with these approaches across a range of source-target language combinations in mono- and multilingual settings. Our results indicate that even simple representation learning-based approaches, such as ***fastText***, can be a viable and less expensive alternative to hand-crafted, linguistically inspired feature-selection methods for the translationese classification task. Interestingly, while the results from ***fastText*** were pretty comparative, deep neural models achieved the highest accuracy in the multilingual translationese classification task. We show that feature and representation learning approaches : (i) comprehensively outperform classical hand-crafted feature engineering based methods and (ii) exhibit better generalisation across languages.

Overall, these findings suggest that it is possible to distinguish between translated and non-translated texts using feature and representation learning-based methods. These methods are effective in identifying translationese without the need for prior linguistic knowledge or intuitions. We believe that the community can benefit from further analysis of what these systems actually learn in such tasks, and we encourage further exploration of representation learning-based approaches for translationese classification. However, it is worth noting that using traditional hand-crafted and linguistically inspired features, along with feature engineering in classification-based approaches to translationese, provides the advantage of using established feature ranking to reason about the linguistic properties identified as important by the classifier (Rubino, Lapshinova-Koltunski, and Genabith, 2016; Volansky, Ordan, and Wintner, 2015). This is more challenging with representation learning-based neural networks, where explainable artificial intelligence (XAI) approaches, such as those referenced in Amponsah-Kaakyire et al. (2022) using Integrated Gradients (IG), must be employed.

4

Debiasing of Translation Artifacts

Translation artifacts have been shown to bias the performance of NLP tasks involving translated data. In an effort to address this issue, this chapter examines the role of debiasing these artifacts in NLP tasks. We adapt the Iterative Null Space Projection (INLP) algorithm, originally designed to mitigate gender attributes, to translationese-induced bias in both word and sentence embedding spaces. INLP requires the direction of the bias in the representation space for it to work. This is challenging for word-based debiasing as unlike in e.g. gender bias, translationse cannot in general be captured in terms of contrastive word pairs. We show how to overcome this challenge. Additionally, we develop two techniques for debiasing translationese at the word level representations. We confirm the effectiveness of our debiasing approach by comparing the classification performance before and after debiasing on the translationese classification task. Additionally, we demonstrate the practical utility of our debiasing method by applying it to a natural language inference (NLI) task involving translated data, where we observed improved accuracy as a result of reduced translation-induced bias.

4.1 Introduction

Translationese artifacts have been found to exert a substantial influence on diverse downstream tasks. In Machine Translation (MT), Toral et al. (2018) found that translating source sentences that are already the result of a translation are easier to translate than original sentences. Similarly, Edunov et al. (2020) show that back-translation results in larger BLEU[1] scores when translationese is on the source side and original text is used as reference. To avoid such artifacts, recent works (Graham, Haddow, and Koehn, 2020; Zhang and Toral, 2019) advise using original source sentences for machine translation (MT) evaluation. Freitag, Caswell, and Roy (2019) draw attention to *translationese* as a factor contributing to the mismatch between BLEU scores and human evaluation, expressing concerns that metrics based on overlap might favor hypotheses containing translationese language over those using more originally authored

1 BLEU, or the Bilingual Evaluation Understudy, is a metric based on n-gram match precision for comparing a candidate translation to one or more reference translations (Papineni et al., 2002).

language. In contrast to the influence of Translationese on evaluation, the effect of Translationese in the training data has been investigated in several studies (Bogoychev and Sennrich, 2019; Kurokawa, Goutte, and Isabelle, 2009; Lembersky, Ordan, and Wintner, 2012). Riley et al. (2020a) train a sentence-level classifier to differentiate translationese from original target text, and use this classifier to tag the training data for an NMT model to produce output that shows fewer translationese effects.

In other cross-lingual applications, Singh et al. (2019) show that substituting segments of original training samples by their translations from another language improves performance on natural language inference (NLI) tasks. Clark et al. (2020) introduce a translation-free question-answering dataset to avoid having inflated gains from translation artifacts in transfer-learning tasks. Artetxe, Labaka, and Agirre (2020) show that cross-lingual models suffer from induced translation artifacts when evaluated on translated test sets. These examples motivate the need to reduce translation artifacts, which leads to our second research question:

RQ2: Is it possible to effectively attenuate translation artifacts from latent representation spaces?

To mitigate the impact of translationese from representation of (word and sentence) embeddings, we propose a new approach: translationese debiasing. However, removing translationese signals is not as straightforward as removing human-like biases such as gender and profanity, which can be captured in terms of simple lists of contrastive word pairs. Translationese signals are complex and multi-faceted, manifesting as a mix of morphological, lexical, syntactic, and semantic phenomena and they cannot in general be captured in terms of contrastive word pairs as (or to the extent it can be done) in gender or profanity debiassing. While there have been recent proposals to remove or attenuate human-like biases in both static and contextualised word embeddings, there has been no study on attenuating and eliminating implicit signals such as translationese in embeddings.
To achieve this, we adapted the Iterative Null Space Projection (INLP) algorithm (Ravfogel et al., 2020), which was originally designed to address gender attributes, to mitigate translationese-induced bias in both word and sentence embedding spaces. Our approach represents the first attempt to debias translationese signals from latent representations, which has not been previously explored, and our findings demonstrate its effectiveness in reducing this type of bias.

Specifically in addition to word-level, we also utilise the INLP method with several neural sentence-level classification architectures, including ***fastText***, mBERT, and XLM, to debias translationese at sentence levels. For INLP to work it requires a vector indicating the direction of the bias to be mitigated. For sentence-level representations, we can compute this as our data are labelled original or translationese. However,

computing the direction of the bias is challenging for word representations as unlike e.g. for gender bias, translationese cannot be captured well in terms of simple contrastive word pairs. Below we show how we can solve this challenge. Furthermore, we propose two alternative methods for detecting and debiasing translationese bias in word embeddings. We confirm the effectiveness of our debiasing approach in terms of an intrinsic evaluation by comparing the classification performance before and after debiasing on the translationese classification task. Our evaluation results show that after debiasing, the performance of the models in classifying translationese decreases to that of a random classifier.

To assess the practical usefulness of our proposed debiasing approach in an extrinsic evaluation, we integrate it into the natural language inference (NLI) task (Bowman et al., 2015), demonstrating that it enhances NLI accuracy by reducing translation-induced bias.

The remainder of this chapter is structured as follows. In Section 4.2, we outline the related work. We describe our debiasing strategies in Section 4.3. In Section 4.4, we describe our experimental settings with data, task and model specifications. We analyse the findings in Section 4.5 and finally, conclude the chapter in Section 4.6.

4.2 Related Work

Biases in language representations can have wide-ranging adverse affects in machine learning applications. Models built from stereotyped associations learned from data propagate these biases into the decisions they produce. As a result, eliminating or at least mitigating biases is essential before using representations. Much previous research (Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020; Zhao et al., 2018a) provides evidence that (i) word embeddings do capture societal biases, (ii) models perpetuate these biases to influence various tasks such as abusive language detection (Park, Shin, and Fung, 2018; Zhao et al., 2018a), co-reference resolution (Rudinger et al., 2018), etc.

There exist a plethora of efforts to debias societal bias in word embeddings. The first work in debiasing gender influence in embeddings was by Bolukbasi et al. (2016) who show that word embeddings trained on news articles exhibit societal gender stereotypes, for example, 'woman' is associated with *nurse* in the same way 'man' is associated with *doctor*. Through the use of principle component analysis (PCA), they identified a gender subspace from gendered word pairs (e.g., "man-woman", "he-she"). Through re-embedding all words such that they do not, or less significantly project onto the gender subspace, they attenuate gender bias while retaining the representation of all neutral words. However, Gonen and Goldberg (2019) makes the case against these debiasing strategies geared towards static word embeddings, claiming that these methods simply cover up the biases - and that they can resurface.

Also related is the work of Niu and Carpuat (2017) that analyses a stylistic subspace to capture the degree of formality in word embeddings.

Following in the footsteps of these authors, Caliskan, Bryson, and Narayanan (2017) illustrates how human-like racial biases exist in data, i.e., black people's names are found to be considerably more associated with unpleasant stereotypes than with pleasant terms when compared to names associated with white people. These biases exhibited in the data might be transmitted on the model built on for the task of sentiment or opinion mining. Similarly, Zhao et al. (2018b) develops a gender-neutral version of (GloVe) called GN-GloVe that tries to preserve gender information in some of the learned dimensions of the word embedding while guaranteeing that other dimensions are free from this gender effect. Similarly, there has been an effort to investigate the amount of ethnic and religious bias within word representations in multi-class settings (Garg et al., 2018; Manzini et al., 2019), and personality stereotypes (Agarwal et al., 2019). In contrast to the approaches discussed above, Kaneko and Bollegala (2020) presents GP-GloVe, a post-processing method for preserving gender-related information with an autoencoder, while eliminating discriminatory biases from stereotyped instances. Dev and Phillips (2019) find that orthogonal projection to gender direction can be used to debias word representations. More recently, Ravfogel et al. (2020) show that iteratively projecting word representations to the null space of the gender direction can improve debiasing performance. In Section 4.3.1, a detailed discussion on this approach is presented.

Despite the extensive literature on debiasing unfair and discriminative biases at the word-level, very little work has focused on sentence representations. For instance, Liang et al. (2020) detect gender and religion bias in sentence representations. Liu et al. (2019a) present a test dataset for dialogue created counterfactually by combining templates and manually created lists of word pairs; this work demonstrates that models generate dialogue that is less diverse when given sentences that contain words that describe individuals of underrepresented groups. May et al. (2019) extend the research in detecting bias in word embedding techniques to that of sentence embeddings. They try to generalize bias-measuring techniques, such as using the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan, 2017) in the context of sentence encoders by introducing their new sentence encoding bias-measuring techniques, the Sentence Encoder Association Test (SEAT). Liang et al. (2020) adapt the Hard-Debias technique for debiasing word embeddings (Bolukbasi et al., 2016) to debias sentence representations. More recently, Schick, Udupa, and Schütze (2021) propose a post-hoc debiasing technique that utilizes a model's internal knowledge to discourage the generation of biased text.

4.3 Translationese Debiasing Strategies

Up till now, researchers have detected, captured and mitigated specific biases (e.g., gender, profanity, etc.) in word embeddings, using lists of contrastive word pairs (e.g., woman-man, she-he). However, attenuating and eliminating a more implicit signal like translationese in embeddings has yet to be studied. Translationese signals are complex and multi-faceted and, unlike e.g. gender and profanity, can in general not be captured in terms of simple lists of contrastive word pairs, but rather manifest as a complex mix of morphological, lexical, syntactic and semantic phenomena.

In order to address this issue, we investigate whether it is possible to debias translationese artifacts and evaluate the results through the task of translationese classification. Given that there are no predefined word lists for translationese, we instead rely on labeled data at the sentence level, which includes both translated (i.e. translationese) and original sentences to identify translationese signals using the INLP algorithm in Section 4.4.2.1. This is followed by exploring debiasing at the word level without using predefined word lists. Specifically, we develop two techniques for debiasing translationese at the word level representations detailed in Section 4.4.2.2. To the best of our knowledge, we are the only ones to deal with translationese at the level of word embedding spaces leveraging distances between graph-based representations of original and translated data.

4.3.1 Iterative Nullspace Projection

INLP was proposed by Ravfogel et al. (2020) to remove linearly decipherable features from vector representations by iteratively projecting onto null-space. Their approach was originally developed for gender bias mitigation. Given a set of labeled data with data points $X = x_1, ..., x_n$ and task labels $Y = y_1, ..., y_n$, INLP uses a standard classification setup with a neural network and a simple classifier τ on top to mitigate the *protected attribute* T. Specifically, encoded representations $h(x_k)$ up to the last hidden layer are taken from the data (i.e., representations that extract effective protected attributes for predicting O/T), and an intervention is performed on these to obtain the debiased last hidden representation.

INLP neutralises the ability of the classifier τ to linearly predict T from h. Concretely, it is assumed that τ is parameterised by a matrix W and trained to predict T from h. Using W, one can collapse the data onto its nullspace $N(W)$[2] with a projection matrix $P_{N(W)}$. This guarantees $WP_{N(W)}h(T) = 0$, i.e., the information used to classify T is linearly removed. By repeating this process i times until no classifier achieves

2 The nullspace represents the subspace of the input feature that is not used by the classifier to predict the protected attribute. In this case, the protected attribute is translationese.

above-majority accuracy, INLP neutralises all features that W_i uses for predicting T from $\tilde{h}$:

$$\tilde{h} := P_{N(W_1)} P_{N(W_2)} ... P_{N(W_i)} h \tag{6}$$

Debiased embedding $d_{INLP}(w)$ is represented as:

$$d_{INLP}(w) := Pw \tag{7}$$

where P is the guarding projection matrix, obtained by iteratively obtaining nullspace projections $PN(W_i)$ and computing $P = P_N(W_i) P_N(W_{i-1}) ... P_N(W_0)$, and $h(x) = Px$ is the guarding function.

The steps of the algorithm are as follows:

Algorithm 1 Iterative Nullspace Projection (INLP)

Require: Training set X of vectors and their corresponding protected attributes Z, with the specified number of iterations denoted as n
Output: The resulting projection matrix P
function GetProjectionMatrix(X, Z, n)
  Initialize $X_{projected} \leftarrow X$ and $P \leftarrow I$
  for $i = 1$ to n **do**
    $W_i \leftarrow$ TrainClassifier($X_{projected}$, Z)
    $B_i \leftarrow$ GetNullSpaceBasis(W_i)
    $P_{N(W_i)} \leftarrow B_i B_i^\top$
    $P \leftarrow P_{N(W_i)} P$
    $X_{projected} \leftarrow P_{N(W_i)} X_{projected}$
  end for
  return P
end function

We are interested in debiasing translation artifacts T from latent representation spaces. To achieve this, we adapt the Iterative Null Space Projection (INLP) algorithm to mitigate translationese-induced bias in both word and sentence embedding spaces. To address bias in sentence representations, we use predefined labels (original vs. translated), where translated labels correspond to protected attributes T. However, unlike sentence-level bias, identifying simple contrastive word pairs to capture translationese is not always possible for word-level bias. Therefore, we develop two techniques to remove translationese bias in word representations. The following sections detail these approaches.

4.3.2 Translationese Directions

To mitigate translationese-induced bias in both word and sentence embedding spaces using the Iterative Null Space Projection (INLP) algorithm, a unique direction in the latent space that corresponds to translationese, our protected attribute, is required. For the direction in sentence representations, we use pre-defined labels (originals vs. translated), where the translated labels correspond to the protected attribute T. However, this poses a challenge for getting translationese direction at the word-level, as unlike gender and profanity, translationese cannot generally be captured using simple contrastive word pairs, such as $\overrightarrow{he} - \overrightarrow{she}$, for gender mitigation. Therefore, to quantify the contrastive subspaces of translationese (i.e., original and translated), we propose two techniques:

(i) **Stepwise Aligned Space Projection**, where we construct the subspaces based on the concept of word-usage between original and translated data. We extract lists of identical word pairs and analyse how their use differs in translated and original data. Instead of contrasting word pairs, we identify differences in original and translated word embedding spaces for the same word, indicating the presence of translationese in the embeddings.

(ii) Alternatively, we suggest a simpler approach, **Direct Joint Space Projection**, that utilises a joint embedding space where words are simply tagged according to their origin (translated or original), without the need for a word list.

In the following sections, we describe these two approaches in detail.

Stepwise Aligned Space Projection. In order to estimate the unique translationese direction, we derive a list of words (G) used differently in translationese T and original O data using the *usage change* concept from Gonen et al. (2020). The same word used in different data sets (original and translated) is likely to have different neighbours in the two embedding spaces. We only use words from the intersection of both vocabularies O and T. We compute the score for context change across the embeddings $\mathcal{O}$ and $\mathcal{T}$ of the two data sets by considering the size of the intersection of two sets where each word in a corpus is represented as its top-k nearest neighbours (NN) in its embedding space:

$$\text{score}^{k}(w) = -|\text{NN}^{k}_{\mathcal{O}}(w) \cap \text{NN}^{k}_{\mathcal{T}}(w)| \tag{8}$$

where $\text{NN}^{k}_{i}(w)$ is the set of k-NN of word w in embedding space i.

The smaller the size of the intersection, the more different the word is used in the two data sets (and words with the smallest intersection can be seen as indicators of translationese). Given $\mathcal{O}$ and $\mathcal{T}$, we collect a ranked list of about 500 words with the smallest intersection as our translationese word list G. G allows us to identify the seed translationese direction for INLP. This translationese direction is later used to obtain the debiased space.

Next, we compute a joint word embedding space $\mathcal{J}$ from the concatenation of the translated and original data, T and O. Since this joint space $\mathcal{J}$ includes both original and translationese signals, we then align the previously unrelated $\mathcal{O}$ and $\mathcal{T}$ spaces to this embedding space $\mathcal{J}$, using VecMap[3] (Artetxe, Labaka, and Agirre, 2018), producing aligned spaces $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{T}}$, and resulting in an extended single embedding space where $\mathcal{T}$ and $\mathcal{O}$ are aligned to $\mathcal{J}$. Next, we compute the translationese direction v of the same word w in the two embeddings spaces, $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{T}}$, using,

$$\begin{aligned} \mathbf{v}(w) &= \tilde{\mathcal{T}}[w] - \tilde{\mathcal{O}}[w], \forall w \in G \\ \overline{\mathbf{v}} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i \end{aligned} \tag{9}$$

where $\tilde{\mathcal{T}}[w]$ is the embedding of the word w in the translated data, $\tilde{\mathcal{O}}[w]$ is the embedding of the word w in the original data, N is the number of words in the wordlist, and $\mathbf{v}_i$ is the translationese direction vector for the ith word in the wordlist. Finally, we compute the similarity of words in $\mathcal{J}$ along the directions v and $-v$, to divide them into two subspaces, translationese and non-translationese, respectively.

Direct Joint Space Projection. This approach involves directly building the embeddings of a specific word w from the original ($\mathcal{O}$) and translated ($\mathcal{T}$) datasets into a single joint space $\mathcal{J}'$. We achieve this by annotating w as either w_o or w_t on surface word forms to distinguish between the two embeddings of the same word coming from the original and translated datasets. By doing so, we can easily track and differentiate between the two embeddings of w in the same embedding space $\mathcal{J}'$, which is a simple concatenation of the two datasets. This approach eliminates the need to compute the translationese direction vector v to group the subspaces, as well as the complexity of maintaining and aligning $\mathcal{O}$, $\mathcal{T}$, $\mathcal{J}$, $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{T}}$ spaces.

4.4 Experimental Setup

4.4.1 Datasets

Data. We use the Europarl corpus annotated with translationese information from Amponsah-Kaakyire et al. (2021) previously detailed in Section 3.3.1. We focus on three languages: English (En), German (De) and Spanish (Es). The corpus provides originals in the three languages (L1) and translations into these three languages that

3 VecMap is a mapping algorithm proposed by (Artetxe, Labaka, and Agirre, 2018) that learns a linear transformation from the source language to the target language using a bilingual dictionary, and it aligns new embeddings from the source and target languages in a common vector space. The mapping algorithm involves a multi-step series of transformations of the source and target spaces that generalises many previous approaches. Their code is open-sourced at-https://github.com/artetxem/vecmap

	fastText Avg.	**mBERT** CLS	**mBERT** pool	**XLM** CLS	**de-biased**
En-De	0.64	0.73	0.79	0.71	0.50
En-Es	0.71	0.78	0.83	0.77	0.50
De-En	0.68	0.78	0.84	0.77	0.50
De-Es	0.69	0.79	0.86	0.79	0.50
Es-De	0.71	0.77	0.85	0.77	0.50
Es-En	0.72	0.76	0.82	0.77	0.50

Table 9: Sentence Embedding Classification accuracy on original versus translationese using different models. After INLP debiasing, translationese classification reduces to random 50% accuracy in all cases.

	Direct Joint	**Stepwise Aligned**		**de-biased**
		INLP.Single	INLP.Avg	
En-De	0.98	0.99	1.00	0.50
En-Es	0.92	1.00	1.00	0.50
De-En	0.96	1.00	0.99	0.50
De-Es	0.91	1.00	1.00	0.50
Es-De	0.93	0.99	0.99	0.50
Es-En	0.95	1.00	1.00	0.50

Table 10: Classification accuracy on original versus translationese with word embeddings using our two approaches, before and after debiasing with INLP.

come from original texts in the other two (L2). We use the notation L1–L2 to refer to the different sets in Table 9 and Table 10. For example, in En-De, L2 refers to English text translated from German. For each corpus, there is an equal number of translated and original sentences: 42k for De–En, De–Es, En–De, En–Es, Es–De and Es–En. We use 70% of the sentences for training, 15% for development and 15% for testing, and note that we do not provide a stylistic analysis as previously discussed in Section 3.3.

4.4.2 Model Specifications

We apply the debiasing strategies described in Section 4.3 using the following model specifications.

4.4.2.1 Translationese in Sentence Representations

We use INLP to study the impact of removing the translationese attributes T from semantic representations, via a binary classification task. The binary classifier learns to distinguish between original and translationese sentences. In our setup, labels Y

correspond to original and translationese and act as protected attributes. We consider three input representations as follows:

(i) ***fastText*** (Joulin et al., 2016): we compute an average of all token vectors in the sentence.

(ii) **mBERT$_{\texttt{CLS}}$** (Devlin et al., 2019): we use the `[CLS]` token in mBERT as sentence representation.

(iii) **mBERT$_{pool}$**(Devlin et al., 2019): we use mean pooling of mBERT's contextualised word embeddings.

(iv) **XLM$_{\texttt{CLS}}$**(Conneau et al., 2019): we use the `[CLS]` token from XLM-RoBERTa.

In our experiments, we use a logistic classifier on top of sentence embeddings h obtained with 4 models defined above, without any additional fine-tuning to the translationese classification task.

4.4.2.2 Translationese in Word Representations

In our word-level experiments, we use the translated and original parts of the data described in Section 4.4.1 to estimate the word embeddings $\mathcal{O}$ and $\mathcal{T}$ and use k=1000 nearest neighbours in Equation 8. We observe a significant value of k results in large neighbour sets for each word in the two corpora, resulting in a more stable translationese wordlist G. For our experiments in Section 4.3.2 we only consider words attested at least 200 times in the data. The extraction of translationese wordlists is shown in Figure 8. Section 4.5.4 shows the top 50 elements for all the word lists.

To initialise the INLP algorithm for the experiments in Section 4.3.2, we use Equation 9 in two ways: (i) **INLP.single**: with a direction vector created from the difference between two aligned spaces for the single highest ranked word in G, and (ii) **INLP.avg**: by averaging the differences of all words in G.

4.5 Evaluation and Analysis

4.5.1 Translationese Classification Performance

First, we evaluate the effectiveness of mitigating translationese bias from latent representations for the internal task of classifying translationese.

The first four columns in Table 9 summarise the translationese classification at sentence level accuracy achieved by the four models. mBERT$_{pool}$ achieves the best performance for all languages, while ***fastText*** trails the pack. The final column in Table 9 shows that INLP is close to perfection in removing translationese signals for the

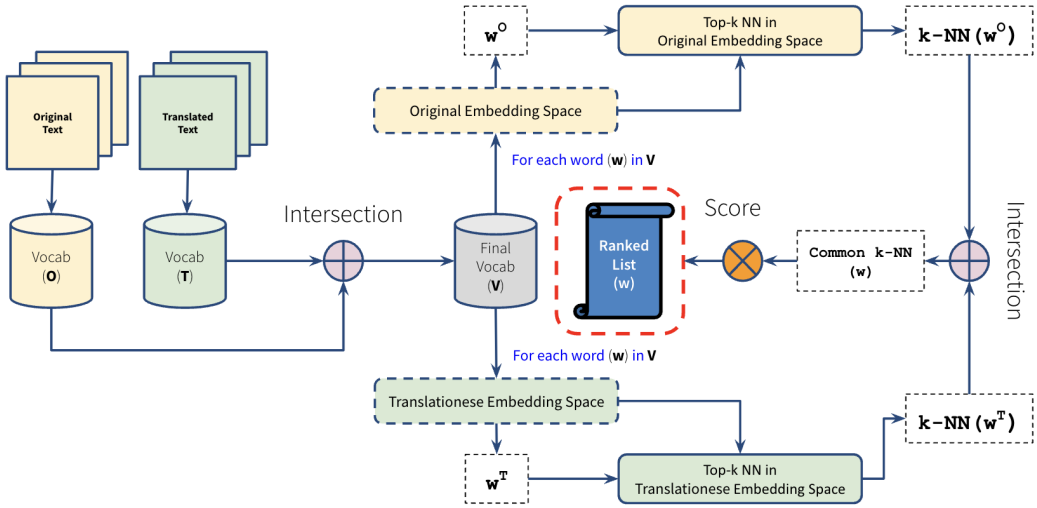

Figure 8: Extraction of the translationese wordlists based on the intersection of their top-k nearest neighbours in Original and Translationese spaces.

linear classifiers, reducing accuracy to a random 50%. This indicates the effectiveness of INLP in debiasing translationese from sentence representations.

We compare the performance before and after debiasing using our two word-level debiasing methods in Table 10. As expected, debiasing reduces classification accuracy for all language pairs from ~100% to ~50% for both methods, indicating the effectiveness of INLP in debiasing translationese from word representations.

4.5.2 Semantic evaluation

To ensure that the quality of the word representations is not compromised by the debiasing process, we conducted an evaluation on the performance of the original and debiased embeddings using the MultiSIMLEX benchmark (Vulić et al., 2020) for word analogy tasks. However, as MultiSIMLEX does not include German, we utilised the German-Simlex dataset from Leviant and Reichart (2015) for semantic evaluation for German.

We found that after debiasing, the Spearman's ρ correlation coefficients showed negligible decreases of only 0.02 on En-De and En-Es, 0.01 on Es-En and Es-De, and a slight decrease of 0.3 on De-En. However, there was an increase of 0.01 on De-Es.

De-En	De-Es	En-De	En-Es	Es-De	Es-En
466	510	429	483	547	504

Table 11: Size of translationese word lists created with the usage change algorithm (Gonen et al., 2020).

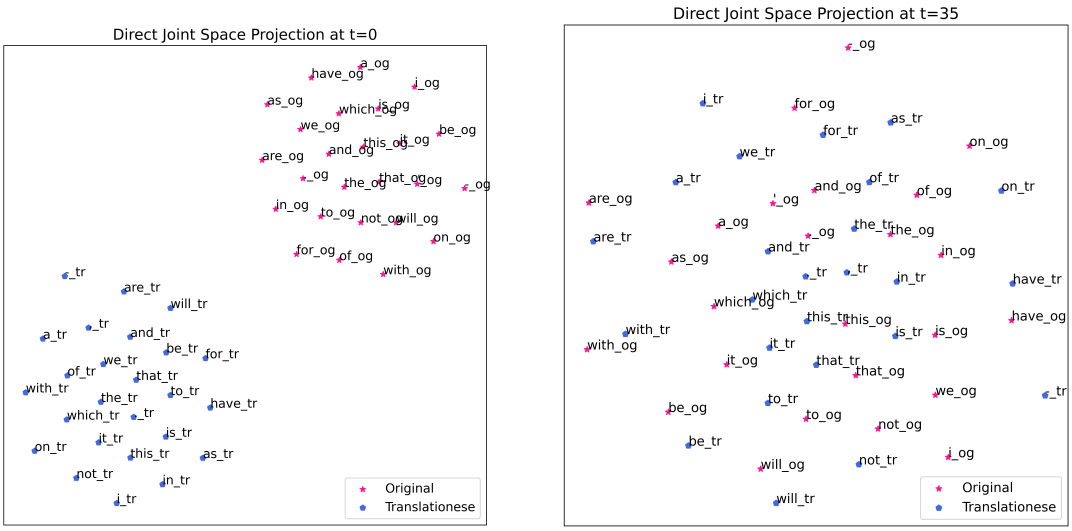

Figure 9: Clustering on Direct Joint Space, before (up-side) and after (down-side) debiasing.

These results indicate that the debiasing process had a minimal impact on the quality of the debiased word representations.

4.5.3 Clustering

The visualization in Figure 9 depicts the t-SNE (Maaten and Hinton, 2008) projection of the tagged tokens in Direct Joint Space Projection before and after debiasing with INLP. It is evident from the results that the original and translationese classes are no longer linearly separable after the debiasing process, which was followed by the projection of the representations on their null-spaces. Similar to Ravfogel et al. (2020), we also quantify this change using V-measure, which evaluates cluster purity. To assess the difference, we perform K-means clustering with K = 2 clusters on the vectors. Specifically, we calculate the V-measure (Rosenberg and Hirschberg, 2007), which measures the degree of overlap between the two clusters and the groups. Figure 9 illustrates that, after intervention, the vectors become increasingly mixed and are no longer clustered by translationese.

4.5.4 Translationese Word Lists

Table 11 displays the size of the translationese word lists produced by the usage change algorithm of Gonen et al. (2020), which varies depending on the language.

N	Stability Score		
50	100	500	1000
0.7	0.71	0.88	0.93

Table 12: Stability of Joint Space Wordlist

4.5.5 Stability of Stepwise Aligned Space Wordlist

We determine the stability of the word lists with the intersection of top-N words in the ranked lists, l_1 and l_2, retrieved from three runs of the usage change algorithm (Gonen et al., 2020), taking the two corpora O and T as input. We formulate this as:

$$\text{N-sized overlap}(l_1, l_2, N) = \frac{|l_2^N \cap l_2^N|}{N} \tag{10}$$

where l_i^N is the set of top N ranked words in ranking l_i. As the value of N increases, the stability score also increases. This suggests that larger joint space wordlists are more stable and consistent across different sources than smaller ones. Specifically, the stability score increases from 0.7 for N=50 to 0.93 for N=1000, which is a substantial improvement.

4.5.6 Extrinsic Evaluation

We demonstrate the practical utility of our translationese debiasing method on the practical task of Natural Language Inference (NLI). To achieve this, we investigate the impact of removing translation-induced bias from the translated data and analyse its effect on NLI performance. In NLI, the goal is to determine the relationship between two sentences - a premise and a hypothesis - and classify it into one of three categories: entailment, contradiction, or neutral.

Recent work by Artetxe, Labaka, and Agirre (2020) has shown that existing NLI datasets contain a significant amount of lexical overlap between the premise and the hypothesis. This overlap is often utilised by neural NLI models to achieve high accuracy in their predictions. However, when the premise and hypothesis are independently paraphrased using translation and back-translation, the lexical overlap is reduced, leading to decreased model performance. Even though reducing lexical overlap may lead to a reduction in model performance, it may well improve model generalisation. If the model is trained on data where the premise and hypothesis are translated together, it may learn to rely on superficial patterns, e.g., a high degree of lexical overlap to make predictions that the premise and hypothesis entail each other. This is also demonstrated in the work of Artetxe, Labaka, and Agirre (2020) on other

	Approach	SNLI Model		
		Original	**Back-translated**	**Debiased**
Sym	Word-Joint	67.2±0.1	64.1±0.2	64.7±0.1
	Word-Aligned	67.8±0.1	64.6±0.2	64. 9±0.2
Asym	Word-Joint	67.2±0.1	64.4±0.2	65.1±0.2
	Word-Aligned	67.8±0.1	65.2±0.2	65.7±0.2

Table 13: Test set accuracies (5 runs average) for models trained and tested on original SNLI, back-translated with German as pivot and back-translated debiased at word-level

tasks such as XQuAD (Artetxe, Ruder, and Yogatama, 2019) and MLQA (Lewis et al., 2019).

To address this issue, we evaluate the impact of our INLP-based translationese debiasing approach on the performance of NLI models trained on translated data. We create a Back-Translated (BT) NLI dataset by translating the premise and hypothesis of the original SNLI data independently. We then train two NLI models: one on the Original data and the other on the BT data. To reduce the impact of translationese artifacts resulting from back-translation, we apply sentence and word embedding debiasing techniques, as described in Sections 4.4.2.1 and 4.4.2.2, to the embeddings generated from the BT data. We then use these debiased-BT embeddings to train a third debiased NLI model (**Debiased**), as shown in Table 13.

To evaluate the effectiveness of our translationese debiasing approach, we conduct experiments in two different scenarios: (i) Symmetric `[Sym]` and (ii) Asymmetric `[Asym]`. In the Symmetric scenario, we test the Original NLI model on an original test set, while the BT and debiased NLI models are tested on BT test data. In the Asymmetric scenario, all models are tested on original test data. We use the Symmetric scenario to determine if debiasing the model trained on the translated train-test data can bring its performance closer to that of the model trained on the original train-test data. In the Asymmetric scenario, we investigate the asymmetry between the original test data and the BT-NLI training data and assess whether our translationese debiasing of the BT training data can compensate for this asymmetry and improve NLI performance. Below, we describe our experimental settings.

4.5.6.1 Experimental Setup

Data. We use the large-scale Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), which contains 570,000 sentence pairs in the training set that have been manually labeled as entailment, contradiction, or neutral. The development and test sets consist of 10,000 examples each, and are used in the same manner as in the original SNLI dataset. The back-translated variant of the data is generated using German as the pivot language. This is accomplished by utilising the pre-trained

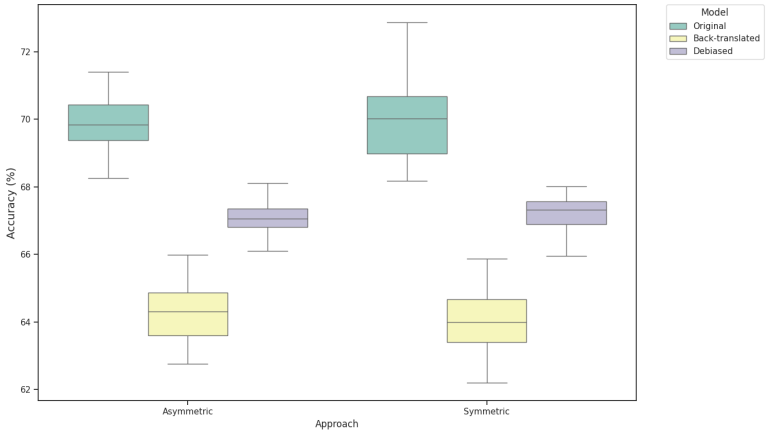

Figure 10: Sentence level NLI debiasing in Symmetric settings (*left*), Asymmetric settings (*right*)

models of Facebook-FAIR's WMT-19 news translation task submission (Ng et al., 2019) to translate the English sentences into German and then back into English. This results in a new set of English sentences that differ slightly from the original sentences.

Models. We train three distinct natural language inference (NLI) models, with each model intended for use with a different set of embeddings. The models were trained using the embeddings from the original data, embeddings from the back-translated data, and debiased embeddings from the backtranslated data, respectively.

For the word-level setup, we use a single hidden BiLSTM layer followed by a standard feedforward output layer on top of frozen ***fastText*** word embeddings for the 3-class NLI classification. The computation of word embeddings and debiasing follows the setup described in Section 4.4.2.2.

In the case of sentence-level debiasing, we use the $\text{BERT}_{\text{pool}}$ approach discussed in Section 4.4.2.1. This involved using a multiclass logistic classifier on top to predict the labels.

Hyperparameters. Each individual run took approximately 1.5 hours to complete using a GTX1080Ti GPU. For the SNLI sentence-level experiment presented in Table 13, each classification and debiasing step required approximately 2 hours on a V100-32 GB GPU. The fastText model was set with a minimum word count of 5 and an embedding dimension of 300. The Logistic Regression model was trained with the 'saga' solver, L2 regularization, a maximum of 7 iterations, and additional settings such as warm start enabled, verbosity set to 5, and a random state of 23. The BiLSTM model used a hidden dimension of 300, a dropout rate of 0.2, and a batch size of 32, optimized with Adam using a learning rate of 0.0001 for 15 epochs. For SNLI experiments, the SNLI-sentence model had 45 classifiers and a maximum of 1500 iterations, while the SNLI-Aligned (word) and SNLI-Joint (word) models had 34 and 35 classifiers, respectively. Each individual run took approximately 1.5 hours to complete on a GTX1080Ti GPU, whereas the SNLI sentence-level experiments required around 2 hours per classification and debiasing step on a V100-32GB GPU.

4.5.6.2 Results

Table 13 displays the classification accuracies at the word-level for (i) and (ii), while Figure 10 depicts the same for the sentence level. As observed, the outcomes are consistent with Artetxe, Labaka, and Agirre (2020) in that models trained on Original data perform better than those trained on BT data in both the `Sym` and `Asym` scenarios. Additionally, Table 13 indicates that debiasing of translationese leads to a modest improvement in classification accuracy over back-translated on SNLI-debiased for all models, with only a minor improvement at the word-level but a significant boost at the sentence level. This may be attributed to the fact that translationese comprises both lexical and syntactic phenomena, which are more accurately captured at the sentence level. The results presented in Table 13 suggest that debiasing translation artifacts aids in reducing the asymmetry between the translated train and the original test set. Consequently, rather than translating the entire test set to align with the training set in transfer-learning tasks, debiasing the training set for translation artifacts appears to be a promising approach for future research.

It is important to acknowledge that when attempting to debias BT data, it is impossible to recreate the same lexical overlap as in the original SNLI dataset. Consequently, the primary objective of this investigation was not to fully restore the accuracy achieved on the original data. Instead, its aim was to demonstrate that translation generates additional biases that can influence the final task, and that debiasing techniques are more effective at minimising these new machine-generated biases.

Finally, for a complex task such as translationese debiasing, linear intervention alone may not be sufficient. As a result, non-linear guarding approaches need to be investigated further.

4.6 Conclusion

This chapter addresses the issue of translationese artifacts, which have been shown to bias the performance of NLP tasks involving translated data. Therefore, to mitigate the impact of translationese, we propose a new approach: translationese debiasing. To this end, we extend the Iterative Null Space Projection (INLP) algorithm (Ravfogel et al., 2020), originally designed to mitigate gender attributes, to address translationese-induced bias in both word and sentence embedding spaces, providing an answer to **RQ2**.

We evaluate the effectiveness of our approach by comparing the classification performance before and after debiasing on the translationese classification task. To this end, we explore debiasing translationese artifacts in sentence-level embeddings with INLP. As expected, the INLP-based linear translationese debiasing results on static word embeddings are as "perfect" as our sentence-level results, reducing the

performance of a linear translationese classifier on the debiased data to chance. Additionally, we also develop two techniques for debiasing translationese at the word level. The Stepwise Aligned subspace approach is akin to the subspace construction method proposed by Bolukbasi et al. (2016) and Ravfogel et al. (2020) for gender debiasing. The Direct Joint subspace approach is a simplified method that operates directly on the joint space without the use of a separate translationese word list and multiple independently computed and subsequently aligned subspaces. We demonstrate that our debiasing strategies effectively attenuate translationese signals in both of these spaces.

In addition, we evaluate the effects of debiasing translation artifacts on a practical NLI task in two settings. Although we achieve perfect performance for the translationese classification-based debiasing task with INLP, this only translates into modest improvements resulting from debiasing translation artifacts resulting from back-translation with neural machine translation in an NLI task, with slightly better but statistically significant results for sentence-level debiasing. This indicates that our debiasing method is effective in reducing translation artifacts, but such bias is more complex than what can be perceived and mitigated by a linear classifier.

Finally, it should be noted that our study represents the first attempt to address translationese biases through latent representations in word and sentence embedding spaces. However, it is important to acknowledge that our research does not analyse the actual surface form of the debiased outputs. Therefore, we plan to incorporate human evaluation of the generated outputs as an essential part of our future research.

5

Unsupervised tracing of Translationese in Semantic Spaces

In this chapter, we present a new, graph structure-based approach to detect translationese in semantic spaces in an unsupervised manner without the need for any explicit linguistic labels. Previous studies use a combination of lexical and syntactic features to show that footprints of the source language remain visible in translations. In this chapter, we sidestep looking at manually defined (often linguistic) features that contribute strongly to translationese classification and introduce a methodology for tracing translationese in semantic spaces based on word usage and the notion of isomorphism. Specifically, we show that using our proposed graph-isomorphism methods, semantic difference between original target language data and translations into this target language can be detected and that such distances between both are far from being random. We find that isomorphism weakens consistently with increased linguistic distance among language families, reflecting the footprints of the source language on the translation. Following this, we show that our proposed methods are robust under a variety of training conditions, encompassing data size, type, and choice of word embedding models. Additionally, our findings indicate that these methods are language-independent, in the sense that they can be applied to multiple languages and are not limited to a specific language or language family.

5.1 Introduction

Parts of linguistics have long been concerned with studying cross-linguistic variation in order to classify languages genetically and typologically. Historical comparative linguistic methods are used to determine the genetic relationships between languages by analysing concept lists of words (Dyen, Kruskal, and Black, 1992; Swadesh, 1952) that share a common origin and have similar meanings and pronunciations in multiple languages. Linguistic typology studies the distinctness of languages and seeks to make generalisations about cross-linguistic variation at different levels of linguistic representations. Computational analysis methods reconstruct language distances based on measurable linguistic patterns.

Studies have also shown that some aspects of language differences are so profound that even when translated into another language, the structure of the source language is still preserved in translation. Rabinovich, Ordan, and Wintner (2017) cluster languages based on linguistically inspired features of their translations into the same target language and show that syntactic footprints of the source language exist in the translations which can be used to estimate phylogenetic similarities between languages. This phenomenon is known as source language interference (Toury, 2012) or *shining-through* (Teich, 2003) of the source language on the translated text. Computational approaches (Bjerva et al., 2019) have been developed to analyse raw words, part-of-speech tags, phrase-structure, or dependency-based input sequence representations of the data to predict language trees and investigate whether predicted trees reflect phylogenetic, geographic proximity, or syntactic (structural) differences. However, it remains unexplored whether interference of the source language or *shining-through* exists in embedding-based semantic space, and if so, whether it can be detected in an unsupervised manner, which leads us to formulate our third research question:

RQ3: Is it possible to track translationese in semantic spaces in an unsupervised manner?

To answer this question, we propose a novel approach for unsupervised tracking of translationese in semantic spaces, which does not rely on explicit linguistic labels. Our method is based on graph-isomorphism approaches that examine departures from isomorphism between embedding spaces built from original target language data and translations into this target language. By comparing the normalised distances between these spaces, we are able to identify systematic evidence of translationese. Specifically, we find that as isomorphism weakens, the linguistic distance between etymologically distant language families increases, providing evidence that the translationese signals are linked to source language interference.

In a multilingual context, non-isomorphic embedding spaces have been attributed to both typological disparities between languages and a poorly conditioned training setup (Vulić, Ruder, and Søgaard, 2020). More specifically, it has been conjectured that there exist typological discrepancies between the languages reflected in them having different structural or syntactical characteristics — making it difficult to align the semantic spaces. One of the other reasons is the poorly conditioned training setup, meaning that the model was not trained on sufficient or diverse enough data to accurately represent the semantics of the language.

In a monolingual setting, the concept of isomorphism seems not directly applicable because there is after all only one language involved. However, here we have two versions of a target language: originally authored target language and translations into the target language, and we investigate divergence from isomorphism between independently computed representations of the two. To explore this, we examine

our graph-isomorphism methods on a variety of data and modelling conditions, encompassing data size, type, and choice of word embedding models. Additionally, we expand our research to non-Indo-European languages and show that our isomorphism approaches are language-independent, in the sense that they can be applied to multiple languages and are not limited to a specific language or language family.
The rest of our chapter is structured as follows. We begin by examining related research in Section 5.2. In this section, we focus on vector language representations and language phylogeny inference. In Section 5.3, we introduce the concept of graph isomorphism and present our novel approach, the Spectral Graph-based measure (SGM), along with Eigenvector Similarity (EV) and Gromov Hausdorff (GH) distance. We explain our experimental setup in Section 5.4. Then, in Section 3.6, we provide an experimental account of the isomorphism metrics, deduce language family relationships, and relate them to linguistic benchmarks. We conduct robustness experiments in Section 5.5.4 and contrast our work with prior research in Section 5.5.3. Finally, in Section 5.5.5, we broaden our analysis to non-Indo-European languages and summarize and conclude our study in Section 3.7.

5.2 Related Work

Quantifying similarities between languages has been a focus in cross-lingual studies. Cross-lingual representations are frequently evaluated by comparing distances in vector spaces with cross-lingual semantic similarity judgments made by humans (Cer et al., 2017). Until recently, there was no formal way to measure the degree to which common word relations[1] and isomorphism, which are underlying assumptions in cross-lingual alignment methods, actually hold true between pairs of languages. Søgaard, Ruder, and Vulić (2018) proposed a measure of the geometric similarity of embedding spaces in the form of Laplacian eigenvalue differences, which was inspired by the approach proposed by Shigehalli and Shettar (2011). Later, Patra et al. (2019) derived another metric which measures the similarity of two embedding spaces a priori using Gromov-Hausdroff distance, and shows a strong correlation with Bilingual Lexicon Induction (BLI) performance. Wendlandt, Kummerfeld, and Mihalcea (2018) investigated the stability of word embeddings through various random initializations. To measure the stability of the space, they utilised the overlap between the top K nearest neighbours of each word in the space. To the best of our knowledge, our work

1 The term "common word relations" refers to the expectation that when translating words between languages, a direct, one-to-one correspondence between words exists, indicating shared semantic or syntactic relationships. However, this assumption does not always hold true. For instance, in closely related languages like English and French, translating a word like "my" from English may result in multiple target translations such as "mon", "ma," and "mes," influenced by factors like gender, case, and number. Similarly, English verb conjugations like "goes" and "go" may correspond to various conjugations in French, depending on factors like person, number, case marking, agglutination, gender, and articles, among others.

is the first instance of spectral graph theory (Tenenbaum, De Silva, and Langford, 2000) being used to compute departures from isomorphism between representations of two forms of a target language: originally authored target language and translations into the target language.

Recent studies have used computational typology to investigate differences between languages. Kolmogorov complexity metrics, for example, were utilised in statistical modelling of linguistic distances (Kettunen et al., 2006). Seminal works of Baroni, Dinu, and Kruszewski (2014), Eger, Hoenen, and Mehler (2016), and Thompson, Roberts, and Lupyan (2018) provide evidence that semantic alignment between languages correlates with the geographical distances between countries as well as with cultural distances among societies speaking the languages.

Relations between languages can also be measured in monolingual original and translationese representation spaces. Rabinovich, Ordan, and Wintner (2017) clustered feature vectors with linguistically inspired features capturing morphological and syntactic information extracted from originally authored and translated texts (in the same target language) and show that phylogenetic language trees can be reconstructed from the clusters. Bjerva et al. (2019) use the same dataset as Rabinovich, Ordan, and Wintner (2017) in their neural language model (NLM)- and sequence-based approach and argue that representation distance between languages can be better explained by structural relatedness than by language genetics. Nikolaev et al. (2020) examined the relationship between the predictability of translated texts and the differences in morpho-syntax between the target and source language. The findings demonstrated that translations from similar and distant languages were both predictable, but in different ways: structurally-similar source languages preferred the use of a smaller range of syntactic patterns limited to those shared by two languages, which represented one type of translational specificity. However, when translating from very diverse languages, translators frequently created non-idiomatic versions that models trained on the target language did not recognize. Several other analyses (Malaviya, Neubig, and Littell, 2017; Oncevay, Haddow, and Birch, 2020) have attempted to disentangle the typological factors that influence language representation distance.

At the same time, the similarity between languages can also be measured using the (dis)similarities between their discrete linguistic properties. Such properties are typically handcrafted and collected in typological databases such as URIEL (Littell et al., 2017) which lists a large inventory of properties for 8000 languages of various typological characteristics, such as overlap in syntactic features, or proximity along phoneme features (Cysouw, 2013).

In this chapter, we investigate whether (i) translationese can be observed in embedding-based semantic space, using (ii) unsupervised approaches in the form of departures from isomorphism between spaces built from original target language

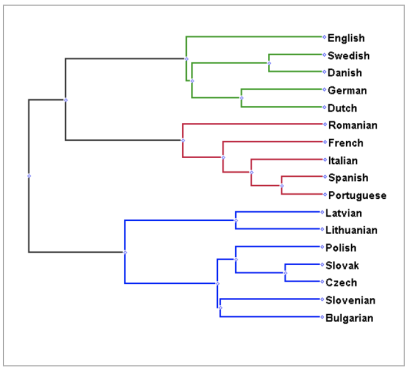

Figure 11: Pruned gold tree of Serva and Petroni (2008), SP08 in text from Rabinovich, Ordan, and Wintner (2017).

and translations into this target language. Thus, in the next section, we introduce graph-isomorphism methods.

5.3 Isomorphism

In graph theory, two graphs X and Y are isomorphic if and only if there exists a bijective function $f : X \to Y$ between them, such that each node $x \in X$ has exactly one counterpart $y \in Y$, and for any two nodes $x_1, x_2 \in X$ with an edge between them, there must be a corresponding edge between $f(x_1)$ and $f(x_2)$ in Y, and vice versa.

Given a matrix of word embeddings, a graph representation can be constructed by defining the nearest neighbor relations for each embedding in the embedding space X. Specifically, for each $x \in X$, a directed graph[2] can be formed such that the edges represent the cosine nearest neighbors to x.

To test for isomorphism between two embedding spaces X and Y, we can represent both spaces as adjacency matrices $A(X)$ and $A(Y)$, respectively. Then, we can examine whether a bijective function exists that can map the adjacency matrix $A(X)$ to $A(Y)$. If such a function exists, then the two embedding spaces are isomorphic.
We introduce our method to quantify a notion of *distance* between languages based on the word-usage using the notion of isomorphism. For a vocabulary $V = v_0, v_1, ..., v_n$ in language ℓ, we define its graph as $G(V, E, w)$, where V denotes the set of vertices corresponding to the vocabulary words; $E = e_0, e_1, ..., e_m$ is a set of edges; and $w(e_i)$ are non-negative edge weights.

After mapping words onto points $\mathbf{v_i}^{\ell}$ as vectors, the distance between words is defined as the distance between their vectors. We quantify the similarity between languages ℓ_1 and ℓ_2 through a distance function between their graphs $d(G^{\ell_1}, G^{\ell_2})$.
The structural congruence in graphs, under isomorphism, naturally extends to metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, highlighting a fundamental mathematical correlation,

2 The importance of directed edges lies in accurately representing asymmetric nearest neighbor relations.

where the vertex set of a graph, coupled with the standard vertex distance function, defines a metric space. Here, $\mathcal{X}$ and $\mathcal{Y}$ represent sets of words, and $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are metric distances. Our goal is to establish an isomorphism $f : \mathcal{X} \to \mathcal{Y}$ that serves as a distance-preserving transformation. This means that for any points x_1 and x_2 in $\mathcal{X}$, the distance between $f(x_1)$ and $f(x_2)$ in $\mathcal{Y}$ mirrors the distance between x_1 and x_2 in $\mathcal{X}$ such that, $d_{\mathcal{Y}}(f(x_1), f(x_2)) = d_{\mathcal{X}}(x_1, x_2)$.

5.3.1 Gromov-Hausdorff (GH) Distance

The first measure we use to quantify the similarity between languages is the Gromov-Hausdorff distance (GH) proposed by Patra et al. (2019).

Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, we start with the Hausdorff distance, defined as:

$$d_H(\mathcal{X}, \mathcal{Y}) = \max \left\{ \sup_{x \in \mathcal{X}} d(x, \mathcal{Y}), \sup_{y \in \mathcal{Y}} d(y, \mathcal{X}) \right\} \tag{11}$$

where $d(a, \mathcal{B}) = \inf_{b \in \mathcal{B}} \| a - b \|_2$ is the distance of point a in $\mathcal{A}$ from set $\mathcal{B}$. Informally, d is the largest distance needed to travel from a point in $\mathcal{A}$ to a point in $\mathcal{B}$.

However, the Hausdorff distance is easily affected by isometric transformations. The GH distance which is the infimum of the Hausdorff distances under all possible isometric transformations is a more robust measure. GH distance reduces the distance over the isometric transforms f and g between $\mathcal{X}$ and $\mathcal{Y}$ as follows:

$$d_{GH}(\mathcal{X}, \mathcal{Y}) = \inf_{f,g} d_H(f(\mathcal{X}), g(\mathcal{Y})) \tag{12}$$

The computation of Hausdorff distance is NP-hard, and hence we follow Patra et al. (2019) and compute the Bottleneck distances (Chazal et al., 2009) which are considered to be reasonable lower-bounds.

5.3.2 Eigenvector Similarity (EV)

Søgaard, Ruder, and Vulić (2018) used this similarity to measure the distance between two embedding matrices corresponding to two languages ℓ_1 and ℓ_2, via their Laplacian matrices, $\mathcal{L}$. They argue that the Laplacian eigenvalues are good compact representations for the graph Laplacian and their comparison can consequently capture the degree of isomorphism.

Let A be an adjacency matrix representing the nearest neighbor relations of the top k most frequent terms in an embedding space. Let D be the degree matrix of this adjacency matrix, which is a diagonal matrix representing the number of edges into each node in the graph. Then the Laplacian matrix $\mathcal{L}$ is then given by $D - A$.

Once the Laplacian matrix is obtained, its eigenvalues are calculated using eigen-decomposition to determine the spectrum of a graph, which provides an estimate of the interconnection between the nodes of the graph. Finally, Søgaard, Ruder, and Vulić (2018) take the square difference of the two sets of eigenvalues corresponding to two languages ℓ_1 and ℓ_2 to obtain the squared eigenvalue difference using the following formula to obtain their EV measure:

$$\Delta = \sum_{i=1}^{k} (\lambda_{1i} - \lambda_{2i})^2 \tag{13}$$

We apply the same idea to characterize distinctions between two spaces, denoted as $\mathcal{X}$ and $\mathcal{Y}$, representing the original data in the target language and translations from various source languages into the same target language. We first find the smallest k_1 in Equation 13 such that the sum of its k_1 largest eigenvalues $\sum_{i=1}^{k_1} \lambda_{1i}$ is at least 90% of the sum of all its eigenvalues. Analogously, we find k_2 and set $k = \min(k_1, k_2)$.

The graph similarity metric returns a value in the half-open interval $[0, \infty)$, where values closer to zero indicate better isometry between the two languages. Or in other words, if the value of Δ is closer to zero, it indicates that the eigenvalues of the two Laplacian matrices are more similar, and therefore the two spaces corresponding to two languages ℓ_1 and ℓ_2 are more likely to be isomorphic. This means that the distance between any two points in one space is approximately equal to the distance between their corresponding points in the other space, indicating a better isometry between the two spaces. On the other hand, if the value of Δ is larger, it indicates that the eigenvalues of the two Laplacian matrices are more dissimilar, and therefore the two spaces are less likely to be isomorphic. This means that the distance between points in the two spaces may not be preserved, leading to a poorer isometry between the languages.

5.3.3 Spectral Graph-based Matching (SGM)

Our third and new measure is based on the `Isomap` algorithm of Tenenbaum, De Silva, and Langford (2000).

Our method differs from that of Søgaard, Ruder, and Vulić (2018) in one key way: our method to build the underlying graphs (G^{ℓ_1} and G^{ℓ_2}) differs and is inspired by the ideas of node representation in contemporary geometric and manifold learning (Cayton, 2005). While EV (Søgaard, Ruder, and Vulić, 2018) samples subgraphs of cross-lingual word pairs from the source and target language by computing the cosine similarity, we build a weighted connected graph over the data points to capture better neighbourhood relations and perform the spectral analysis on these.

Weights w_{ij} correspond to the distance between points i and j in the input space $(\mathcal{X}, d_{\mathcal{X}}(i, j))$. We connect each point only to its K nearest neighbours to consider more

Algorithm 2 Isomap (Tenenbaum, De Silva, and Langford, 2000)

Input: Data vectors $x_1, x_2, ..., x_n \in \mathbb{R}^d$
Output: Lower-dimensional representation preserving data structure

Step 1: Construct Neighborhood Graph
Define graph G by connecting points i and j if $d_X(i,j) < \epsilon$ (ϵ-Isomap) or if i is one of the K nearest neighbors of j (K-Isomap). Set edge lengths equal to $d_X(i,j)$.

Step 2: Compute Shortest Paths
Initialize $d_G(i,j) = d_X(i,j)$ if i,j are linked by an edge; $d_G(i,j) = \infty$ otherwise.
for $k = 1, 2, ..., N$ **do**
Update $d_G(i,j) = \min(d_G(i,j), d_G(i,k) + d_G(k,j))$ for all i,j
end for
$D_G \leftarrow$ matrix of final shortest path distances between points in G

Step 3: Construct d-dimensional Embedding
Compute eigenvalues λ_p and eigenvectors v_p^i of matrix $\tau(D_G)$
for $p = 1, 2, \ldots, d$ **do**
Set the p-th component of the d-dimensional coordinate vector y_i equal to $\sqrt{\lambda_p} v_p^i$ for all data points.
end for

geometrical information on the interaction between all vectors within the initial space to improve the graph characterisation of the spaces. The value $K(= 6)$ is chosen to have similar edge density for all graphs. This is done for convenience, and while this selection may seem somewhat arbitrary compared to epsilon neighbors[3], its advantage lies in capturing more contextual information from the interaction between data points and their neighborhoods. We estimate the geodesic distances between vertices (points) in the input space using shortest-path distances obtained with Dijkstra's algorithm (Dijkstra et al., 1959) on the constructed graph to minimize the sum of the weights of their constituent edges. This resulting distance matrix captures the geometrical information on the interaction between all vectors within the initial space, and is used as the basis for our graphs. From this point onwards, the computation of the Laplacian matrices and the final measure Δ is performed in a similar fashion as Section 5.3.2.

5.4 Experimental Setup

In this section, we provide information on the data, and the vector spaces used for computing deviation from isomorphism.

3 Epsilon neighbors define a neighborhood for a data point based on a specified radius or distance threshold ϵ. Data points within this radius become its neighbors. Thus, the epsilon neighbors approach adapts to the local data density, potentially capturing intricate local structures more effectively.

5.4.1 Datasets

We use the same setup as Rabinovich, Ordan, and Wintner (2017), Bjerva et al. (2019), and use the comparable portion of Europarl (Koehn, 2005) with translations from 21 European Union languages into English to minimize the impact of domain difference. The tokens per language vary, ranging from 67 k tokens for Maltese to 7.2 M for German. We refer to the multiple translations into English as L_j's, where j=1,2,...,n; and to originally written text in English as L_e. This dataset is a superset of the one described in Section 3.3.1; therefore, we do not conduct the stylistic analysis carried out in Section 3.3.

We select a subset of translations from 16 languages covering four families: *Romance* (French (fr), Italian (it), Spanish (es), Romanian (ro), Portugese (pt)), *Germanic* (Dutch (nl), German (de), Swedish (sv), Danish (da)), *Slavic* (Czech (cs), Slovak (sk), Slovenian (sl), Polish (pl), Bulgarian (bg)) and *Baltic* (Latvian (lv) and Lithuanian (lt)) into English and English original (en) text.

For these 17 data sets, we define two settings, the *full data* condition and the *small data* condition. The former makes use of the complete data in the Europarl edition available for a language (recall that data size differs widely); for the latter, we randomly extract m sentences, where m corresponds to the lower-bound data-size of our translated data, i.e., the size of the Latvian corpus (118,525 words). We report results for the *full data* setting and use the *small data* for robustness checks and comparisons with existing literature.

5.4.2 Vector Spaces

Our data are original English L_e or translations from language j into English L_j's. To obtain monolingual word embeddings for each dataset, we train models under two conditions: *full* and *small data*. We tokenize and lowercase the data using Moses (Koehn et al., 2007), and then apply ***fastText*** (Bojanowski et al., 2017) to generate the embeddings. We select words that appear more than 5 times in the corpus and train 300-dimensional embeddings. We use skip-gram with negative sampling (Mikolov et al., 2013) with standard hyperparameters, such as character n-grams of sizes 3 to 6 and a learning rate of 0.025. Finally, we mean center and unit normalise the embeddings.

5.4.3 Typological Benchmarks

We estimate our isomorphism based results against the following two benchmarks:

SP08 (Serva and Petroni, 2008). SPO8 is constructed by employing a pure lexicostatistic method to measure distances between pairs of words in different languages. To construct a distance between pairs of languages, the Levenshtein (edit) distance is computed between words of an open cross-lingual list (Dyen, Kruskal, and Black, 1992) which contains the Swadesh list of 200 words for 96 languages for each pair of words with the same meaning in one language pair. The distance of each language pair is defined as the average of the distances between the word pairs resulting in a number between 0 and 1 and is considered to be the lexical distance between the two languages. These distances are then used to construct phylogenetic trees.

Rabinovich, Ordan, and Wintner (2017) use the pruned gold-tree of Serva and Petroni (2008) as their gold standard, as shown in Figure 11. Their comparison metric, is based on the L2 norm, which basically is the sum of squared deviations between each pair's gold-tree distance g (SP08) and their computed distance P based on linguistic features:

$$\mathrm{Dist}(P, g) = \sum_{i,j} \left(D_P(l_i, l_j) - D_g(l_i, l_j)\right)^2 \tag{14}$$

URIEL. The URIEL knowledge base is a compilation of various linguistic resources, including the World Atlas of Language Structure (WALS) (Dryer, 2009), PHOIBLE (Moran, McCloy, and Wright, 2014), Ethnologue (Lewis, Simons, and Fennig, 2015), and Glottolog (Nordhoff and Hammarström, 2011). These resources provide information on various linguistic features of languages, such as phonology, syntax, and grammar. Based on these linguistic feature vectors, URIEL provides precomputed distance statistics between any language pairs stored in the database in terms of various metrics including genetic, geographical, syntactic, and phonological inventory distances. These distances are often deemed to be very useful for guiding cross-lingual transfer tasks (Agić, 2017; Cotterell and Heigold, 2017; Lin et al., 2019).

5.5 Evaluation and Analysis

Our analysis below provides evidence that translationese, here specifically the interference of source languages in translated text, can be traced in semantic word embedding spaces. If we observe a weakening of isomorphism going hand in hand with an increase in the linguistic distance between etymologically distant language families, this provides evidence that the signals of translationese are linked to source language interference, answering **RQ3**.

We employ our isomorphism measures for two tasks: (i) inferring language families and (ii) conducting correlation analyses against two benchmarks, SP08 and URIEL. In addition, we compare our graph-isomorphism methods to previous approaches that

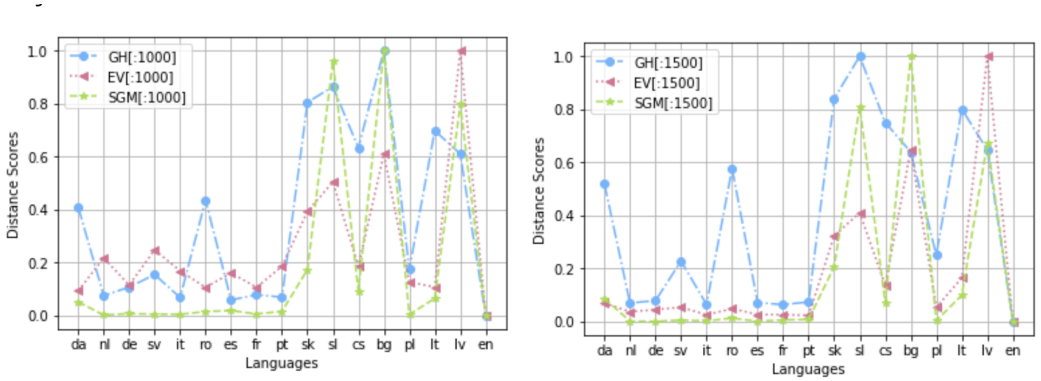

Figure 12: Normalised distances between embedding spaces for original en and translations into en from 16 languages given by our three distance measures using 2 different numbers of data points.

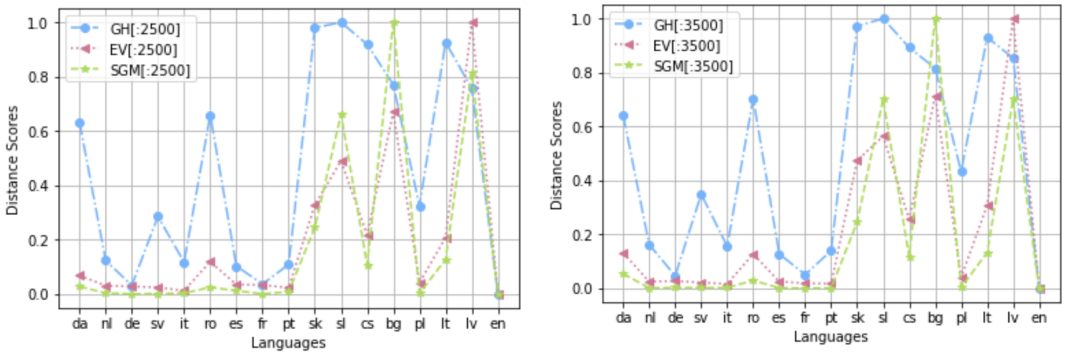

Figure 13: Normalised distances between embedding spaces for original en and translations into en from 16 languages given by our three distance measures using 2 different numbers of data points.

rely on linguistic assumptions. Furthermore, we assess the robustness of our proposed methods and investigate their applicability to non-European languages.

5.5.1 Language Distance Measures

First, we conduct an experiment to determine the distance between vector spaces created from original target language (English) data and translations into the target language. Specifically, we calculate each metric over the top-n-most common words shared between originals in the target language and translations into the target language, where we explore the impact of different graph sizes by varying n between 1000, 1500, 2500, and 3500. It should be noted that achieving isomorphism requires having the same number of components (i.e., vertices and edges) in the graphs.

Figures 12 and 13 depict the results of our analysis using normalised distances. We observe that the behavior of the metrics varies depending on the number of top-n most frequent data points considered. Specifically, SGM is the most stable measure across all configurations, with the highest variance observed for 1000 points. EV exhibits greater variability in distinguishing language similarity, whereas GH results are relatively

stable when considering larger datapoints (2500 and 3500 points).
These variations can be attributed to the different nature of the metrics. GH computes the distance between spaces solely based on the subset of n words, while SGM weights the nearest neighbors of a word, which may lie outside the top-n, to construct the initial graph as the selection of nearest neighbours is not constrained to the top-n words. SGM connects each data point (word) to its K nearest neighbors in the input space, regardless of whether these neighbours fall within the top-n words or not. Thus, this approach takes more context into account for each point and requires fewer datapoints to successfully describe the space. As expected, the results of EV and SGM are more similar to each other than to GH since they follow a similar methodology, except that SGM retains more context than EV.
Across all plots, we note the smallest differences between original English and translations from Germanic, followed by Romance languages, indicating that vector spaces of these languages are more similar to each other in terms of semantic embedding space based isomorphism measures. We also observe consistent weakening of isomorphism with increased linguistic distances, specifically for Baltic and Slavic families, regardless of the method used, providing evidence that the distance between languages in semantic space is greater for etymologically distant language pairs.
However, there are some outliers that vary from one measure to another. GH places Romanian far from other Romance languages, while Danish and Swedish are located relatively far from other Germanic relatives. In contrast, SGM (and to a lesser extent EV) places Polish close to English. These discrepancies might point towards influences beyond linguistic family relations, potentially stemming from geographical proximity or distance, cultural influences, and historical interactions.

5.5.2 Correlation with Typology, Geography and Phylogeny Benchmarks

In this section, we estimate the Kendall correlations [4] between our results based on the difference from isomorphism and SP08 (which represents genetic similarities), as well as between our results and the averaged URIEL features introduced in Section 5.4.3 (which represent other rich typological similarities beyond genetics). The Kendall correlation between SP08 and the selection of URIEL features is 0.56 reflecting the different nature of the two benchmarks. Although both attempt to capture genealogical differences, the source for this kind of information is different. Our results, summarised in the top rows of Table 14, show that correlations with URIEL are higher than with SP08, demonstrating that other factors besides genetics are reflected in the semantic spaces. SGM reproduces the genetic SP08 benchmark better than EV and GH, while

4 The Kendall correlation coefficient (Kendall, 1938) is a measure of the strength and direction of association between two variables. It is a non-parametric statistical test, which means it does not assume that the data follows a particular distribution. It is often used to assess the relationship between ordinal (ranked) data.

# Points	*SPo8*			*URIEL*		
	GH	**SGM**	**EV**	**GH**	**SGM**	**EV**
Full data condition						
1000	0.32	0.52	0.43	0.39	0.38	0.24
1500	0.40	0.30	0.26	0.51	0.31	0.36
2500	0.42	0.38	0.39	0.57	0.36	0.43
3500	0.40	0.39	0.31	0.58	0.45	0.36
FW	0.40	0.30	0.32	0.44	0.45	0.30
Swadesh	0.30	0.32	0.11	0.36	0.43	0.09
Small data condition						
1000	0.11	0.45	0.21	0.21	0.39	0.20
1500	0.29	0.37	0.21	0.49	0.27	0.21
2500	0.39	0.45	0.23	0.40	0.39	0.30
3500	0.27	0.46	0.11	0.36	0.35	0.12

Table 14: Mean Kendall correlations with SPo8 and average URIEL for various number of datapoints and the function words experiment (FW).

	Rabinovich *et al.* 2017	**Bjerva** *et al.* 2019	**Our proposal** Full	Small
Words	–	0.53	✓	✓
FW	0.43	–	–	–
POS	0.35	0.52	–	–
FW+POS	0.36	0.56	–	–
PS	–	0.36	–	–
DepRel	–	0.32	–	–
GH	–	–	0.37	0.38
EV	–	–	0.57	0.56
SGM	–	–	0.54	0.58

Table 15: Mean distance between SPo8 and reconstructed phylogenetic trees as compared to previous literature using words, function words (FW), parts of speech (POS), phrase structures (PS) and dependency relations (DepRel) as features.

GH clearly correlates better with structural URIEL features, followed by SGM and EV. This corroborates NLM-based findings of Bjerva et al. (2019) in our semantic word embedding-based spaces: the differences and similarities between languages and language representations go beyond genetic (dis)similarities.

5.5.3 Comparison with Previous Approaches

In order to compare our results with previous work, we calculate tree distances using the leaf-node distance in Equation 14 previously defined in Rabinovich, Ordan, and

Wintner (2017) and compare with the best results on SP08 in Rabinovich, Ordan, and Wintner (2017) and Bjerva et al. (2019). We report our results for 1500 most frequent datapoints obtained with different metrics in Table 15. Notice that the results of Table 14 and Table 15 cannot be directly compared as the later one is computed after summing over all possible pairs of the leaves (languages), while Table 14 shows the association with benchmarks (SP08 and URIEL) keeping only originally authored English as its source. Table 14 correlates the results with the benchmarks and Table 15 compares the findings to Bjerva et al. (2019) and Rabinovich, Ordan, and Wintner (2017). All distances are normalised to a zero-one scale.

According to the mean distance, our embedding and graph-based approaches, especially GH can reproduce genetic trees on-par with previous work without requiring any explicit linguistic information. Unlike previous methods which rely on surface-level features of the source language, our graph-based isomorphism analysis is unsupervised and still is able to detect important language differences related to linguistic distances. Of all the methods, GH is the closest to SP08, followed by SGM and EV in the *full data* settings while the trend for SGM-EV is reversed under the *small data* condition.

5.5.4 Robustness Analysis

After showing that exploiting departures from isomorphism between spaces can be used to predict relations between languages, we analyse the impact of various modeling assumptions and different training conditions that might have an effect in skewing the results of Table 15.

5.5.4.1 Data Size Effects

Large differences in data sizes between high and low-resource languages have played a pivotal role in the performance of monolingual embeddings (Sahlgren and Lenci, 2016; Vulić, Ruder, and Søgaard, 2020). In our work, to some extent this is already minimised by taking only the most frequent n words to estimate the distances between embedding spaces, but still the quality of even these embeddings might differ. To examine the impact of the data size for our experiments, we use the embeddings obtained under the *small data* condition (see Section 5.4) and compare the results in the bottom rows of Table 14.

The results show that SGM correlates best with SP08 under all training conditions (number of datapoints and corpus size), but the correlation decreases with respect to URIEL features. GH shows good correlation in some instances (1500 and 2500 datapoints) for both SP08 and URIEL, while EV shows no consistent correlation. For EV, we consider frequent words and mutual nearest neighbors, thus in the *small data* condition, it has even less access to contexts. Our spectral graph-based measure SGM,

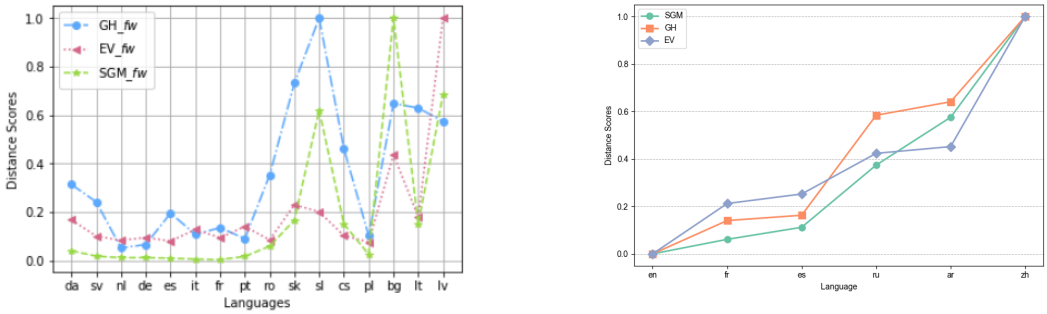

Figure 14: Normalised distances between original texts and their translationese counterparts in English (en). Embedding spaces are constructed using only function words from the Europarl corpus (left) and the UN parallel corpus (Tolochinsky et al., 2018) (right).

which is inspired by the ideas of node representation in contemporary geometric and manifold learning (Cayton, 2005), provides better robustness with respect to measuring linguistic distances under varied data settings.

5.5.4.2 Word Embedding Models

Köhn (2015) showed that different methods to obtain word embeddings (CCA, skip-gram, CBOW, GloVe, etc.) behave similarly when capturing syntactic and morphological information. We check that this is also the case with our distance methods by comparing the performance obtained with skip-gram and CBOW architectures, and observe small variations with similar global trends. To give an example, correlation results for SGM under the full data condition with CBOW and skip-gram vary only in the ± 0.05 range.

We also performed experiments with lower embedding dimensions (50,100,200) which may lead to reduced expressivity, but, very interestingly, we obtained similar performance as we did with 300 dimensions. For example, comparing 100-dimensional monolingual word embeddings with 300 dimensional embedings, the differences are: GH (± 0.077), SGM (± 0.013), and EV (± 0.088).

5.5.4.3 Function Words

Function words, also known as closed class or grammatical words, are words that do not carry lexical meaning but perform important syntactic functions in a sentence. By experimenting with function words, we can get a better understanding of how our different approaches to constructing graphs affect the measurement of isomorphism between languages and whether these approaches are able to capture the nuances of language use beyond just distributional patterns of word usage. Thus, we focus on function words which e.g. introduce and identify key discourse referents and represent relationships between entities but are considered to be not well-modelled

by distributional semantics (Bernardi et al., 2015). Specifically, we use the list of function words defined in Koppel and Ordan (2011) to construct the language distance measures of Section 5.5.1 in Figure 14. In this case, the number of data points is 468, well below the minimum number of points used with content words (1000). However, the performance of all three methods (GH, EV and SGM) show similar trends as in Figures 12 and 13. Figure 14 demonstrates that function words are able to capture departures from isomorphism in a similar way as the complete set of words, indicating that source languages carry over grammatical constructs into the translation product, corroborating prior findings of Rabinovich, Ordan, and Wintner (2017) and Bjerva et al. (2019) with function words.

Additionally, we explore the much smaller cognate collection of Swadesh word lists (Swadesh, 1952) to capture the relatedness between languages in Table 14. As this concept-aligned resource ensures a consistent set of word lists across all our languages, thereby enhancing comparability, these findings are particularly important. The results show that the large context (6 neighbours) exploited by SGM estimations exceeds other isomorphism methods while highlighting the limitations of EV in low-data regimes with limited access to contexts.

5.5.5 Results on Non-European Languages

Previous research focused on investigating translationese and source language interference for European language families Bjerva, Plank, and Bos (2016) and Rabinovich, Ordan, and Wintner (2017). Here we extend this work, for the first time, to the best of our knowledge, to translations from non-Indo-European languages into English. We investigate the language distance measures of Section 5.5.1 on the UN corpus (Tolochinsky et al., 2018) which consists of translations covering typologically different languages such as Arabic (ar) and Chinese (zh), as well as Indo-European languages (i.e., Russian (ru), Spanish (es) and French (fr)).

We present results with 2500 words in Figure 14. Figure 14 shows that the vector spaces constructed from translations of Romance languages (i.e., fr and es) are closer to original English in terms of semantic embedding space based isomorphism measures. However, the difference between original English and translations from non-Romance languages (i.e., ar, zh, and ru) are larger, indicating a larger distance between their vector spaces and original English. This is in line with our previous results in Figures 12 and 13 on the same-domain monolingual Europarl data under different data settings, demonstrating that the language distance measures we used are language-independent and can be applied to multiple languages. These findings suggest that our methods are not limited to a specific language or language family, and can be used to analyse the translations of typologically different languages.

Of all measures, SGM, which captures more context from the interaction between data-points and their neighborhoods, is the best indicator of linguistic (dis)similarities and explains more variance than previous isomorphism measures, as shown in Figure 14. Among the non-Romance languages, translations from Chinese (*zh*) has the highest values for all three measures, indicating a larger distance between its vector space and original English compared to Arabic (*ar*) and Russian (*ru*).

5.6 Conclusion

In this chapter, we introduce a novel unsupervised approach to tracking signals of translationese in word embedding spaces. Our method utilises graph-isomorphism measures to trace translationese without relying on explicit linguistic labels or surface-level features, answering **RQ3**. This new approach enables us to trace translationese in embedding spaces that have not been explored before. Specifically, our approach lets as much as possible "lets the data speak for itself" uncoloured by manual feature engineering in detecting important language differences related to linguistic typology. We provide systematic evidence of translationese in embedding-based semantic spaces, which is a novel contribution to the field.

To demonstrate the effectiveness of our approach, we compare the distances between graphs for original English and translations into English from 16 different languages. Our results align with expectations, showing a greater distance between English and translations from Slavic languages and a smaller distance between English and translations from Germanic languages providing evidence of source language interference. We also compare our results to two benchmarks: a language phylogeny tree based on structural features (URIEL) and a tree based on a combination of genetic and structural features (SP08) and found that our novel spectral graph-based approach (SGM) performs as well as and in low resource scenarios better than previous approaches that rely on linguistic features.

Finally, we perform robustness tests to ensure that our methods are stable and reliable under different modelling and training conditions. This includes testing the robustness of our results to variations in the size and type of the data (function words used, as well as the robustness of our methods to different choices of word embedding models, data size and type. By demonstrating the stability of our methods, we show greater confidence in the validity of our results and the generalisability of our findings. In addition to this, we not only investigate the phenomenon of translationese in Indo-European languages but also extend our analysis to non-Indo-European languages (Chinese and Arabic). This allows us to determine if the tracing of translationese is specific to certain language families.

6

Influence of Domain on Translationese using Multi-View Semantic Spaces

In this chapter, we evaluate the impact of domain on our unsupervised tracking of translationese in semantic spaces. We aim to determine whether the results presented in the previous chapter are primarily a result of domain variations between the original and translated data, rather than, or to a lesser extent, from proper translationse signals. Translationese signals are subtle and compete with other signals in the data, particularly those related to domain. To address this, we mask domain information by applying our graph-isomorphism methods to different delexicalized representations using various views, including words, parts of speech, semantic tags, and synsets. Our results demonstrate that while source-language interference is most pronounced in lexicalised embeddings (reported in the previous chapter), it is also present in delexicalised views. This indicates that our lexicalised findings are not just the result of possible domain differences between original and translated texts.

Regardless of the level of linguistic representation, we present clustering experiments that show that various translationese "dialects" in the same language, arising from translations originating from different source languages, exhibit a clustering pattern akin to linguistically motivated phylogenetic trees, aligned with the source languages. In addition, we show that the distinction between language families in the translated texts partially originates from the typological characterisation of the source languages.

6.1 Introduction

In the previous chapter (Chapter 5), we explore the impact of source language on translations from this source language into the same target language as compared to originally authored text in the target language using departures of isomorphism in semantic spaces. We find that embedding spaces created from raw word representations exhibit traces of the source languages of the translations. However, translationese signals are subtle and can compete with other signals in the data, particularly those related to domain. Domain is often reflected in lexical signals. Thus it is important to examine the robustness of the results obtained in our previous research under

delexicalisation in order to determine whether the patterns identified in the data are truly indicative of translationese, or whether they may be influenced by other factors such as domain differences between the original and translated texts. This leads us to formulate our next research question.

RQ4: To what extent can the outcomes observed in response to RQ3 be attributed to variations in domain between original and translated texts, as opposed to true translationese signals?

To address this, we mask lexical domain information and apply our graph-isomorphism methods to different delexicalized representations using various views and quantify departure from isomorphism for each of the multiple views of the same underlying data. Specifically, we are using different manifestations of the same data based on the words (Raw), Part-of-Speech (PoS), Semantic Tags (ST), and Synsets (SS) to create a multi-view of the embedding spaces. Using these multi-views can help to reduce the influence of the lexical aspects of domain on the results. For example, using part-of-speech tags as a view may be less sensitive to domain-specific words than using Raw word representations because PoS tags capture grammatical information rather than lexical content. This means that even if the texts (i.e., original and translated) are about different domains and might use different sets of words corresponding to those domains, the part-of-speech tags would be more similar across the texts, allowing us to compare the non-lexical translationese characteristics of the texts. The further apart the embedding spaces for a given view are, the more predominant the information captured by this view is in translationese.

Our results demonstrate that while source-language interference is most pronounced in lexicalised embeddings (word), it is also present in delexicalised views. This supports our claim that the delexicalised embeddings are just as able as lexical ones to identify important translationese artifacts, providing evidence that our lexicalised results capture more than just possible domain divergences between originals and translated texts. We observe that in particular, a larger departure from isomorphism in semantic space computed from delexicalised views, by and large, corresponds to a larger distance in terms of linguistically motivated language families and that language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from the isomorphism distances even for delexicalised views.

The rest of the chapter is organised as follows. We review related work in Section 6.2. Section 6.3 briefly outlines data settings, views and methods used in the study. In Section 6.4, we describe our evaluation methods. Specifically, Section 6.4.1 outlines the classification task using different views of the data and Section 6.4.2 shows the hierarchical clustering of the results obtained from our approaches. Finally, we analyse

the source language interference properties in Section 6.4.3 and draw conclusions in Section 6.5.

6.2 Related Work

A significant amount of research in historical linguistics and linguistic typology seeks to quantify how languages are related to one another. One way of visualizing the relations between languages are phylogenetic trees. Computational analysis techniques attempt to automatically construct phylogenetic relationships between languages based on linguistic patterns. For example, Nouri and Yangarber (2016) constructs phylogenetic relations based on the lexical overlap between languages using manually compiled lists of cognates. There are a large number of techniques for automatically extracting cognates (Serva and Petroni, 2008), however, these methods rely on the surface structure of words and are a time-consuming and somewhat subjective procedure. Other than quantifiable lexical overlap, structural similarity across languages are often used to measure how (dis)similar they are (Cysouw, 2013).

Rabinovich, Ordan, and Wintner (2017) showed that differences in languages are so strong that vestiges of the structure of a language are approximately preserved even when translating into another language, that is, translations show *translationese* effects from the source language. They show that footprints of the source language remain visible in translations, to the extent that it is possible to predict the original source language from the translation and reconstruct phylogenetic trees. To show this, they experimented with English and French translations of European parliamentary speeches from languages derived from mainly three language families: Germanic, Romance and Balto-Slavic. Bjerva et al. (2019) explored the structural similarity between languages using phrase structure trees and dependency relations. Both approaches were able to generate relatively faithful phylogenetic trees using surface-oriented lexical, morphological and syntactic similarities. The similarity between languages is often measured by their semantic similarities. Eger, Hoenen, and Mehler (2016) induced bilingual vector spaces for 21 Europarl languages and calculated a representation distance between them by averaging pairwise similarity between word representations. They discovered that geographic factors of the countries where these languages are spoken explain the differences better than phylogenetic factors.

Similar findings have been made by Lupyan and Winter (2018) who found that the semantic similarity of languages is correlated with the cultural distance between them. They propose the idea that the structure and content of language is shaped by the culture in which it is used, and that this cultural influence can be measured by a concept they refer to as "cultural distance". However, the (Youn et al., 2016) note that there are other factors, such as polysemy (the ability of a word to have multiple meanings) and context-sensitivity (the way in which the meaning of a word or phrase

can change depending on the context in which it is used) that can influence semantic similarity across languages.

Recent research that learns cross-lingual word embeddings for downstream tasks such as bilingual dictionary induction or machine translation suggests that phylogenetic features are encoded in the final representations, even if language information is not explicitly provided during training. For example, (Beinborn and Choenni, 2020) demonstrated that phylogenetic relationships across languages can be reconstituted from cross-lingual representations if the training objective optimizes monolingual semantic constraints for each language individually, as in the multilingual MUSE model (Conneau et al., 2017). Along the same line, (Malaviya, Neubig, and Littell, 2017) also probed missing features in typological databases from cross-lingual representations. More recently, Oncevay, Haddow, and Birch (2020) used multi-view language representations for typological feature prediction and language clustering and adapted the ranking of related languages for multilingual transfer.

Linguists usually determine phylogenetic relationships between languages using concept lists of words in multiple languages with a common origin that share a similar meaning and a similar pronunciation (Dyen, Kruskal, and Black, 1992; Swadesh, 1952). Rabinovich, Ordan, and Wintner (2017) reconstructed phylogenetic trees by performing hierarchical language clustering with Ward's variance minimization algorithm (Ward Jr, 1963) with Euclidean distance as a linkage method. Specifically, they leverage interference features (POS trigrams and function words) and one translation universal features (cohesive markers) through the use of classifiers to compute the trees. More recently, Bjerva et al. (2019) built on this work and investigated causal relationships between language representations and similarities from structural similarities (Pearl et al., 2009).

The evolution of phylogenetic trees remains a highly debated domain in the history of linguistics and the quality of phylogenetic language trees has been evaluated using a variety of methodologies (Nouri and Yangarber, 2016; Pompei, Loreto, and Tria, 2011; Wichmann and Grant, 2012). However, there exists no standard agreed on the method that covers all reconstructions (Ringe, Warnow, and Taylor, 2002). Rabinovich, Ordan, and Wintner (2017) proposed comparing their generated trees to a binary Levenshtein-based approximation developed by Serva and Petroni (2008) as a reference tree. This reference tree was constructed by computing the Levenshtein (edit) distance between words of an open cross-lingual list (Dyen, Kruskal, and Black, 1992). Lexicostatistic strategies such as this one typically depend on lists of cognates (Swadesh, 1952) and are not well suited for isolated languages, and are inaccurate in capturing the more subtle variations for the Indo-European language family (Fortson IV, 2011).

6.3 Experimental Setup

6.3.1 Datasets

As in the previous Chapter 5, we experiment with the comparable portion of Europarl (Koehn, 2005) with translations from 21 European Union languages into English (refer to Section 5.4) to minimise the impact of genre, style and to some extent domain difference. The amount of tokens per language in this dataset varies, ranging from 67 k tokens for Maltese to 7.2 M for German. We refer to the multiple translations into English as L_j's, where j=1,2,...,n; and to originally written text in English as L_e. We select the subset of translations from 16 languages covering three language families: *Romance* (French, Italian, Spanish, Romanian, Portugese), *Germanic* (Dutch, German, Swedish, Danish) and *Balto-Slavic* (Latvian, Lithuanian, Czech, Slovak, Slovenian, Polish and Bulgarian) into English and English original text.

Further, to control the impact of corpus size we modify the training setup of a high-resource language to simulate a low-resource scenario. For the departure from isomorphism experiments, for each view, we choose the size of the common overlapped vocabulary list corresponding to the range of low-resource translated data (see Table 16).

6.3.2 Views

We make use of the data described in Section 7.3.1 to create multiple sources of input (views) at the morphological (PoS), syntactic (ST) and conceptual-semantic (SS) levels. Each view provides a different perspective on the data.

To create Part-of-Speech (PoS) annotations, we utilised the SpaCy toolkit (Honnibal and Johnson, 2015). This toolkit provides us with fine-grained morphological information, enabling us to assign PoS tags to each word in the text. There are 37 possible PoS tags, which provide detailed information about the grammatical role of each word in the sentence.

For extracting semantic tags (SemTag), we employed the best model of Abzianidze and Bos (2017), which is language-neutral and abstracts over PoS and named-entity classes. Their implementation achieves an accuracy of around 95% when evaluated on short Parallel Meaning Bank (Abzianidze et al., 2017) sentences.

To obtain conceptual semantic (Synset) annotations, we followed the approach of España-Bonet and Genabith (2018), retrieving synsets based on the PoS of a token using the WordNet knowledge base (Miller, 1998). We selected a subset of PoS tags, namely *NN, ADV, ADJ*, and *VB*, and use the first synset for each word/tag combination.

An overview and examples of each type of annotation used in this study are presented in Table 16. We utilised these multi-view datasets throughout this chapter to

Feature	Annotated Output	Overlapped Vocab
PoS	DET NOUN VERB DET NOUN ADP ADJ NOUN	37
SemTag	DEF CON EPS DIS CON REL HAS CON	57
Synset	ministry.n.04 send.v.02 answer.n.04 inquiry.n.01.	1667
Raw	the ministry sent an answer to our inquiry	6739

Table 16: Examples of the level of abstraction

examine the influence of domain to the task of unsupervised tracing of translationese in semantic spaces.

6.3.3 Multi-view Embedding Representations

We generate embedding representations for each view by creating separate monolingual word embedding spaces for both L_e and L_j's. To accomplish this, we treat each tag as a word and use ***fastText*** (Bojanowski et al., 2017). We produce embeddings with 300 dimensions and only include words that occur more than 5 times. We use skip-gram with negative sampling (Mikolov et al., 2013) with default hyper-parameters for training. This process allows us to create embedding spaces for each view, which is important for addressing **RQ4**.

6.3.4 Isomorphism

We now provide a brief overview of the methods for measuring approximate isomorphism between the embedding spaces computed from our different views in order to quantify the (dis)similarity between languages. Further technical details of these measures are discussed in Section 5.3.

Gromov Hausdorff distance (GH). This distance was described in Patra et al. (2019) to test how well two language embedding spaces are aligned under an isometric transformation. Particularly, it calculates the worst-case distance between two metric spaces, which is the maximum distance of a set of points to the nearest point in another set.

Eigenvector Similarity measure (EV). This measure was introduced by Søgaard, Ruder, and Vulić (2018) based on spectral analysis of the Laplacian eigenvalues of the nearest neighbourhood graphs that result from the initial language embedding spaces. They argue that these eigenvalues are compact representations of the graph Laplacian and their comparison indicates the degree of (approximate) isomorphism.

Spectral Graph Based Measure (SGM). Although similar in spirit to the Eigen-Vector Similarity approach, we introduce this measure (Dutta Chowdhury, España-

Families	Raw	PoS	ST	SS
Germanic	0.72	0.65	0.63	0.59
Romance	0.76	0.67	0.65	0.60
Balto-Slavic	0.78	0.69	0.67	0.62

Table 17: Classification accuracy (%) of English O vs. English T on different levels of abstractions

Bonet, and Genabith, 2021) to consider more geometrical information on the interaction between all vectors within the initial space. We do this by connecting each point to its K nearest neighbours.

6.4 Evaluation and Analysis

6.4.1 Classification Analysis

First, we investigate the impact of interference on different levels of linguistic abstractions, such as lexical, morphological and semantic, and determine whether originals and translations can be easily distinguished based on these views.

To achieve this, we select 1000 test sentences from both the English original (L_e) and each of the translated languages (L_j) as test data, and down-sample the training data to the lowest overall number of instances in order to ensure balance. We then apply logistic regression as the classification algorithm, using the default settings of the scikit-learn toolkit (Pedregosa et al., 2011). We perform 10-fold cross-validation to evaluate the results, reporting accuracy as the percentage of sentences correctly classified as either English originals or translations. Since the classification task is binary and the training corpus is balanced, the baseline accuracy for this experiment is 50%.

We further extend this experiment to include 16 languages, representing 3 language families in our study. Table 17 presents the classification accuracy of English originals versus translations using each level of abstraction (view) across each language family. Our results indicate that our selected delexicalised features can capture source language interference a.k.a. shining-through across texts, even with our simple logistic regression based classifier. Notably, the highest accuracy is achieved with the lexicalised feature (Raw), indicating a possible influence of domain bias.

6.4.2 Hierarchical Clustering

In this section, we investigate whether, and if so to what extent, source interference (Toury, 2012) or shining-through (Teich, 2003) is traceable in lexical as well as delex-

icalised embedding spaces. We perform a cluster-based analysis using each of the isomorphism metrics described in Section 7.3.2. To evaluate our isomorphism methods on multi-view data representations, we cluster the language representations for each view of the data using the isomorphism measures in Figures 15-26, resulting in 3 x 4 (number of isomorphism metrics x number of views) cluster matrices.

While this is similar in concept to the evaluation presented in Section 5.5, the phylogenetic reconstruction presented here is based on a 17x17 distance matrix, whereas in Section 5.5, we calculate normalised distances between embedding spaces for original English and translations into English from 16 languages using three different distance measures.

Figures 15-26 illustrate heatmaps with dendograms representing phylogenetic trees using agglomerative clustering with variance minimisation (Ward Jr, 1963). For each method and every view of the data, we include the 17*17 matrix. This matrix accounts for distances between original English (L_e) and translations into English from 16 non-English languages (L_j), as well as the distances between each pair of these translations (L_j to L_k, where $j \neq k$).

Each heatmap tile represents the deviations in isomorphism values between the corresponding source and target languages. In this context, the source languages encompass both original English (L_e) and translations into English from 16 non-English languages (L_j). The target languages cover both original English and translations into English from the same set of languages (L_k, where $j \neq k$). The color scale indicates correlation, a value closer to 1 implies a higher correlation while a negative correlation is indicated using the contrasting colour in each graph. Additionally, we plot dendrograms to cluster the language-language relationships, i.e. they represent possible phylogenetic trees.

This evaluation is similar in approach to the evaluation method outlined in Section.5.5.1 but with more views of the data and here using all possible combinations of L_e and L_j's. From the heatmaps, we see that language family groupings can be inferred from both lexicalised and delexicalised embedding data. The language clusters obtained from the lexical (*Raw*) embedding spaces are the most distinctive ones for all divergence from isomorphism metrics. This reconfirms our results with lexicalised representations in Chapter 5. Furthermore, we find that the clusters generated using *Synset*, *ST* and *PoS* representations also exhibit coarse-grained language family traces. This provides evidence that translationese is reflected in semantic spaces generated from different representations. The differences are traced without reliance on fine-grained linguistic knowledge (annotations) and the comparison shows that the lexicalised results are not just due to possible lexical differences in domain between texts.

Next, we study what the dendrograms imply in the Figures 15-26. In all the predicted phylogeny trees, regardless of the representations, we observe trends that indicate

groupings based on important language-language relationships. Specifically, we find that translationese dialects generated by typologically similar languages tend to demonstrate more similarities than those derived from typologically distant source languages. Or in other words, cross-linguistic transfer in translation is guided by the typological properties of the source language. Germanic languages, for example, are almost perfectly clustered together, with only a minor association with the Baltic and Slavic language families.

This demonstrates clear intra-family linguistic ties, lending support to the fact that translations from related sources tend to resemble each other more than translations from more distant languages. Finally, we show that our constructed phylogentic trees show language family ties that resemble linguistically motivated phylogenetic trees. Unlike supervised lexicostatistic approaches relying on aligned multilingual cognate lists, our isomorphism analysis is unsupervised and still able to trace important language differences related to linguistic typology. In a sense, and compared to some previous approaches, departures from isomorphism in embedding spaces lets "the data speak more for itself".

Among the three metrics, we find that EV fails to reasonably group individual languages into three language-family branches as compared to the trees generated by GH and SGM. While in general we identify some well-known language–language relationships in all twelve trees such as clustering of the Germanic and Romance languages closer to each other than the Balto-Slavic, there are many divergences visible. For example, in most of cases, Romanian is far from the other Romance languages, but close to Slavic languages (Bulgarian, Czech, Slovak, Slovenian and Polish). Such a divergence could be from the influence of geographical factors such as language contact or structural interactions (*Balkan Sprachbund*, BS) as the geography of Romanian opens it to cross-pollination with the other languages of the BS area and the figures provide evidence for that. Portuguese in many instances is located far away from its other Romance counterparts under EV and GH reconstructions. Polish is always misplaced into the Germanic language group, despite its Slavic origin. As future work we intend to include a thorough quality assessment by native speakers on the data before performing clustering divergence from isomorphism based analyses. Another possible reason for such divergences could be adherence to the target language (i.e., English in this case) norms or *explicitation*. This is something that would need to be further explored in future work.

Finally, our analysis demonstrates that all embedding views exhibit source language interference a.k.a. shining-through. In turn, this allows us to conclude that the lexicalised results are not just due to possible domain differences between original and translated texts, thus addressing **RQ4**.

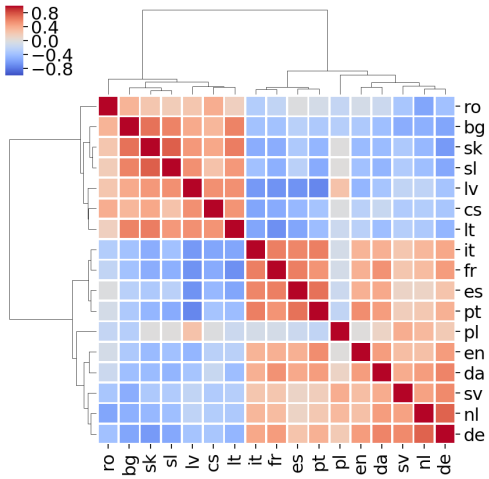

Figure 15: GH-RAW

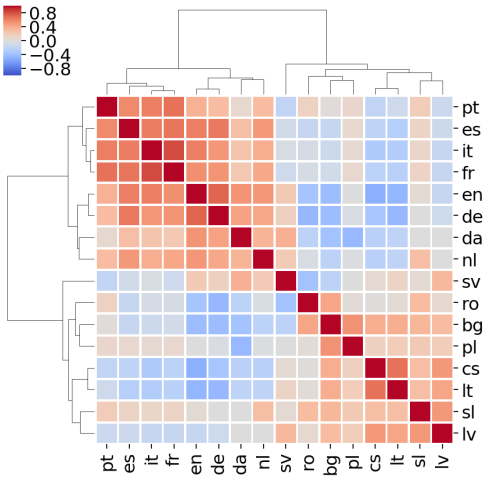

Figure 16: GH-PoS

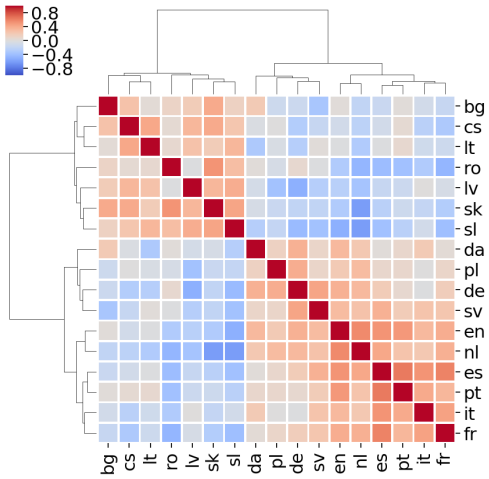

Figure 17: GH-SemTag

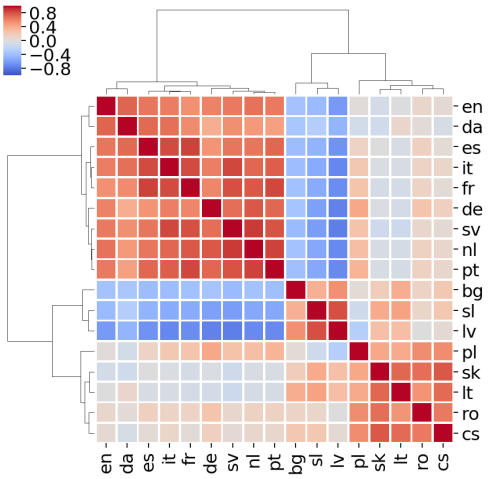

Figure 18: GH-Synset

6.4.3 Typological Characterisation

We demonstrated in Section 6.4.2 that source-language traces (*shining-through*) exist in semantic spaces that can establish genetic relations between languages without the need for any etymological information. Building on this, our next step is to analyse how the source languages are reflected in the translations, as represented by the PoS and Semantic tagsets described in Section 7.3.1. Specifically, we aim to explore whether differences in the frequency of these features in translations into English from different source languages correspond to the usage of the same features in the original source languages.

To do this, we calculate the frequency of these structural characteristics (as opposed to embeddings) in texts translated into English from each particular source language and average the measures across each language family group (Germanic, Romance, and Balto-Slavic). The results are shown in Tables 18 (PoS) and 19 (ST).

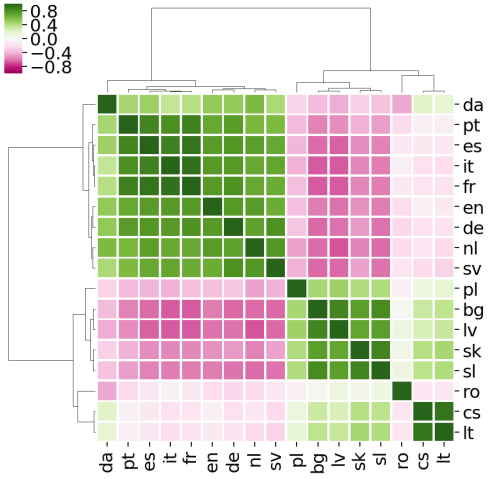

Figure 19: EV-RAW

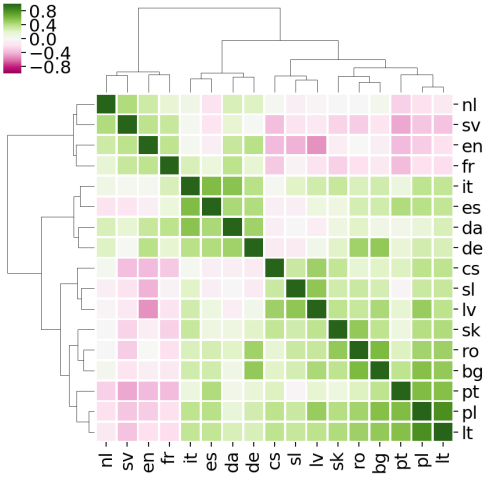

Figure 20: EV-PoS

Figure 21: EV-SemTag

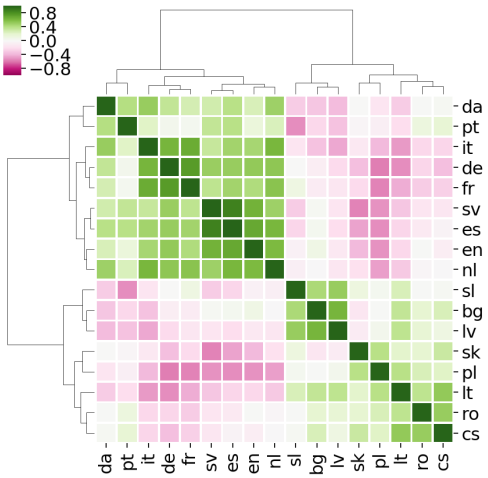

Figure 22: EV-Synset

Analysis. All Romance, nearly all Germanic, and some Balkan languages have both indefinite and definite articles. Balto-Slavic languages traditionally lack articles (Rabinovich, Ordan, and Wintner, 2017). The acquisition of article usage in English is notoriously challenging, resulting in errors among non-native speakers (Han, Chodorow, and Leacock, 2006). Notably, individuals with Slavic language backgrounds, as native speakers, often exhibit an inclination to overuse definite articles when expressing themselves in Germanic languages (Hirschmann et al., 2013). Similarly to this, we expect an excessive usage of the indefinite articles in translations into English from Balto-Slavic languages. The findings reveal a notable overuse of "DET" as one single class in Table 18 and fine-grained "DEF" in Table 19 in translations from Balto-Slavic languages into English as well as a minor overuse in translations from Romance languages.

A contributing factor to the higher frequency of adjectives (ADJ) (Table 18) in translations from Romance languages into English may be that particles are frequently

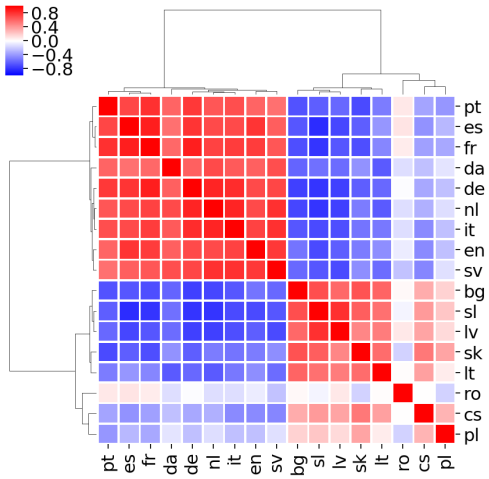

Figure 23: SGM-RAW

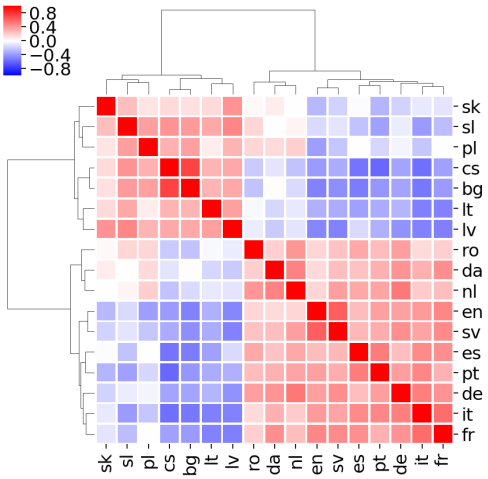

Figure 24: SGM-PoS

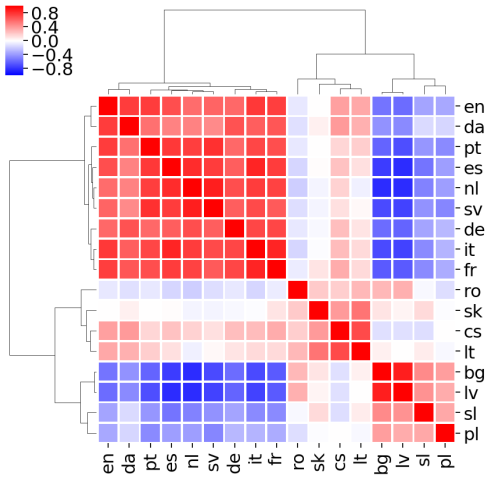

Figure 25: SGM-SemTag

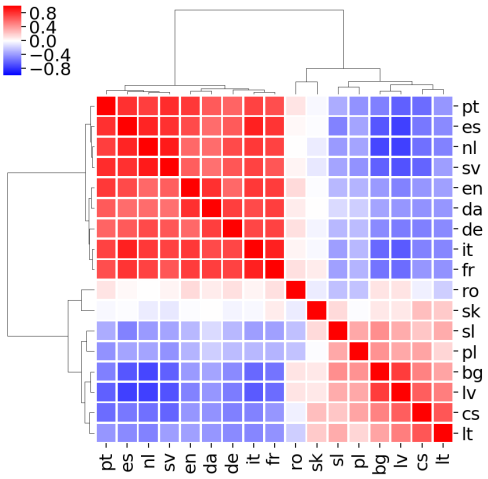

Figure 26: SGM-Synset

expressed through adjectives or verbs, unlike in Romance languages where they are separate. In contrast, Balto-Slavic languages tend to use nouns or participles to modify the meaning of other nouns, relying less on adjectives. This linguistic distinction may contribute to the lower frequency of adjectives in Balto-Slavic languages, a characteristic also reflected in translations into English from this language family.

Nouns (NOUN) have a higher frequency in the translations into English from Balto-Slavic language group (20.82) than in the Germanic (18.75) and Romance (20.22) language groups. This could be influenced by a range of factors such as cultural, historical, and linguistic factors. For example, some linguists (Beck, 2013; Nichols, 1992) have suggested that the Slavic branch of the Balto-Slavic language family may have been more influenced by the structure of ancient Indo-European languages, which tended to have more complex noun declension systems. Additionally, the morphology of Balto-Slavic languages may also contribute to the frequency of nouns, as the system of noun declension in these languages is complex, with different cases and declension

Tag	Description	Germanic	Romance	Balto-Slavic
ADJ	adjective	7.21	7.69	1
NOUN	noun	18.75	20.22	20.82
VERB	verb	11.25	11.05	10.82
ADV	adverb	5.02	3.91	4.07
ADP	adposition	10.80	11.55	11.67
AUX	auxiliary	5.82	4.59	4.43
CCONJ	coordinating conjunction	3.41	3.59	3.89
DET	determiner	12.43	12.69	13.06
INTJ	interjection	0.027	0.020	0.013
NUM	numeral	0.83	0.97	0.87
PART	particle	2.84	2.60	2.50
PRON	pronoun	4.87	3.83	3.57
PROPN	proper noun	4.50	4.62	5.17
PUNCT	punctuation	9.69	9.54	9.21
SCONJ	subordinating conjunction	2.61	2.21	2.10
PUNCT	punctuation	9.79	9.67	9.21

Table 18: Average frequency distributions of Parts-Of-Speech tags for translations into English across language families. Cell colors indicate higher frequencies.

patterns, which allows for the formation of compound words that combine several noun stems. On the other hand, Germanic and Romance languages have undergone significant simplification of noun declension systems over time, resulting in a relatively smaller number of distinct noun forms and fewer opportunities for the formation of complex compounds.

Table 18 shows that the frequencies of "AUX" and "VERB" have are the highest for translations into English from Germanic languages, followed by Romance and Balto-Slavic languages respectively. A similar trend is observed with "NOW" in Table 19. The order of Germanic and Romance language "EPS" (past simple) in Table 19 is swapped. This could be because verbs play a more central role in the syntax of the Germanic languages, where the main verb usually comes second in the sentence and auxiliary verbs are used to show tense, mood, and aspect. Languages vary greatly in their use of tense and aspects. Germanic languages generally use auxiliary verbs to form compound tenses and express aspect and modality, rather than relying on inflectional endings as the Romance languages do. For example, in German, the auxiliary verb "haben" is used to form the present perfect tense (e.g. "Ich habe gegessen") and the auxiliary verb "werden" is used to express future tense (e.g. "Ich werde essen"). In contrast, in Romance languages, verb endings are often used to indicate the tense and mood of the verb, rather than through auxiliary verbs. For example, in French, the verb "parler" (to speak) can be conjugated as "je parle" (present tense), "j'ai parlé" (past tense), or "je parle rais" (conditional), where the different endings indicate the tense and mood. In Balto-Slavic languages, such as Lithuanian and Polish, the distinction between perfective and imperfective aspect is marked by using different lexical verbs rather than by using auxiliary verbs as in Germanic languages. For

Tag	Description	Germanic	Romance	Balto-Slavic
PRO	pronoun	4.31	3.50	3.11
DEF	definite	6.53	7.26	7.50
HAS	possessive	1.02	0.98	1.04
REF	reflexive	0.16	0.10	0.10
EMP	emphasizing	0.09	0.06	0.05
GRE	greeting	0.003	0.004	0.007
ITJ	interjection	0.005	0.004	0.003
QUE	interrogative	0.52	0.37	0.31
UOM	measurement	0.30	0.33	0.33
IST	intersective	7.41	7.90	8.29
REL	relation	14.21	15.36	14.93
INT	intensifier	0.59	0.41	0.43
ALT	alternative	0.91	0.71	0.77
IMP	implication	0.15	0.09	0.09
AND	conjunct./univ.	1.98	2.08	1.91
BUT	contrast	0.50	0.42	0.36
LES	comparative neg.	0.038	0.035	0.04
TOP	pos. superlative	0.19	0.21	0.23
BOT	neg. superlative	0.001	0.001	0.001
ORD	ordinal	0.15	0.15	0.16
PRX	proximal	1.29	1.27	1.16

Tag	Description	Germanic	Romance	Balto-Slavic
SUB	subordinate	1.42	1.14	1.04
COO	coordinate	0.85	0.92	0.93
NOT	negation	1.11	0.86	0.83
NEC	necessity	0.82	0.68	0.67
POS	possibility	0.48	0.35	0.33
CON	concept	20.56	21.92	22.96
ROL	role	1.79	1.87	1.97
GPE	geo-political ent.	0.005	0.004	0.003
MOR	comparative pos.	0.41	0.36	0.39
PER	person	0.02	0.02	0.02
ORG	organisation	0.004	0.003	0.004
ART	artifact	0.001	0.001	0.001
HAP	happening	0.0007	0.0004	0.0003
EXS	untensed simple	4.82	4.50	4.43
ENS	present simple	3.49	3.25	3.16
EPS	past simple	1.40	1.50	1.38
EFS	future simple	0.028	0.025	0.021
EXG	untensed prog.	1.43	1.78	1.67
NOW	present tense	3.00	2.41	2.27
PST	past tense	0.46	0.38	0.40
FUT	future tense	0.92	0.85	0.98

Table 19: Averaged frequency distributions for translations into English across language families using Semantic Tags from Bjerva, Plank, and Bos (2016). Cell colors highlight higher frequencies.

example, in Lithuanian, the perfective aspect is marked by using the prefix "su-" with the verb, while the imperfective aspect is marked by using a different verb stem altogether. This is different from Germanic languages, where the distinction is often marked using auxiliary verbs such as "have" or "be", exhibiting the highest frequency in translations from Germanic languages.

The majority of Slavic languages follow the original Proto-Slavic pattern of seven case forms (nominative, genitive, dative, accusative, locative, instrumental, vocative), which exist in both the single and plural. The genitive case, in particular, is head-marked, which alters the order of the two nouns in comparison to the standard English clitic "s" construction (Eetemadi and Toutanova, 2014). Thus, we anticipate translations of Balto-Slavic languages to overuse phrases containing 'of' — NP constructions. Table 19 shows the overuse of possessive constructions "IST" and "REL" in translation into English from Balto-Slavic languages, as also observed within the single class "ADP" in Table 18.

Finally, we also identify several spurious correlations in Tables 18 and 19 that cannot be described using the interference features. It is worth noting that the illustrating linguistic examples used here (collected from various sources) are not meant to be exhaustive or representative of all the ways that these PoS and Semantic tags can be used in each language group. They are simply intended to provide a general idea of the ways that these tags are used in each language family.

Furthermore, it is worth acknowledging that since we took average counts across language families for our analysis, this could potentially result in the loss of individual characteristics of each source language. Each language is unique in its own right, and

averaging them together could have obscured important nuances and variations.
It is also worth noting that this analysis relies on rather high-level linguistic features, either at a fine-grained (PoS) or coarse-grained (SemTag) level. Further exploration using additional features based on deeper linguistic analyses should provide a more comprehensive understanding of these inconsistencies.

We collected the illustrating examples used in our analysis from various online sources and subsequently validated or refined them using commonly available language tools[1]. Furthermore, these examples were verified with the help of four language experts to ensure accuracy and consistency with established linguistic principles.

6.5 Conclusion

This chapter focuses on exploring whether and if so to what extent the results obtained in Chapter 5 are due to possible differences in domains between original and translated data as reflected in terms of lexical differences rather than on proper translationese signals. To achieve this, we mask domain information by applying our graph-isomorphism methods to different delexicalized representations based on word level, morphological, structural, and semantic views of the data. We demonstrate that translationese can be traced in semantic spaces even when lexical information is removed, and this indicates that the influence of domain is not critical for the success of unsupervised tracing of translationese, answering **RQ4**.

We find that lexicalized embeddings show the most pronounced source-language interference, and compared to all levels of abstraction explored in our work. Our results suggest that translated texts retain typological characteristics from the source language, and our isomorphism-based results can infer phylogenetic language family relations for all views without requiring any externally supplied etymological information. This demonstrates that translated texts retain typological characteristics from the source language and that our method is able to capture these characteristics even at different levels of abstraction.

As future work, we intend to extend our experiments to capture the geometric properties of the embedding features and work on isolated languages.

1 https://www.dict.cc/?s=it

7

Divergence from Isomorphism and Surprisal in Translationese

In the previous chapter, we presented implicit evidence that divergence from isomorphism between embedding spaces indicates structural surface differences between language families that are the source languages of the translations. Higher divergence from isomorphism between embedding spaces suggests greater linguistic distance in terms of language families and, consequently, supposedly reflects surface structural linguistic differences (e.g., morphology, syntax) between languages, as indicated in the linguistic literature. In this chapter, we explicitly explore the relationship between embedding spaces and surface string representation (where the latter is measured by LM entropy) by computing the correlation between (a) differences in surface string entropy of original vs. translated data computed by language models trained on originally authored data and (b) divergence from isomorphism between embedding spaces computed on the same text data. Our results show that the two measures (a surface measure and an embedding measure) correlate, and that higher departure from isomorphism between embedding spaces corresponds with a higher difference in surface entropy.

7.1 Introduction

The relationship between linguistic distance and surprisal has been extensively studied in the context of receptive multilingualism (Vanhove, 2014) or intercomprehension studies (Jágrová et al., 2019). Linguistic distance is commonly assessed across various descriptive levels of languages, encompassing lexical, orthographic, and morphological dimensions. These measurements are typically derived from parallel sets of words or texts. Golubović and Gooskens (2015) demonstrated that the mutual intelligibility of two languages is often predicted by their linguistic distance. A key observation in their work is that if a text has reduced linguistic distance, then the transfer of knowledge from a language to an unknown language is possible, for instance, Czech and Slovak allow for mutual intelligibility without significant issues (Nábělková, 2007). On the other hand, surprisal scores assigned by statistical models can also serve as reasonable

indicators of the cognitive effort involved in human natural language comprehension (Hale, 2001; Levy, 2008). High surprisal scores indicate greater processing difficulty and higher information content.
Unlike intercomprehension studies, the interplay between these two orthogonal dimensions—linguistic distance and surprisal—in the specific context of translation remains under-explored. Translation is a complex linguistic process that entails conveying information in a language distinct from that of the original language (Nikolaev et al., 2020). From an information-theoretic perspective, two hypotheses in the literature propose conflicting views about translationese: (1) that it is more predictable than the original language and (2) that it includes non-native linguistic patterns, making it less predictable. For example, Pastor et al. (2008) show that translated texts retain traces of their source language, rendering them less predictable and more challenging to process. In another example, research by Bjerva et al. (2019), building on the work of Rabinovich, Ordan, and Wintner (2017) has demonstrated that in translations into English, it is often possible to identify the original language based on morphosyntactic properties alone. The assumption that translated text is more predictable also seems to be supported by the empirical observation that translated texts tend to exhibit lower levels of surprisal in contrast to their original counterparts (Bizzoni et al., 2020). On the other hand, Baker et al. (1993) highlighted both a tendency to simplify the language used in translation and also a tendency to amplify features of the target language and adhere to its typical patterns, where the former would lead to lower and the latter to higher perplexity scores using a language model trained on originally authored language.

In the previous chapters, we provide some implicit evidence that divergence from isomorphism between embedding spaces indicates structural surface differences between language families that serve as source languages for the translations. Higher divergence from isomorphism between embedding spaces indicates higher linguistic distance in terms of language families and with that supposedly surface structural linguistic distance (e.g. morphology, syntax) between languages themselves as we know from the linguistic literature (Haspelmath et al., 2005). This is shown indirectly in Chapter 6 by the Part-of-Speech (PoS) (and the other different view) experiments that abstract from lexical information in the original and translationese embedding experiments where masked views (POS etc.) still show reasonable original vs. translationese differences in isomorphism between the embedding spaces and based on this, phylogentic family tree results.

This raises the question of whether there is any explicit evidence to support the connection between embedding spaces and structural surface differences, deviating from the implicit approach observed in Chapter 6. This, in turn, leads us to formulate our last research question:

RQ5: To what extent can graph-based divergence from isomorphism in embedding spaces act as a proxy for surprisal at the level of surface texts?

To address this question, we explore the relationship between embedding spaces and surface strings further by explicitly computing the correlation between (a) differences in surface string entropy of original vs. translated data computed by language models trained on originally authored data and (b) divergence from isomorphism between embedding spaces computed on the same text data. We do this in two ways: (i) general correlation between differences in LM entropy on surface strings between original and translated and divergence from isomorphism between embedding spaces for the same data; and (ii) a more fine-grained version of (i) to verify whether larger differences in entropy correspond with larger divergence in isomorphism. We find that the two (a surface measure and an embedding measure) correlate, with a higher departure from isomorphism between embedding spaces corresponding to a greater difference in surface entropy. In fact, from our results, we find that the observed differences in terms of surface string representations (as measured by LM entropy) between the original and translated datasets correspond, at a high level, with surface structural linguistic distance (e.g., morphology, syntax) between the source languages themselves, as characterized by WALS (Haspelmath et al., 2005). This is different from the indirect approach in Chapter 6 where the link between embeddings and structural surface differences was only implicit through what is generally assumed in the linguistic literature in terms of language family relations and surface structural differences. In contrast, in this chapter, we provide hard evidence for this via linking two explicit measures: divergence from isomorphism between original and translated embedding spaces and entropy differences in the surface strings of the same text data.

The rest of the chapter is organised as follows. We review related work in Section 7.2. Section 7.3 briefly outlines data settings, and methods used in the study. In Section 7.4, we describe our evaluation methods, present the results and conclude in Section 7.5.

7.2 Related Work

Several investigations have utilized information-theoretic metrics like entropy and perplexity to automatically classify and contrast translated and non-translated texts. For example, Rubino, Lapshinova-Koltunski, and Genabith (2016) employ information density encoded as surprisal at the word, part of speech, and syntax levels to help build a state-of-the-art model for detecting mixed-domain translationese. Martínez and Teich (2017) focus on disparities in lexical choices between professional and student translators, leveraging word translation entropy indicating how many equally likely translations may be produced for a source word in a given context. Higher translation

entropy implies a broader range of lexical choices for the translator, indicating higher cognitive effort on their part. Toral (2019) use PoS perplexity to quantify the degree of translationese (interference of the source language) in translated text T (human translation, post-editing, and machine translation). Given a PoS-tagged translation T, a language model of a PoS-tagged corpus in the source language, and a language model of a PoS-tagged corpus in the target language, they calculate the difference in perplexities of T with respect to both language models. They show a high difference in perplexity suggests the translation is dissimilar to the source language but similar to the target language, while a low difference in perplexity indicates the opposite. Bizzoni and Lapshinova-Koltunski (2021) uses perplexity to explore translationese in English-to-German translations by professionals and students across various registers. Despite expectations, they found professional translations show higher perplexity scores than student translations, indicating increased stylistic diversification and register sensitivity in the former. Teich, Martínez, and Karakanta (2020) proposes that entropy can effectively serve as a suitable index for various translation-related factors, including (dis)similarity between languages, level of expertise, and translation mode (i.e., interpreting vs. translation), and demonstrating the level of production effort. Nikolaev et al. (2020) explores whether the degree of difference in morphosyntactic entropies between original texts and texts translated from English correlates with the extent of morphosyntactic divergence between the languages of these texts and English. In a more recent study, Przybyl et al. (2022) examines distinctive features of translated and interpreted texts at the European Parliament, exploring the effects of mediation (translation vs. interpreting) and discourse mode (written vs. spoken) using word-based n-gram language models and relative entropy. Their study identifies linguistic features typical of translation/interpreting, confirming the influence of written vs. spoken mode on translation and interpreting.

7.3 Experimental Setup

7.3.1 Datasets

We use a comparable subset of Europarl (Koehn, 2005) used in Chapters 5 and 6, consisting of translations from 21 European Union languages into English (refer to Section 5.4). Our focus is specifically on a subset of translations from 16 languages, covering three language families: *Romance* (French, Italian, Spanish, Romanian, Portuguese), *Germanic* (Dutch, German, Swedish, Danish), and *Balto-Slavic* (Latvian, Lithuanian, Czech, Slovak, Slovenian, Polish, and Bulgarian), both into English and the original English text for this study.

7.3.2 Divergence from isomorphism as a measure for distance

Here, we provide a brief overview of the methods we used in Chapter 5 to quantify a notion of *distance* between languages based on the word embedding spaces using the notion of divergence from isomorphism. Additional technical details about these measures are covered in Section 5.3.

Gromov Hausdorff distance (GH). This distance, as detailed in Patra et al. (2019) tests how well two language embedding spaces are aligned under an isometric transformation. Specifically, it computes the worst-case distance between two metric spaces, representing the maximum distance of a set of points to the nearest point in another set.

Eigenvector Similarity measure (EV). This measure, proposed by Søgaard, Ruder, and Vulić (2018), analyzes the Laplacian eigenvalues of nearest neighborhood graphs from the initial language embedding spaces. These eigenvalues, viewed as compact representations of the graph Laplacian, are compared to assess the degree of (approximate) graph-isomorphism.

Spectral Graph Based Measure (SGM). Similar to the Eigen-Vector Similarity approach, we introduce this measure (Dutta Chowdhury, España-Bonet, and Genabith, 2021) to account for more geometric information about the interactions among all vectors within the initial space. This is acheived by linking each embedding to its K nearest neighbors.

7.3.3 Surprisal as a measure for information density

To assess differences in entropy between original and translated texts, we use Shannon's entropy (Shannon, 1948) as an estimate of surprisal. Surprisal, denoted as $-\log_2 p(x)$, is the measure of how unexpected an event x is. Here, $-\log_2 p(x)$ is the surprisal of x, measured in bits. This value quantifies the surprise or uncertainty associated with the occurrence of event x based on its probability $p(x)$. Higher values indicate more surprising events.
Formally, the entropy of a discrete random variable X is defined as the weighted average of the surprisal values of this distribution. In other words, entropy is the expectation of the surprisal. The formula is given by:

$$H(X) = -\sum_i P_X(x_i) \log P_X(x_i)$$

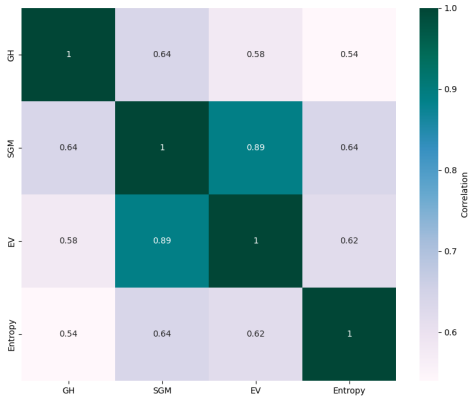

Figure 27: Correlation between Graph-Based Divergence Measures and Entropy

Here, $H(X)$ denotes the entropy, and $P(X)$ the probabilities associated with different words in the corpora.

We calculate the surprisal of each word in the corpus using a single-layer LSTM-based language model (Hochreiter and Schmidhuber, 1997). This involves feeding the word into the model and then measuring the negative log probability of the word given the model's predictions. The surprisals were then averaged to calculate the entropy of the surprisal distributions for both corpora: the original English data and the translations into English language data from 16 languages covering three language families. The resulting differences between recorded entropies ($H_{\text{translated}} - H_{\text{original}}$) is recorded for further analysis.

7.4 Evaluation and Analysis

To find if graph-based divergence from isomorphism at the level of embedding spaces can act as a proxy for surprisal at the level of surface texts, we estimate entropies to investigate whether the entropies of the originally authored target language text significantly differ from translated texts into this target language. Figure 27 shows the correlation coefficients between graph-based divergence from isomorphism (GH, SGM, and EV) and entropy differences between original and translated texts. We employ Pearson correlation analysis to perform the correlation analysis. Figure 27 shows that SGM exhibits the strongest positive correlation with surprisal, followed by EV and GH. The plot suggests that there is a moderate positive connection between the predictability of translated texts and the divergence in isomorphism between the target language and the source language.

Next, we check if bigger divergences from isomorphism correspond to bigger differences between respective entropies of the translated and original-language

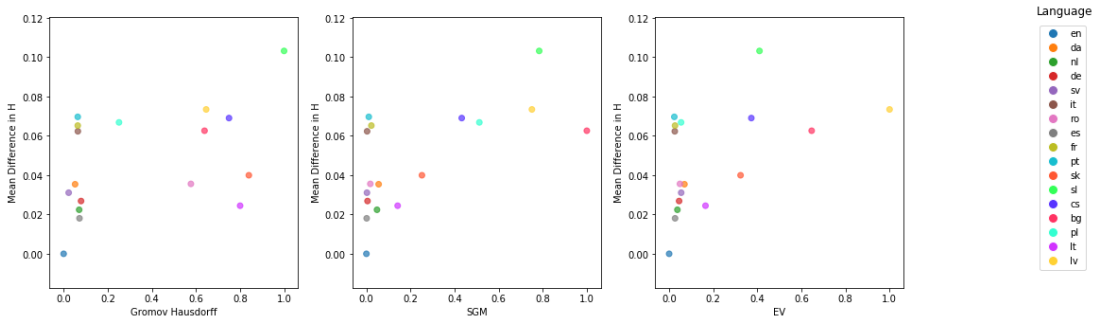

Figure 28: Pair Plot between Isomorphic Divergence and Entropic Differences (H) in Translated Texts

corpora. Specifically, we compare the mean difference in entropy between the original English texts and translations into English from 16 source languages across three families and then compare this to the divergences from isomorphism between the corresponding pairs to understand whether these differences in entropy align with the differences in isomorphism between corresponding language pairs.

The results are shown in Figure 28. Overall, the results support the hypothesis of a positive correlation between the difference in entropies of translated and original texts and the degree of divergence in isomorphism between texts: bigger divergences in isomorphic measures correspond to bigger differences between the respective entropies of the translated and original-language corpora. We see that in most cases when texts are translated from topologically divergent languages into English, higher levels of entropic differences are seen than those translated from structurally similar languages. For example, from Figure 28, we see that two source languages structurally close to English (in this case, German and Dutch) show mean difference values that are closer to zero, while more structurally divergent languages (Latvian, Bulgarian) demonstrate a noticeable boost in the entropy in translated texts. These results mirror the patterns observed in divergence from isomorphism between embedding spaces, where translations from structurally more divergent languages lead to increased divergence in isomorphism. However, some interesting divergences, with respect to Balkan Sprachbund (BS) are visible as well.

By combining the results from both analyses, we address the research question that explores to what extent graph-based divergence from isomorphism at the level of embedding spaces can act as a proxy for surprisal at the level of surface strings as measured by a language model. The first analysis explores the correlation between divergence from isomorphism measures (GH, SGM, and EV) and entropy differences, indicating the predictability of translated texts compared to the original language. The second analysis provides evidence supporting the hypothesis that larger divergence from isomorphism are associated with more significant differences in linguistic predictability, indicating the presence of non-native linguistic patterns in translationese. Specifically, we observe that, on average, translations from languages that to a first

approximation are more structurally divergent from English tend to exhibit higher levels of entropy difference. On the other hand, translations from languages structurally similar to English show smaller entropy differences. At a higher level, this aligns with the surface structural linguistic distance (e.g., morphology, syntax) between the languages themselves, as characterized by WALS (Haspelmath et al., 2005). However, the precise nature of this correlation remains a topic for future investigation.

7.5 Conclusion

This chapter explores whether the graph-based divergence from isomorphism at the level of embedding spaces, as discussed in Chapters 5 and 6, can serve as a viable proxy for surprisal at the level of surface texts as measured by a language model. Our approach involves employing entropy estimation and correlating it with three graph-based divergence from isomorphism measures (GH, SGM, and EV) to shed light on the intricate relationship between surface string predictability in translated texts and divergence from isomorphism between embedding spaces. We perform this in two ways: (i) general correlation between differences in LM entropy on surface strings between original and translated and divergence from isomorphism between embedding spaces for the same data; and (ii) a more fine-grained version of (i) to verify whether larger differences in entropy correspond with larger divergence in isomorphism.

Our results exhibit a positive correlation between divergence from isomorphism measures (GH, SGM, EV) and entropy differences, indicating that a higher departure from isomorphism between embedding spaces corresponds with higher differences in surface entropy. This explicitly links embedding space with surface string representation (where the latter is measured by LM entropy). Additionally, our findings show that on average, translations from languages with greater structural divergence from English tend to exhibit higher levels of entropy difference. In contrast, translations from languages structurally similar to English show smaller entropy differences. These outcomes align with the observed trends in divergence from isomorphism results from Chapters 5 and 6, where embedding spaces of translations into English from structurally similar source languages are more isomorphic nature to embedding spaces computed from original English. Conversely, translations from structurally more divergent languages result in increased divergence from isomorphism. At a broader level, this aligns with the surface structural linguistic distance (e.g., morphology, syntax) between the source languages themselves, as characterized by WALS (Haspelmath et al., 2005). However, the precise nature of this correlation remains a subject for future exploration. Collectively, these findings provide hard evidence that our two explicit measures: divergence from isomorphism between original and translated embedding

spaces and entropy differences of the surface strings of the same text data are linked, addressing the overarching question in **RQ5**.

Finally, it is an open research question whether just entropy could be used as effectively for translationese classification and phylogenetic tree reconstruction tasks. Entropy mainly focuses on statistical language properties and might not be a strong predictor by itself. Future directions could explore combining entropy with linguistic features or neural methods to achieve a more comprehensive understanding that encompasses both statistical language aspects and the intricate semantic and syntactic relationships between languages.

8

Conclusion & Future Work

8.1 Summary of Dissertation

Translationese refers to systematic linguistic differences between translated texts and texts originally authored in the same language. These differences can be categorised as either source language dependent or universal. Basic research on translationese aims to capture and comprehend the language-specific and language-independent aspects of translationese. Additionally, translationese has practical implications for natural language processing tasks that involve translation, as it can lead to biased results and artificially inflated or decreased performance in cross-lingual tasks. Therefore, understanding, analysing, and mitigating translationese is critical to improving the accuracy and effectiveness of cross-lingual natural language processing. This dissertation focuses on representation learning and addresses both foundational and practical aspects of translationese. In this final chapter, we summarize our main findings, identify the limitations of our work, and suggest future directions.

In Chapter 3, we investigate the effectiveness of representation learning-based methods for mono- and multilingual translationese classification. Initially, a variety of stylistic features based on text statistics, richness, and readability measures are examined using a broadly corpus linguistics based approach. The results show that translated texts exhibit different characteristics compared to texts originally authored in the same language. In fact, we found that stylistic feature based results can differ substantially for specific language pairs, and even for the same target language with translations from different source languages into this target language. This highlights the need for a language-agnostic approach to translationese classification.

We then focus on translationese classification in both mono- and multilingual settings using representation-learning-based models. Previous studies on translationese classification had almost exclusively used hand-crafted, linguistically inspired features and feature-engineering approaches. In contrast, in our work we use word representations derived from various word embedding architectures. Our findings show that word representation learning-based approaches, particularly those using ***fasttext*** representations, are a superior alternative to traditional hand-crafted, linguistically

inspired feature-selection methods for translationese classification across a wide range of tasks.

Furthermore, we compare these approaches to fine-tuned `BERT` and `LSTM` architectures, and we find that ***fasttext*** representations perform fairly close to these advanced models, with `BERT` producing the best results. Most of the research presented in Chapter 3 is published in Pylypenko et al. (2021).

In Chapter 4, we shift our focus from using representation learning-based methods to classify original and translated texts to debiasing translationese in those representations. Translationese artifacts are found to have a significant impact on various downstream NLP tasks. Therefore, we propose a new approach to mitigate this impact: translationese debiasing.

Specifically, we extend the INLP algorithm, originally designed to mitigate gender attributes, to debias translationese. At the sentence level, we use several neural architectures and show that classification scores reduce to that of a random classifier after debiasing. For word representations, we propose two methods. First, we extract translationese word lists based on their usage. A translationese word list is a list of words that are used maximally different in original and translated data. To find these words in the data, we consider the size of the intersection of two sets where a word is represented as its top-k nearest neighbors (NN) once in original data and once in the translated data embedding spaces. Words with the smallest intersection are considered indicators of translationese as they differ most between original and translated data. Second, we use explicit tagging to distinguish the same word used in original and translated data when mapping them onto a shared embedding space. We then employ INLP to linearly remove translationese features from its representation. As expected, we observe that the INLP-based linear translationese debiasing results on static word embeddings are as perfect as our sentence-level results. We evaluate the effectiveness of our approach by comparing the classification performance before and after debiasing on the translationese classification task.

Finally, we conduct an extrinsic evaluation of this debiasing method in the context of the natural language inference (NLI) task, where we integrate the proposed translationese debiasing method. Our results show that the debiasing approach leads to improved NLI accuracy due to reduced translation-induced bias. Most of the research presented in Chapter 4 is published in Chowdhury et al. (2022).

In the second part of this thesis, we focus on understanding and characterising translationese from a foundational perspective. We explore whether it is possible to observe signals of translationese in embedding spaces and investigate the practical implications of such observations. In Chapter 5, we introduce a novel technique based on departures from isomorphism, which enables us to track translationese in word embedding based semantic spaces. We model the distance between languages based

on the departures from isomorphism between embedding spaces constructed from original target language data and translations into that target language.

We experiment with state-of-the-art metrics to quantify divergence from isomorphism in embedding spaces and propose an alternative graph-based distance metric called Spectral Graph-based Matching. By comparing the distance between embedding spaces, we identify systematic evidence of translationese. Our findings suggest that as isomorphism weakens, the linguistic distance between etymologically distant language families increases, providing evidence that translationese signals are linked to source language interference.

Furthermore, we demonstrate that our unsupervised data-driven graph and embedding-based distance algorithms are as effective as earlier methods that relied on linguistic features (lexical and/or syntactic) to identify source language interference in translations. Additionally, we show that our proposed methods are robust under a variety of training conditions, encompassing data size, type, and choice of word embedding models. Finally, our findings also indicate that these methods are language-independent and can be applied to multiple languages, not limited to a specific language or language family. Most of the research presented in Chapter 5 is published in Dutta Chowdhury, España-Bonet, and Genabith (2021).

In Chapter 6, we extended our work on unsupervised tracking of translationese in word embedding based semantic spaces to evaluate the impact of domain on distinguishing between original and translated data. To achieve this, we use various perspectives (views) of the same data including words, parts of speech, semantic tags, and synsets to detect translationese in semantic spaces by measuring departures from isomorphism. Our findings indicate that while lexicalised embeddings demonstrate a significant presence of source-language interference, other delexicalised levels of abstraction also display similar tendencies, implying that lexicalised results are not solely due to possible domain differences between original and translated texts.
Additionally, we observe that language family ties with characteristics similar to linguistically motivated phylogenetic trees can be inferred from divergence from isomorphism without relying on external etymological information. In particular, a greater departure from isomorphism in semantic spaces typically corresponds to a larger linguistic distance among language families. Lastly, we demonstrate that distinctive typological characteristics of the source languages of translations are retained in the translation process, which helps in clustering the language families. Most of the research presented in Chapter 6 is published in Dutta Chowdhury, España-Bonet, and Genabith (2020).

Finally, in Chapter 7, we explore the relationship between embedding spaces and surface strings further by explicitly computing the correlation between (a) differences in surface string entropy of original vs. translated data computed by language models

trained on originally authored data and (b) divergence from isomorphism between embedding spaces computed on the same text data. We do this in two ways: (i) general correlation between differences in LM entropy on surface strings between original and translated and divergence from isomorphism between embedding spaces for the same data; and (ii) a more fine-grained version of (i) to verify whether larger differences in entropy correspond with larger divergence in isomorphism. In (i) we find that the two (a surface measure and an embedding measure) correlate, and in (ii) we find that a higher departure from isomorphism between embedding spaces corresponds with a higher difference in surface entropy. This explicitly links embedding space with surface string representation (where the latter is measured by LM entropy). This is different from the indirect approach in Chapter 6 where the link between embeddings and structural surface differences was only implicit through what is generally assumed in the linguistic literature in terms of language family relations and surface structural differences. In contrast, in Chapter 7, we provide hard evidence for this via linking two explicit measures: divergence from isomorphism between original and translated embedding spaces and entropy differences of the surface strings of the same text data.

8.2 Future Directions

We would like to highlight some limitations of our work, and suggest future research directions in this section. Chapter 3 presents various approaches for learning word representations for translationese classification. While we find that pre-trained *fastText* models perform close to other advanced neural models (such as BERT), our experiments show that using Gaussian distributions over a latent embedding space (Nikolentzos et al., 2017) only leads to fairly modest results. However, we are yet to explore an alternative approach that involves embedding objects in a hyperbolic space (Nickel and Kiela, 2017), which can more naturally represent hierarchical relationships (Krioukov et al., 2010). Therefore, we believe that investigating embedding spaces with different topologies (Li et al., 2018; Vilnis et al., 2018) holds great potential as an exciting future direction for multilingual translationese classification.

Expanding the scope and data to include non-European languages is a potential avenue for further research in multilingual translationese classification. To achieve this, collecting and curating multilingual data for translationese would be a valuable step, as evaluating the extent to which the classification model is influenced by new language features and representations is not a trivial matter. Unfortunately, due to the lack of comparable translationese corpora in multiple languages, this was not possible during the thesis. Another interesting area of research would be to explore translationese classification in low-resource settings. This could e.g. involve training multilingual language models purely on low-resource languages, without relying on any high-resource transfer, in order to assess their competitiveness.

In Chapter 4, we initiated a research line on mitigating translationese bias through attenuation in the latent representation of (word and sentence) embeddings. However, we only scratched the surface, and the effect of this on the actual output generated (as opposed to the latent spaces) was not explored. Translationese signals are multi-faceted and complex, encompassing a mix of morphological, lexical, syntactic, and semantic phenomena. Linear interventions alone may not be sufficient for a complex task such as debiasing translationese. In future work, we plan to explore non-linear guarding mechanisms (Dev and Phillips, 2019) for such tasks. In the meantime, we have taken the first step in this direction published in Jalota et al. (2023), but it is not included as part of this thesis.

We also acknowledge the various other challenges involved in translationese detection and classification, such as explicitly quantifying the direction of translationese to determine or create translationese subspaces. Additionally, there may be intersectional subspaces that need to be addressed specifically. For instance, a text that contains both original language and translated signals could be classified as belonging to an intersectional subspace, rather than purely original language or translationese. Identifying and understanding these intersectional subspaces is crucial for research on translationese, as they may have distinct characteristics or impacts on natural language processing tasks.

As a direct NLP application, we intend to expand our research to comprehend the extent to which debiasing translation artifacts can improve other downstream tasks, such as slot-filling or named entity recognition. This could be especially valuable as translation artifacts may introduce additional complexity and uncertainty to the text, making it more difficult for models to accurately identify and classify named entities or fill in predefined slots.

Chapter 5 and Chapter 6 both focus on exploring translationese in semantic spaces using self-and-unsupervised methods. However, there is still potential for further research in this area, including investigating alternative fully unsupervised approaches for detecting translationese, such as optimal transport or density matching methods (Alvarez-Melis and Jaakkola, 2018; Zhou et al., 2019).

One promising direction for future work is to use the estimated similarities and dissimilarities between languages and original and translated texts from our research to shed light on transfer behavior in other downstream tasks between specific language pairs. For instance, our results could help identify which languages are most likely to influence each other in the context of machine translation. It may also involve analysing factors such as the linguistic distance between languages, the cultural similarities and differences between language communities, or the cognitive processes involved in language learning and use.

In Chapter 7, we show that translations into English from source languages structurally more divergent from English tend to exhibit higher entropy differences, while

those from structurally similar languages show smaller differences. At a broader level, this corresponds to the surface structural linguistic distance (e.g., morphology, syntax) between the source languages themselves, as characterized by WALS (Haspelmath et al., 2005). However, the precise nature of this correlation remains a subject for future investigation.

We are also interested in exploring how different forms of language transfer can be distinguished from an information-theoretic perspective. By examining the efficiency of the language transfer process, we can determine how much information is lost or gained during the transfer and how this affects the overall quality of the transferred information. Information-theoretic metrics, such as mutual information and conditional entropy, can be used to measure the efficiency of such language transfer process.

Bibliography

[1] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. "The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 242–247. URL: https://www.aclweb.org/anthology/E17-2039.

[2] Lasha Abzianidze and Johan Bos. "Towards universal semantic tagging." In: *arXiv preprint arXiv:1709.10381* (2017).

[3] Oshin Agarwal, Funda Durupınar, Norman I Badler, and Ani Nenkova. "Word embeddings (also) encode human personality stereotypes." In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*. 2019, pp. 205–211.

[4] Željko Agić. "Cross-lingual parser selection for low-resource languages." In: *Proceedings of the NoDaLiDa 2017 workshop on universal dependencies (UDW 2017)*. 2017, pp. 1–10.

[5] David Alvarez-Melis and Tommi S Jaakkola. "Gromov-Wasserstein alignment of word embedding spaces." In: *arXiv preprint arXiv:1809.00013* (2018).

[6] Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. "Do not Rely on Relay Translations: Multilingual Parallel Direct Europarl." In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*. online: Association for Computational Linguistics, May 2021, pp. 1–7. URL: https://aclanthology.org/2021.motra-1.1.

[7] Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. "Explaining Translationese: why are Neural Classifiers Better and what do they Learn?" In: *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 281–296. DOI: 10.18653/v1/2022.blackboxnlp-1.23. URL: https://aclanthology.org/2022.blackboxnlp-1.23.

[8] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. "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)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 789–798. DOI: 10.18653/v1/P18-1073. URL: https://aclanthology.org/P18-1073.

[9] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. "Translation Artifacts in Cross-lingual Transfer Learning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. URL: https://aclanthology.org/2020.emnlp-main.618.

[10] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. "On the cross-lingual transferability of monolingual representations." In: *arXiv preprint arXiv:1910.11856* (2019).

[11] Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. "Identifying translationese at the word and sub-word level." In: *Digital Scholarship in the Humanities* 31.1 (2016), pp. 30–54.

[12] James Baker. "The DRAGON system–An overview." In: *IEEE Transactions on Acoustics, speech, and signal Processing* 23.1 (1975), pp. 24–29.

[13] Mona Baker et al. "Corpus linguistics and translation studies: Implications and applications." In: *Text and technology: In honour of John Sinclair* 233 (1993), p. 250.

[14] Richard Bamberger and Erich Vanacek. *Lesen-Verstehen-Lernen-Schreiben*. Diesterweg, 1984.

[15] Antonio Valerio Miceli Barone. "Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders." In: *arXiv preprint arXiv:1608.02996* (2016).

[16] Marco Baroni and Silvia Bernardini. "A new approach to the study of translationese: Machine-learning the difference between original and translated text." In: *Literary and Linguistic Computing* 21.3 (2005), pp. 259–274.

[17] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 238–247.

[18] David Beck. *The typology of parts of speech systems: The markedness of adjectives*. Routledge, 2013.

[19] Lisa Beinborn and Rochelle Choenni. "Semantic drift in multilingual representations." In: *Computational Linguistics* 46.3 (2020), pp. 571–603.

[20] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623.

[21] Raffaella Bernardi, Gemma Boleda, Raquel Fernández, and Denis Paperno. "Distributional semantics in use." In: *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*. 2015, pp. 95–101.

[22] Yuri Bizzoni, Tom S Juzek, Cristina Espana-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. "How human is machine translationese? comparing human and machine translations of text and speech." In: *Proceedings of the 17th International Conference on Spoken Language Translation*. 2020, pp. 280–290.

[23] Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. "Measuring Translationese across Levels of Expertise: Are Professionals more Surprising than Students?" In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. 2021, pp. 53–63.

[24] Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. "What do language representations really represent?" In: *Computational Linguistics* 45.2 (2019), pp. 381–389.

[25] Johannes Bjerva, Barbara Plank, and Johan Bos. "Semantic Tagging with Deep Residual Networks." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 3531–3541. URL: https://www.aclweb.org/anthology/C16-1333.

[26] Nikolay Bogoychev and Rico Sennrich. "Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation." In: *CoRR* abs/1911.03362 (2019). arXiv: 1911.03362. URL: http://arxiv.org/abs/1911.03362.

[27] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information." In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[28] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Advances in Neural Information Processing*

Systems 29 (2016). Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, pp. 4349–4357. URL: https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[29] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

[30] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[31] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." In: *Science* 356.6334 (2017), pp. 183–186.

[32] Lawrence Cayton. "Algorithms for manifold learning." In: *Univ. of California at San Diego Tech. Rep* 12.1-17 (2005), p. 1.

[33] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. "Semantic textual similarity-multilingual and cross-lingual focused evaluation." In: *Proceedings of the 2017 SEMVAL International Workshop on Semantic Evaluation (2017). https://doi. org/10.18653/v1/s17-2001*. 2017.

[34] Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. "Gromov-Hausdorff stable signatures for shapes using persistence." In: *Computer Graphics Forum*. Vol. 28. 5. Wiley Online Library. 2009, pp. 1393–1403.

[35] Andrew Chesterman. "Beyond the particular." In: *Translation universals: Do they exist* (2004).

[36] Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. "Towards Debiasing Translation Artifacts." In: *arXiv preprint arXiv:2205.08001* (2022).

[37] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages." In: *Transactions of the Association for Computational Linguistics* 8 (July 2020), pp. 454–470. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00317. eprint: https://direct.mit.

edu/tacl/article-pdf/doi/10.1162/tacl_a_00317/1923348/tacl_a_00317.pdf. URL: https://doi.org/10.1162/tacl_a_00317.

[38] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised cross-lingual representation learning at scale." In: *arXiv preprint arXiv:1911.02116* (2019).

[39] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word translation without parallel data." In: *arXiv preprint arXiv:1710.04087* (2017).

[40] Ryan Cotterell and Georg Heigold. "Cross-lingual, character-level neural morphological tagging." In: *arXiv preprint arXiv:1708.09157* (2017).

[41] Michael Cysouw. "Chapter Predicting language-learning difficulty." In: *Approaches to measuring linguistic differences*. De Gruyter, 2013.

[42] Rajarshi Das, Manzil Zaheer, and Chris Dyer. "Gaussian LDA for topic models with word embeddings." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 795–804.

[43] Orphée De Clercq, Gert De Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French." In: *Translation Quarterly* 101 (2021), pp. 21–45.

[44] Sunipa Dev and Jeff Phillips. "Attenuating bias in word vectors." In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 879–887.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[46] Edsger W Dijkstra et al. "A note on two problems in connexion with graphs." In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[47] Matthew S Dryer. "Problems testing typological correlations with the online WALS." In: *Linguistic Typology* 13.1 (2009), pp. 121–135.

[48] Alan Duff. *The third language*. Pergamon Press New York, NY, 1981.

[49] Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. "Understanding Translationese in Multi-view Embedding Spaces." In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 6056–6062. DOI: 10.18653/v1/2020.coling-main.532. URL: https://aclanthology.org/2020.coling-main.532.

[50] Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. "Tracing Source Language Interference in Translation with Graph-Isomorphism Measures." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online: INCOMA Ltd., Sept. 2021, pp. 375–385. URL: https://aclanthology.org/2021.ranlp-main.43.

[51] Isidore Dyen, Joseph B Kruskal, and Paul Black. "An Indoeuropean classification: A lexicostatistical experiment." In: *Transactions of the American Philosophical society* 82.5 (1992), pp. iii–132.

[52] Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. "On The Evaluation of Machine Translation Systems Trained With Back-Translation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2836–2846. DOI: 10.18653/v1/2020.acl-main.253. URL: https://www.aclweb.org/anthology/2020.acl-main.253.

[53] Sauleh Eetemadi and Kristina Toutanova. "Asymmetric features of human generated translation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 159–164.

[54] Steffen Eger, Armin Hoenen, and Alexander Mehler. "Language classification from bilingual word embedding graphs." In: *arXiv preprint arXiv:1607.05014* (2016).

[55] Cristina España-Bonet and Josef van Genabith. "Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems." In: *Proceedings of the LREC 2018, MLP-MomenT Workshop*. Miyazaki, Japan, 2018, pp. 8–13.

[56] José Fernández Huerta et al. "Tres decenios de innovación didáctico-experimental (1943-1973)." In: (1983).

[57] Emilio Ferrara. "Should chatgpt be biased? challenges and risks of bias in large language models." In: *arXiv preprint arXiv:2304.03738* (2023).

[58] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. "Placing search in context: The concept revisited." In: *Proceedings of the 10th international conference on World Wide Web*. 2001, pp. 406–414.

[59] Benjamin W Fortson IV. *Indo-European language and culture: An introduction*. John Wiley & Sons, 2011.

[60] William Frawley. *Translation: Literary, linguistic, and philosophical perspectives*. Newark: University of Delaware Press; London: Associated University Presses, 1984.

[61] Markus Freitag, Isaac Caswell, and Scott Roy. "APE at Scale and Its Implications on MT Evaluation Biases." In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 34–44. DOI: 10.18653/v1/W19-5204. URL: https://www.aclweb.org/anthology/W19-5204.

[62] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes." In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[63] Martin Gellerstam. "Translationese in Swedish novels translated from English." In: *Translation studies in Scandinavia* 1 (1986), pp. 88–95.

[64] Jelena Golubović and Charlotte Gooskens. "Mutual intelligibility between West and South Slavic languages/." In: *Russian linguistics* (2015), pp. 351–373.

[65] Hila Gonen and Yoav Goldberg. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them." In: *arXiv preprint arXiv:1903.03862* (2019).

[66] Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 538–555. DOI: 10.18653/v1/2020.acl-main.51. URL: https://aclanthology.org/2020.acl-main.51.

[67] Antoine Gourru, Julien Velcin, and Julien Jacques. "Gaussian Embedding of Linked Documents from a Pretrained Semantic Space." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (July 2020). Ed. by Christian Bessiere. Main track, pp. 3912–3918. DOI: 10.24963/ijcai.2020/541. URL: https://doi.org/10.24963/ijcai.2020/541.

[68] Yvette Graham, Barry Haddow, and Philipp Koehn. "Translationese in Machine Translation Evaluation." In: *arXiv preprint arXiv:1906.09833* (2019).

[69] Yvette Graham, Barry Haddow, and Philipp Koehn. "Statistical Power and Translationese in Machine Translation Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online:

Association for Computational Linguistics, Nov. 2020, pp. 72–81. DOI: 10.18653/v1/2020.emnlp-main.6. URL: https://www.aclweb.org/anthology/2020.emnlp-main.6.

[70] Robert Gunning et al. "Technique of clear writing." In: (1952).

[71] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. "Large-scale learning of word relatedness with constraints." In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, pp. 1406–1414.

[72] John Hale. "A probabilistic Earley parser as a psycholinguistic model." In: *Second meeting of the north american chapter of the association for computational linguistics*. 2001.

[73] Hans van Halteren. "Source Language Markers in EUROPARL Translations." In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, Aug. 2008, pp. 937–944. URL: https://www.aclweb.org/anthology/C08-1118.

[74] Na-Rae Han, Martin Chodorow, and Claudia Leacock. "Detecting errors in English article usage by non-native speakers." In: *Natural Language Engineering* 12.2 (2006), pp. 115–129.

[75] Zellig S Harris. "Distributional structure." In: *Word* 10.2-3 (1954), pp. 146–162.

[76] Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. *The world atlas of language structures*. OUP Oxford, 2005.

[77] Felix Hill, Roi Reichart, and Anna Korhonen. "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." In: *Computational Linguistics* 41.4 (2015), pp. 665–695.

[78] Hagen Hirschmann, Anke Lüdeling, Ines Rehbein, Marc Reznicek, and Amir Zeldes. "Underuse of syntactic categories in Falko. A case study on modification." In: *Granger, S., Gilquin, G. und Meunier, F.(Hgg.), Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead* (2013), pp. 223–234.

[79] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[80] Matthew Honnibal and Mark Johnson. "An improved non-monotonic transition system for dependency parsing." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1373–1378.

[81] Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. "Identification of Translationese: A Machine Learning Approach." In: *Computational*

Linguistics and Intelligent Text Processing. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 503–511. ISBN: 978-3-642-12116-6.

[82] Klára Jágrová, Tania Avgustinova, Irina Stenger, and Andrea Fischer. "Language models, surprisal and fantasy in Slavic intercomprehension." In: *Computer Speech & Language* 53 (2019), pp. 242–275.

[83] Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. "Translating away Translationese without Parallel Data." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7086–7100. DOI: 10.18653/v1/2023.emnlp-main.438. URL: https://aclanthology.org/2023.emnlp-main.438.

[84] Frederick Jelinek. "Continuous speech recognition by statistical methods." In: *Proceedings of the IEEE* 64.4 (1976), pp. 532–556.

[85] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. zip: Compressing text classification models." In: *arXiv preprint arXiv:1612.03651* (2016).

[86] Masahiro Kaneko and Danushka Bollegala. "Autoencoding improves pre-trained word embeddings." In: *arXiv preprint arXiv:2010.13094* (2020).

[87] M. G. Kendall. *A New Measure of Rank Correlation*. 1938.

[88] Dorothy Kenny. "Creatures of habit? What translators usually do with words." In: *Meta: journal des traducteurs/Meta: Translators' Journal* 43.4 (1998), pp. 515–523.

[89] Paul Kerswill. "Children, adolescents, and language change." In: *Language variation and change* 8.2 (1996), pp. 177–202.

[90] Kimmo Kettunen, Markus Sadeniemi, Tiina Lindh-Knuutila, and Timo Honkela. "Analysis of EU languages through text compression." In: *International Conference on Natural Language Processing (in Finland)*. Springer. 2006, pp. 99–109.

[91] Philipp Koehn. "Europarl: A parallel corpus for statistical machine translation." In: *MT summit*. Vol. 5. Citeseer. 2005, pp. 79–86.

[92] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. "Moses: Open source toolkit for statistical machine translation." In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics. 2007, pp. 177–180.

[93] Arne Köhn. "What's in an embedding? Analyzing word embeddings through multilingual evaluation." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 2067–2073.

[94] Moshe Koppel and Noam Ordan. "Translationese and Its Dialects." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 1318–1326. URL: https://www.aclweb.org/anthology/P11-1132.

[95] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. "Hyperbolic geometry of complex networks." In: *Physical Review E* 82.3 (2010), p. 036106.

[96] David Kurokawa, Cyril Goutte, and P. Isabelle. "Automatic Detection of Translated Text and its Impact on Machine Translation." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, 2009.

[97] Guillaume Lample and Alexis Conneau. "Cross-lingual language model pretraining." In: *arXiv preprint arXiv:1901.07291* (2019).

[98] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A lite bert for self-supervised learning of language representations." In: *arXiv preprint arXiv:1909.11942* (2019).

[99] Ekaterina Lapshinova-Koltunski. "Variation in translation: Evidence from corpora." In: *New directions in corpus-based translation studies* 1 (2015), p. 93.

[100] Rémi Lebret and Ronan Collobert. "" The Sum of Its Parts": Joint Learning of Word and Phrase Representations with Autoencoders." In: *arXiv preprint arXiv:1506.05703* (2015).

[101] Gennadi Lembersky, Noam Ordan, and Shuly Wintner. "Language models for machine translation: Original vs. translated texts." In: *Computational Linguistics* 38.4 (2012), pp. 799–825.

[102] Ira Leviant and Roi Reichart. "Separated by an un-common language: Towards judgment language informed vector space modeling." In: *arXiv preprint arXiv:1508.00106* (2015).

[103] Roger Levy. "Expectation-based syntactic comprehension." In: *Cognition* 106.3 (2008), pp. 1126–1177.

[104] M Paul Lewis, GF Simons, and CD Fennig. "Ethnologue: Languages of the world [Eighteenth." In: *Dallas, Texas: SIL International* (2015).

[105] Mike Lewis, Yinyin Liu, Naman Goyal, Maruan Ghazvininejad, Aravind Mohamed, Omer Levy, and Veselin Stoyanov. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *International Conference on Machine Learning*. 2020.

[106] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. "MLQA: Evaluating cross-lingual extractive question answering." In: *arXiv preprint arXiv:1910.07475* (2019).

[107] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. "Smoothing the geometry of probabilistic box embeddings." In: *International Conference on Learning Representations*. 2018.

[108] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. "Towards debiasing sentence representations." In: *arXiv preprint arXiv:2007.08100* (2020).

[109] Stephanie Lin, Jacob Hilton, and Owain Evans. "Truthfulqa: Measuring how models mimic human falsehoods." In: *arXiv preprint arXiv:2109.07958* (2021).

[110] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. "Choosing transfer languages for cross-lingual learning." In: *arXiv preprint arXiv:1905.12688* (2019).

[111] Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. "Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, pp. 8–14.

[112] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. "Does gender matter? towards fairness in dialogue systems." In: *arXiv preprint arXiv:1910.10486* (2019).

[113] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692* (2019).

[114] Gary Lupyan and Bodo Winter. "Language is more abstract than you think, or, why aren't languages more iconic?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1752 (2018), p. 20170137.

[115] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[116] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. "Learning language representations for typology prediction." In: *arXiv preprint arXiv:1707.09569* (2017).

[117] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings." In: *arXiv preprint arXiv:1904.04047* (2019).

[118] José Manuel Martínez Martínez and Elke Teich. "Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations." In: (2017).

[119] Anna Mauranen. "Universal tendencies in translation." In: *Incorporating corpora: the linguist and the translator* (2008), pp. 32–48.

[120] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. "On measuring social biases in sentence encoders." In: *arXiv preprint arXiv:1903.10561* (2019).

[121] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[122] George Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[123] Jeff Mitchell and Mirella Lapata. "Vector-based models of semantic composition." In: *proceedings of ACL-08: HLT*. 2008, pp. 236–244.

[124] Steven Moran, Daniel McCloy, and Richard Wright. "PHOIBLE online." In: (2014).

[125] Mira Nábělková. "Closely-related languages in contact: Czech, Slovak,"Czechoslovak"." In: (2007).

[126] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. "Facebook FAIR's WMT19 News Translation Task Submission." In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 314–319. DOI: 10.18653/v1/W19-5333. URL: https://aclanthology.org/W19-5333.

[127] Johanna Nichols. *Linguistic diversity in space and time*. University of Chicago Press, 1992.

[128] Maximillian Nickel and Douwe Kiela. "Poincaré embeddings for learning hierarchical representations." In: *Advances in neural information processing systems* 30 (2017).

[129] Dmitry Nikolaev, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Saeboe, and Omri Abend. "Morphosyntactic predictability of translationese." In: *Linguistics Vanguard* 6.1 (2020).

[130] Giannis Nikolentzos, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. "Multivariate gaussian document representation from word embeddings for text categorization." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, pp. 450–455.

[131] Xing Niu and Marine Carpuat. "Discovering stylistic variations in distributional vector space models via lexical paraphrases." In: *Proceedings of the Workshop on Stylistic Variation*. 2017, pp. 20–27.

[132] Sebastian Nordhoff and Harald Hammarström. "Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources." In: *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*. 2011.

[133] Javad Nouri and Roman Yangarber. "Modeling language evolution with codes that utilize context and phonetic features." In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 2016, pp. 136–145.

[134] Arturo Oncevay, Barry Haddow, and Alexandra Birch. "Bridging linguistic typology and multilingual machine translation with multi-view language representations." In: *arXiv preprint arXiv:2004.14923* (2020).

[135] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. "Training language models to follow instructions with human feedback." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[136] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[137] Ji Ho Park, Jamin Shin, and Pascale Fung. "Reducing gender bias in abusive language detection." In: *arXiv preprint arXiv:1808.07231* (2018).

[138] Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. "Translation universals: do they exist? A corpus-based NLP study of convergence and simplification." In: *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*. 2008, pp. 75–81.

[139] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. "Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 184–193. DOI: 10.18653/v1/P19-1018. URL: https://www.aclweb.org/anthology/P19-1018.

[140] Judea Pearl et al. "Causal inference in statistics: An overview." In: *Statistics surveys* 3 (2009), pp. 96–146.

[141] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python." In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[142] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[143] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://aclanthology.org/N18-1202.

[144] Simone Pompei, Vittorio Loreto, and Francesca Tria. "On the accuracy of language trees." In: *PloS one* 6.6 (2011), e20109.

[145] Marius Popescu. "Studying Translationese at the Character Level." In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria: Association for Computational Linguistics, Sept. 2011, pp. 634–639. URL: https://www.aclweb.org/anthology/R11-1091.

[146] Heike Przybyl, Alina Karakanta, Katrin Menzel, and Elke Teich. "Exploring linguistic variation in mediated discourse: Translation vs. interpreting." In: *Mediated discourse at the European Parliament: Empirical investigations* (2022), pp. 191–218.

[147] Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. "Comparing Feature-Engineering and Feature-Learning Approaches for Multilingual Translationese Classification." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Lan-*

guage Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8596–8611. DOI: 10.18653/v1/2021.emnlp-main.676. URL: https://aclanthology.org/2021.emnlp-main.676.

[148] Ella Rabinovich, Noam Ordan, and Shuly Wintner. "Found in Translation: Reconstructing Phylogenetic Language Trees from Translations." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 530–540. DOI: 10.18653/v1/P17-1049. URL: https://www.aclweb.org/anthology/P17-1049.

[149] Ella Rabinovich and Shuly Wintner. "Unsupervised Identification of Translationese." In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 419–432. DOI: 10.1162/tacl_a_00148. URL: https://www.aclweb.org/anthology/Q15-1030.

[150] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving language understanding by generative pre-training." In: (2018).

[151] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. "Language models are unsupervised multitask learners." In: *OpenAI blog* 1.8 (2019), p. 9.

[152] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. URL: https://aclanthology.org/2020.acl-main.647.

[153] Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. "Translationese as a Language in "Multilingual" NMT." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7737–7746. URL: https://www.aclweb.org/anthology/2020.acl-main.691.

[154] Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. "Translationese as a Language in "Multilingual" NMT." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7737–7746. DOI: 10.18653/v1/2020.acl-main.691. URL: https://aclanthology.org/2020.acl-main.691.

[155] Don Ringe, Tandy Warnow, and Ann Taylor. "Indo-European and computational cladistics." In: *Transactions of the philological society* 100.1 (2002), pp. 59–129.

[156] Andrew Rosenberg and Julia Hirschberg. "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure." In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Ed. by Jason Eisner. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 410–420. URL: https://aclanthology.org/D07-1043.

[157] Herbert Rubenstein and John B Goodenough. "Contextual correlates of synonymy." In: *Communications of the ACM* 8.10 (1965), pp. 627–633.

[158] Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. "Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification." In: *Proceedings of NAACL-HLT 2016, Association for Computational Linguistics*. San Diego, California, 2016, pp. 960–970.

[159] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. "Gender bias in coreference resolution." In: *arXiv preprint arXiv:1804.09301* (2018).

[160] Magnus Sahlgren and Alessandro Lenci. "The effects of data size and frequency range on distributional semantic models." In: *arXiv preprint arXiv:1609.08293* (2016).

[161] Federica Scarpa. "Corpus-based quality-assessment of specialist translation: A study using parallel and comparable corpora in English and Italian." In: *Insights into specialized translation–linguistics insights. Bern: Peter Lang* (2006), pp. 155–172.

[162] Timo Schick, Sahana Udupa, and Hinrich Schütze. "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp." In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 1408–1424.

[163] Maurizio Serva and Filippo Petroni. "Indo-European languages tree by Levenshtein distance." In: *EPL (Europhysics Letters)* 81.6 (2008), p. 68005.

[164] Claude E Shannon. "A mathematical theory of communication." In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[165] Vijayalaxmi S Shigehalli and Vidya M Shettar. "Spectral techniques using normalized adjacency matrices for graph matching." In: *International Journal of Computational Science and Mathematics* 2.4 (2011), pp. 371–378.

[166] Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. "XLDA: Cross-lingual data augmentation for natural language inference and question answering." In: *arXiv preprint arXiv:1905.11471* (2019).

[167] Anders Søgaard, Sebastian Ruder, and Ivan Vulić. "On the Limitations of Unsupervised Bilingual Dictionary Induction." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 778–788. DOI: 10.18653/v1/P18-1072. URL: https://www.aclweb.org/anthology/P18-1072.

[168] Ilia Sominsky and Shuly Wintner. "Automatic Detection of Translation Direction." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sept. 2019, pp. 1131–1140. DOI: 10.26615/978-954-452-056-4_130. URL: https://www.aclweb.org/anthology/R19-1130.

[169] Andreas Stolcke. "SRILM – An extensible language modeling toolkit." In: *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*. 2002, pp. 901–904.

[170] Morris Swadesh. "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos." In: *Proceedings of the American philosophical society* 96.4 (1952), pp. 452–463.

[171] Ahmad Taie, Raphael Rubino, and Josef van Genabith. "INFODENS: An Open-source Framework for Learning Text Representations." In: *CoRR* abs/1810.07091 (2018). arXiv: 1810.07091. URL: http://arxiv.org/abs/1810.07091.

[172] Elke Teich. *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter, 2003.

[173] Elke Teich, José Martínez Martínez, and Alina Karakanta. "Translation, information theory and cognition." In: *The Routledge handbook of translation and cognition* (2020), pp. 9781315178127–24.

[174] Joshua B Tenenbaum, Vin De Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction." In: *science* 290.5500 (2000), pp. 2319–2323.

[175] Bill Thompson, Sean Roberts, and Gary Lupyan. "Quantifying semantic similarity across languages." In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*. 2018.

[176] Elad Tolochinsky, Ohad Mosafi, Ella Rabinovich, and Shuly Wintner. "The UN parallel corpus annotated for translation direction." In: *arXiv preprint arXiv:1805.07697* (2018).

[177] Antonio Toral. "Post-editese: an exacerbated translationese." In: *arXiv preprint arXiv:1907.00900* (2019).

[178] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation." In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 113–123. DOI: 10.18653/v1/W18-6312. URL: https://aclanthology.org/W18-6312.

[179] Gideon Toury. "Interlanguage and its manifestations in translation." In: *Meta: journal des traducteurs/Meta: Translators' Journal* 24.2 (1979), pp. 223–231.

[180] Gideon Toury. *In Search of a Theory of Translation*. Tel Aviv University, Tel Aviv: The Porter Institute for Poetics and Semiotics, 1980.

[181] Gideon Toury. *Descriptive translation studies and beyond: Revised edition*. Vol. 100. John Benjamins Publishing, 2012.

[182] Jan Vanhove. "Receptive multilingualism across the lifespan: Cognitive and linguistic factors in cognate guessing." PhD thesis. Université de Fribourg, 2014.

[183] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017).

[184] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. "Probabilistic embedding of knowledge graphs with box lattice measures." In: *arXiv preprint arXiv:1805.06627* (2018).

[185] Vered Volansky, Noam Ordan, and Shuly Wintner. "On the features of translationese." In: *Digital Scholarship in the Humanities* 30.1 (2015), pp. 98–118.

[186] Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. "Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity." In: *Computational Linguistics* (2020), pp. 847–897.

[187] Ivan Vulić, Sebastian Ruder, and Anders Søgaard. "Are All Good Word Vector Spaces Isomorphic?" In: *arXiv preprint arXiv:2004.04070* (2020).

[188] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. "Aligning large language models with human: A survey." In: *arXiv preprint arXiv:2307.12966* (2023).

[189] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function." In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[190] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. "Finetuned language models are zero-shot learners." In: *arXiv preprint arXiv:2109.01652* (2021).

[191] Laura Wendlandt, Jonathan K Kummerfeld, and Rada Mihalcea. "Factors influencing the surprising instability of word embeddings." In: *arXiv preprint arXiv:1804.09692* (2018).

[192] Søren Wichmann and Anthony P Grant. *Quantitative approaches to linguistic diversity: commemorating the centenary of the birth of Morris Swadesh*. Vol. 46. John Benjamins Publishing, 2012.

[193] Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. "On the universal structure of human lexical semantics." In: *Proceedings of the National Academy of Sciences* 113.7 (2016), pp. 1766–1771.

[194] C Udny Yule. *The statistical study of literary vocabulary*. Cambridge University Press, 2014.

[195] Mike Zhang and Antonio Toral. "The Effect of Translationese in Machine Translation Test Sets." In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 73–81. DOI: 10.18653/v1/W19-5208. URL: https://www.aclweb.org/anthology/W19-5208.

[196] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: https://aclanthology.org/N18-2003.

[197] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. "Learning gender-neutral word embeddings." In: *arXiv preprint arXiv:1809.01496* (2018).

[198] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. "A survey of large language models." In: *arXiv preprint arXiv:2303.18223* (2023).

[199] Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. "Density Matching for Bilingual Word Embedding." In: *CoRR* abs/1904.02343 (2019). arXiv: 1904.02343. URL: http://arxiv.org/abs/1904.02343.