Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09726-y

ORIGINAL PAPER

An aligned corpus of Spanish bibles

Gerardo Sierra[1] · **Gemma Bel-Enguix**[1] · **Ameyali Díaz-Velasco**[1] · **Natalia Guerrero-Cerón**[1] · **Núria Bel**[2]

Accepted: 26 January 2024
© The Author(s) 2024

Abstract
We present a comprehensive and valuable resource in the form of an aligned parallel corpus comprising translations of the Bible in Spanish. Our collection encompasses a total of eleven Bibles, originating from diverse centuries (XVI, XIX, XX), various religious denominations (Protestant, Catholic), and geographical regions (Spain, Latin America). The process of aligning the verses across these translations has been meticulously carried out, ensuring that the content is organized in a coherent manner. As a result, this corpus serves as a useful convenient resource for various linguistic analyses, including paraphrase detection, semantic clustering, and the exploration of biases present within the texts. To illustrate the utility of this resource, we provide several examples that demonstrate how it can be effectively employed in these applications.

Keywords Aligned corpus · Paraphrase detection · Semantic clustering · Subjective bias · Bibles corpus · Dialectal differences

1 Introduction

The Bible holds the distinction of being the most translated book in the world, whether in its complete version or in parts, like the New Testament. Although it is difficult to have an exact inventory, Wikipedia states[1] that there are fully translated versions in at least 700 languages.

The extension, multiplicity and systematicity of translations make the Bible an ideal book for typological (Cysouw & Wälchli, 2007; Wälchli, 2007), dialectal and even evolutionary analysis of many languages. Moreover, as mentioned by de Vries

[1] https://en.wikipedia.org/wiki/Bible_translations.

✉ Gerardo Sierra
gsierram@iingen.unam.mx

[1] Instituto de Ingeniería, Universidad Nacional Autónoma de México, Mexico City, Mexico

[2] Universitat Pompeu Fabra, Barcelona, Spain

Published online: 15 March 2024

(2007), the Bible is the resource par excellence for the creation of parallel corpora, or MPT (massive parallel texts) (Cysouw & Wälchli, 2007). The Bible is, in fact, a text on which different translation techniques have been studied, perfected and tested. However, its use as a reference corpus for translation or as MPT is not simple, since there are religious, political and theological criteria that condition translations beyond stylistic or geographical choices (de Vries, 2007; Lavidas, 2022).

When it comes to Bible translations, the goal is to capture the intended message of the text. There are different perspectives on translation, ranging from the literal version, also called formally equivalent, to the free or functionally equivalent.

Klein et al. (2017) mention that formally equivalent translations aim to be faithful to the original text by providing a literal word-for-word translation. The goal is to be the closest possible to the original, preferring accuracy to clarity or intelligibility for current readers. At the other extreme are the dynamically or functionally equivalent translations. This is a phrase to phrase technique that is more concerned with the clarity of the sentences, rewording the texts, and interpreting meaning in context. Between these two extremes, optimally equivalent translations seek to find the best possible translation for each passage, leaning toward one approach or the other.

These different perspectives about how the best biblical translation should be performed favor the diversity of the resultant texts, and make it more useful the task of compiling corpora of aligned versions of the book.

This paper introduces a corpus consisting of 11 Spanish translations of the Bible, covering different centuries and geographical regions. It is important to emphasize that this is not an MPT, since only the different translations of the Bible are included, and not their source texts. The resource is specifically designed for monolingual studies. The eleven different versions have been aligned based on verse distribution and sense equivalence. This resource serves as a foundation for conducting paraphrase experiments and suggests further studies in lexical, semantic, and ideological aspects.

The rest of the paper is structured as follows. Section 2 introduces some related work in the area. Section 3 explains the elaboration of the corpus, including the compilation, labeling and alignment. Section 4 explains the different phenomena that have been used to generate paraphrases in the corpus, and the methods to identify them. In Sect. 4 we show three examples of applications: paraphrase detection and analysis (4.1) semantic clustering (4.2) and bias (4.3). The paper closes with the conclusion and suggestions for future work in Sect. 5.

2 Related work

Due to its status as the most translated book in history, the Bible has been extensively explored in computational linguistics as a valuable corpus, particularly as a parallel corpus.

Resnik et al. (1999) were among the pioneers in this field, creating the first annotated parallel corpora for the 66 books of the Protestant canon of the Bible in eight languages, and the New Testament in twelve languages.

At the University of Oslo, the project PROIEL (Pragmatic Resources in Old Indo-European Languages) started to build a family of parallel treebanks of the oldest Indo-European New Testament translations, annotated with morphological, syntactic and discourse annotation (Eckhoff et al., 2018).

In a similar vein, Christodouloupoulos and Steedman (2015) presented a multilingual corpus comprising translations into 100 languages. However, while 55 have the whole text, 45 languages only have partial texts, mostly focusing on the New Testament. Some languages, such as Armenian, Chamorro, Gaelic, Manx, and Tamajaq, have even smaller fragments. The diversity of alphabets used in the corpus poses a challenge for researchers.

Mayer and Cysouw (2014) created a multilingual corpus consisting of 994 aligned translations spanning 76 language families. McCarthy et al. (2020) elaborated a verse-aligned corpus comprising 4,272 Bible translations in 1,611 languages, including 27 English versions. This resource is specially designed for typological analysis.

All these multilingual resources serve as convenient tools for various tasks in computational linguistics, including comparative linguistics, language typology, automatic translation and lexical extraction, among others.

Moreover, the creation of aligned corpora for different versions of the Bible in the same language proves to be very useful for computational linguistics tasks related to language change, bias and ideology studies, and variants detection. For instance, the Biblia Medieval project (Enrique-Arias and Pueyo Mena, 2008) focuses on 14 medieval aligned translations of the Bible in Spanish, aligned with their source text. In English, specific parts of McCarthy et al. (2020), for example, can be taken in order to compare the versions in this language.

Beyond Bibles' alignment, the field of computational linguistics shows a growing interest in constructing aligned text for various tasks, including paraphrase detection. In Spanish, there are a few resources with such features. Herrera et al. (2007) created a dataset comprising 393 labeled pairs of Spanish questions from the QA@CLEF dataset. Queralt et al. (2018) developed a corpus based on 50 journal notes divided into five topics. Gómez-Adorno et al. (2020) assembled the Sushi Corpus through manual paraphrasing, providing a baseline for research in this area.

3 Corpus

Linguistic work with the Bible has advantages and disadvantages. Among the advantages, it stands out that it is a text long enough (more than 700 k words in Spanish) to carry out typological experiments on it. Furthermore, it is a closed corpus, to which no more texts are added, and therefore it is a controlled experimental environment. Another characteristic is that all translators want to convey the essence of the text as faithfully as possible. This means that, despite the divergent perspectives of translation (free vs. literal), the differences in the final result are not as great as one would expect. Finally, as it is a text that has been translated incessantly over the centuries, in different geographies, various types of variation can be explained from the translations, for example diachronic, geographical, ideological.

For many purposes, the Bible has an additional advantage: its division into verses. Verses are the minimum units of the text, which can correspond to a verse of the original, or a sentence. All translations preserve this division, and this makes preliminary alignment much easier. That is, in theory, Genesis chapter 1 verse 1 has to align with that same chapter and verse in all versions of Genesis. However, this is not always the case. Because of different perspectives on translation, preservation of the originals, theological divergences, and other factors, not all verses line up perfectly.

Sometimes a verse or a complete chapter can be omitted, for example, because they are not accepted in the canon, because the origin is not clear or because they are considered late additions.

When translating, there are also stylistic disparities that can lead to the merging of two verses into one version or, conversely, the division or reordering of a single verse. These modifications are sometimes prompted by the desire to adapt the expression to the target language or to enhance the story's expressiveness.

The Christian Bible is a collection of books that varies in number depending on the specific tradition and is divided into two main groups: the Old Testament and the New Testament. The canon of the Old Testament differs among Protestant, Catholic, and Orthodox traditions, with 39 books, 46 books, and 51 books respectively. Catholic Bibles include additional books known as the Deuterocanonical books, which are not found in Protestant Bibles. These books are considered part of the Old Testament by Catholics and include Tobit, Judith, Baruch, Ecclesiasticus (also known as Sirach), Wisdom of Solomon, and First and Second Maccabees. The Catholic Church considers these books to be inspired and authoritative for teaching and doctrine. The New Testament is common to all Christian traditions and consists of 27 books, written in Greek and universally accepted across Christian denominations.

The Old Testament was written between the ninth century BC and 1st AD by multiple authors using different languages. Most parts of the Old Testament were originally written in Hebrew, with some sections in Aramaic, while the original language of certain later books is Koine Greek.The New Testament, on the other hand, was written entirely in Koine Greek. Despite the problems of text fixation and the many interferences between one adaptation and another, we will call the widely accepted version that contains the three aforementioned languages 'original'. Thus, when authors want to translate the Bible from the original they have to use the three languages source.

Three important versions emerged from this original. A) Septuagint. A translation of the Old Testament into Greek, dating from the 3rd to the first centuries BC. B) Vulgate. A translation of the original and Septuagint into Latin made by Jerome in the fourth century AD. This version was adopted as the official one by the Catholic Church in the sixteenth century. C) Masoteric. The Hebrew version of the Old Testament used by Jewish communities. It is adapted not only linguistically, but conceptually. It was copied and spread between the seventh and tenth centuries AD.

All Bible translations follow one of these four sources, three of them being versions of the original texts in turn. Although in the last decades the tendency to translate from the original has been consolidated, this is not valid for all the Bibles in our

Table 1 Editions of the Bible in the corpus, ordered by year

Title	Code	Verses	Books	Words	Year	Place	Translator	Source
La Biblia, que es, los Sacros Libros del Viejo y Nuevo Testamento	OSO	31102	66	726,859	1569	Switzerland	de Reina	Masoteric
La Santa Biblia	REV	30952	65	734,435	1602	Netherlands	de Valera	Masoteric
La Sagrada Biblia	PET	31169	66	796,828	1825	Spain	Torres Amat	Vulgate
La Sagrada Biblia	JNM	46743	73	782,697	1928	Chile	Jünemann	Septuagint
La Sagrada Biblia	NAC	31069	66	783,346	1944	Spain	Nacar and Colunga	Original
La Biblia de Jerusalén	JER	35155	73	769,850	1956	Paris	Jerusalem Bible Sch	Original
La Santa Biblia	EMN	35130	73	827,570	1961	Spain	Martín Nieto	Original
La Nueva Biblia—Edición pastoral para Latinoamérica	LAT	35432	73	785,199	1972	Chile	Ricciardi and Hurault	Original
Nueva Biblia Española	ESP	30946	66	657,346	1975	Spain	Schökel and Mateos	Original
La Biblia	SER	30862	65	722,481	1975	Spain	Serafín de Ausejo	Original
La Biblia de las Américas	AME	27761	65	720,596	1986	USA	Lockman Foundation	Original

corpus. The last column of Table 1 indicates the main source each of them claims to follow.

The Bible shows a great diversity of literary genres, including instructions, narratives, wisdom literature, poetry, prophecies, laws, and myths (Lawrenz, 2014). These diverse genres contribute to the richness and depth of the biblical text.

Our corpus contains 11 different versions of the Bible in Spanish.

The editions that integrate our corpus were created during different time periods, but most of them were published or edited during the XX century and in different Spanish speaking countries (Table 1).

The corpus comprises a total of 754 books, encompassing a voluminous collection of 7,970,269 words. This extensive dataset is accessible on the online platform Github[2] as well as on the corpus manager GECO4,[3] facilitating convenient access and retrieval for scholarly purposes. GECO features various user functionalities and applications, including a concordance viewer and options for downloading texts in plain format or tagged with part-of-speech and lemmas (Sierra et al., 2017). To

[2] https://github.com/GIL-UNAM/SpanishParaphraseCorpora/tree/main/Biblias

[3] http://www.geco.unam.mx/geco3/proyecto/CPBE

enhance efficiency in book localization and referencing, a unique code has been assigned to each Bible, assuming that they share the same book names.

The first two Bibles, in chronological order, translated by Casiodoro de Reina (OSO) and revised by Cipriano de Valera (REV), are well-known Protestant versions in Spanish from the sixteenth and seventeenth centuries, respectively. Casiodoro de Reina's translation was first published in Switzerland in 1569 and Cipriano de Valera later revised and published an updated version in 1602. Casiodoro de Reina's translation was primarily based on the Hebrew Masoretic text but also took into consideration other versions in Latin, Greek, and even Judeo-Spanish. These, being Protestant versions, do not include the deuterocanonical books such as Tobit, Judith, Baruch, Sirach, 1 Maccabees, 2 Maccabees, and Wisdom. However, it's worth noting that, in our corpus, there are several Catholic versions that do not present them either; among them, Torres Amat (PET), La Biblia de las Américas (AME), Nueva Biblia Española (ESP) and La Biblia (SER) are also in this group.

In 1823, Torres Amat (1772-1847) released his translation of the Bible known as PET. This translation draws heavily from the work of the Jesuit Petisco, who had made a version towards the close of the eighteenth century. Torres Amat's translation is primarily based on the Vulgate, while also considering the Hebrew and Greek versions of the Bible.

Our next text is a Latin American Translation made in Chile by Wilhelm Jünemann (JUN) (1855–1938), a Chilean Catholic priest of German origin. The translation of the New Testament was published in 1928, and the Old Testament was released in 1992, many years after Jünemann's death. Junemann's translation of the Old Testament was not modified in the published version. Therefore, the volume must be, for all intents and purposes, dated to the 1920s or 1930s (twentieth century). Junemann's version is the first completed in Latin American, and the first in Spanish translated from the Septuagint.

Nacar and Colunga's Sagrada Biblia (NAC) was published in Spain in 1944. This translation derives from the original Hebrew, Aramaic, and Greek texts. It holds a prominent position among Catholic versions, having been disseminated through more than thirty editions. The authors were two Dominica priests, Eloíno Nacar and Alberto Colunga. They express their intention to maintain fidelity to the original in the prologue (Nacar and Colunga, 1944: XLI). However, they "do not believe that fidelity obliges the translator to slavishly follow the letters of the original, reproducing it exactly with Spanish words… The translator has to pay attention to the words of the text, but more than them he has to pay attention, mainly, to the meaning of the phrases, to give it with scrupulous fidelity to the language into which it translates". Notwithstanding the inclusion of the 73 canonical books in NAC, our specific version only comprises 66 books.

The Biblia de Jerusalén (JER) is a collaborative translation project undertaken by the École biblique et archéologique française de Jérusalem, led by Dominican scholars. The primary objective of this translation endeavor is to facilitate biblical exegesis, and as such, it is guided by principles of historical criticism. Notably, this work is distinguished by the inclusion of historical and linguistic comments within the text, which serve to provide valuable insights and analysis. Since its initial publication in 1966, the Biblia de Jerusalén has undergone several revisions. However,

the version referred to in this context is the original 1966 edition. The translation draws upon the original Hebrew, Aramaic, and Greek source texts, adhering to the same guidelines employed in the French translation. Additionally, the comments and critical apparatus present in the Biblia de Jerusalén are directly derived from the corresponding French edition.

La Santa Biblia (EMN) is the first Spanish version made by a team of translators, led by Evaristo Martín Nieto. This translation aimed to provide a faithful representation of the original texts, adhering closely to the wording and structure of the source languages. The focus on formal equivalence means that the translation seeks to maintain a close correspondence to the original text, emphasizing accuracy and consistency. The first edition was released in 1961.

The Biblia Latinoamericana (LAT) is translated from Hebrew, Aramaic and Greek with a functionally equivalent style. It is specifically oriented to Latin American readers. It was printed for the first time in 1972, after the Second Vatican Council (1962–1965), the world council of the Catholic Church that promoted a deep updating of the theology and structure of the Church. It has a progressive ideology in the framework of the Catholic Church. It was branded as a tool of liberation theology. It is very popular among people from Latin America. The translation was led by the priests Bernardo Hurault and Ramón Ricciardi.

Nueva Biblia Española (ESP) is a functionally equivalent translation from the original texts. The translators, Luis Alonso Shöckel and Juan Mateos, had the goal to make the text more accessible to the people. To achieve this, the translation employed a free style that used expressions and language that could bridge the cultural gap between the modern world and the historical context in which the biblical texts were written. The intention was to help readers connect with the message of the Bible in a way that resonated with their own cultural and linguistic background.

A team of translators under the coordination of Serafín de Ausejo (SER) published in 1975 a version that is considered to be a more literal Spanish version that follows the principles of formal equivalence as closely as possible. This translation aims to maintain a faithful representation of the original texts, emphasizing accuracy in its rendering of the source languages.

The title Biblia de las Americas (AME) aptly reflects the intended audience and purpose of this edition, as it primarily targets Spanish-speaking individuals residing in the Americas. The translation itself is a collaborative effort involving individuals from diverse Protestant denominations, all working from the original text. The Biblia de las Americas was first published in 1986, making it a relatively recent addition to the array of Spanish Bible translations available.

3.1 Compilation and labeling

The primary criterion for book selection was the availability of digital versions. The initial search involved locating complete editions that could be downloaded and converted into text documents. Subsequently, the Bibles were transformed into plain text files (.txt). However, these documents retained subtitles, notes,

Table 2 Codes of the Bibles and books. Every book has a 6 characters code with the form BIBLE + BOOK

Books	CODE	Books	CODE	Books	CODE
Genesis	GEN	Ecclesiastes	ECL	Luke	LUC
Exodus	EXD	Song of Songs	CNT	John	JUN
Leviticus	LEV	Wisdom	SAB	Acts Apostles	HCH
Numbers	NUM	Sirach	ECS	Romans	ROM
Deuteronomy	DET	Isaiah	ISA	1 Corintians	ICO
Joshua	JOS	Jeremiah	JER	2 Corintians	2CO
Judges	JCS	Lamentations	LAM	Galatians	GAL
Ruth	RUT	Baruch	BAR	Ephesians	EFS
1 Samuel	1SM	Ezekiel	EZQ	Philippians	FIL
2 Samuel	2SM	Daniel	DAN	Colossians	COL
1 Kings	IRY	Hosea	OSE	1 Thessalonians	1TL
2 Kings	2RY	Joel	JOL	2 Thessalonians	2TL
1 Chronicles	1CR	Amos	AMS	1 Timothy	1TM
2 Chronicles	2CR	Obadia	ABD	2 Timothy	2TM
Ezra	ESD	Jonah	JON	Titus	TIT
Nehemiah	NHM	Micah	MIQ	Philemon	FLM
Tobit	TBS	Nahum	NAH	Hebrews	HBR
Judith	JDT	Habakkuk	HAB	Santiago	SNT
Esther	EST	Zephaniah	SOF	1 Peter	1PD
1 Macabees	1MC	Haggai	HAG	2 Peter	2PD
2 Macabees	2MC	Zechariah	ZAC	1 John	1JN
Job	JOB	Malachi	MAL	2,3 John	2JN, 3JN
Psalms	SAL	Mathew	MAT	Jude	JUD
Proverbs	PRV	Mark	MAR	Revelation	APC

HTML symbols, and other extraneous elements, which were excluded to retain only the verse content. Separate plain text files were created for each book within every Bible.

To facilitate easy identification of the Bible version and book, each document was assigned a unique six-letter code. The first three letters denote the specific Bible (refer to Table 1, CODE column), while the last three letters indicate the book. For instance, the book of Genesis from the Latinoamerican Bible would be designated as "LATGEN." The books included in the corpus are detailed in Table 2, although some books may not be present in all eleven versions.

Furthermore, within each document, each verse has been assigned a unique code comprising nine digits. The first three letters of the code represent the book, followed by three numbers indicating the chapter, and concluding with three numbers corresponding to the verse number. As an example, the first verse of Genesis appears in every version as "GEN001001." Hence, "OSOGEN001001" denotes the first verse of the Valera version.

In annex 1 we include a table with the number of words of each book in each one of the translations.

3.2 Corpus alignment

The number of books in the various translations ranges from 64 to 73, indicating discrepancies in the inclusion or exclusion of certain books across editions. Additionally, differences in the number of chapters within a book and the number of verses within each chapter can also be observed across different editions. These variations stem from factors such as differing translation approaches, challenges in interpreting the text, the use of different source texts, or ideological differences among translators.

These disparities pose challenges when attempting to compare the same verse across different editions, as the content may not align due to variations in the number of verses. This divergence in verse count makes it impractical to automatically align texts from different editions. To address this issue, a Python program was used to compare the number of verses in the different translations of each book and categorize them based on differences in verse count within chapters. For example, the analysis revealed that certain books, like Jude or Lamentations, exhibit matching verse counts and content across editions. However, books like Psalms present significant variations with up to 10 different variants observed.

Following the initial automatic alignment, a manual review of the 11 Bibles and their chapters was conducted. The alignment process involved referencing the version with the highest number of verses as a guide for each book and chapter. This means that no single edition served as the alignment reference throughout; instead, it varied depending on the specific book and chapter being considered. Verses that were omitted in a particular version were marked and labeled with the text "(TEXTO OMITIDO)" to indicate the omission. In cases where an entire chapter was omitted, the reviewer would write the complete chapter code followed by "(TEXTO OMITIDO)" on each line.

At times, there were instances where the enumeration of the last verse in a chapter was missing, but the omitted content actually existed at the beginning of the chapter. In such cases, efforts were made to identify the omitted verse and reorganize the information to align with all the different versions. The English translation of each of the examples that appears in the subsequent tables has been done using Google in order to preserve as much as possible the differences that can be seen in the Spanish text.

In Table 3, it is evident that there is a lack of alignment among the translated text across different versions. Upon careful examination of the preceding text within the chapter, it was discovered that the passage identified as LUC017027 in the ESP Bible was divided into two verses, resulting in a discrepancy in the subsequent information in chapter 17. Furthermore, the corresponding passages identified as LUC017036 and LUC017037 in other translations were merged into a single verse in order to maintain consistent numbering, but without considering the alignment of content. To address this issue, the divided portions of LUC017027 were combined,

Table 3 Example of non-coincident alignment of a passage from Luke in OSO, JER, LAT, ESP

	OSO	JER	LAT	ESP
LUC017034	Os digo que aquella noche estarán dos en una cama; el uno será tomado, y el otro será dejado	Yo os lo digo: aquella noche estarán dos en un mismo lecho: uno será tomado y el otro dejado	Yo les declaro, que aquella noche, de dos personas que estén durmiendo en una misma cama, una será llevada y la otra dejada	Esto les digo: Aquella noche estarán dos en una cama, a uno se lo llevarán y al otro lo dejarán
	I tell you that that night there will be two in one bed; the one will be taken, and the other will be left	I tell you: that night there will be two in the same bed: one will be taken and the other left	I declare to you, that that night, of two people who are sleeping in the same bed, one will be taken and the other left	I tell you this: That night there will be two in one bed, one will be taken away and the other will be left
LUC017035	Dos mujeres estarán moliendo juntas; la una será tomada, y la otra será dejada. Dos estarán en el campo; el uno será tomado, y el otro será dejado	habrá dos mujeres moliendo juntas: una será tomada y la otra dejada.»	dos mujeres estarán moliendo juntas, pero una será llevada y la otra dejada”	estarán dos moliendo juntas, a una se la llevarán y a la otra la dejarán’
	Two women will be grinding together; the one will be taken, and the other will be left. Two will be in the field; the one will be taken, and the other will be left	there will be two women grinding together: one will be taken and the other left."	two women will be grinding together, but one will be taken and the other left”	there will be two milling together, one will be taken away and the other will be left’
LUC017036	Y respondiendo, le dicen: ¿Dónde, Señor?	Y le dijeron: «¿Dónde, Señor?»	Entonces preguntaron a Jesús: “¿Dónde sucederá eso, Señor?”	Ellos le preguntaron: ¿Dónde será, Señor?
	And answering, they say to him: Where, Lord?	And they said to him: "Where, Lord?"	So they asked Jesus: “Where will that happen, Lord?”	They asked him: Where will it be, Lord?
LUC017037	Y él les dijo: Donde estuviere el cuerpo, allá se juntarán también las águilas	El les respondió: «Donde esté el cuerpo, allí también se reunirán los buitres.»	Y él respondió: “Donde esté el cuerpo, allí se juntarán los buitres”	El contestó: Donde se reúnen los buitres, allí está el cuerpo
	And he said to them: Where the body is, there the eagles will also gather	He answered them: "Where the carcass is, there also the vultures will gather."	And he replied: "Where the body is, there the vultures will gather."	He replied: Where the vultures gather, there is the body

the remaining passages were renumbered, and the final passage was split to achieve proper alignment of the information.

In contrast, the OSO Bible contained additional information in LUC017036 that was not present in the other translations, resulting in a gap in content. Additionally, the verses corresponding to LUC017036 and LUC017037 in the other Bibles were merged into a single verse. To align this, the information from LUC017036 was moved to LUC017035, and the content of LUC017037 was divided into LUC017036 and LUC017037 (Table 4).

Furthermore, in the JER Bible, chapter 17 consisted of 36 verses because the information from verses LUC017036 and LUC017037 was incorporated into LUC017036. To align this Bible with the others, verse LUC017036 was divided.

On the other hand, the LAT Bible did not require any alterations as it shared both numbering and content with the other Bibles in the corpus.

In cases where a verse is missing in one chapter, it may be found in the following chapter, and the information will be aligned according to the other books, or the majority of them. It should be noted that verses cannot be deleted, but the information can be reorganized, and verses can be divided or combined. Once the review and corrections are completed, all the translations of a book should have the same number of verses and closely agree in their content.

Furthermore, the aligned books are divided into folders with the name of the book in Spanish, where you can find the books of the different translations.

4 Applications of the aligned corpus of Bibles

The aligned parallel corpus, encompassing 11 distinct Spanish translations of the Bible, represents an invaluable resource for sophisticated linguistic analyses and algorithmic advancements within the realm of NLP. In this section, we present three distinct tasks that can be undertaken using this corpus: Firstly, paraphrase detection (referred to as PARAPHRASING) encompasses an exploration of the diverse categories of paraphrastic phenomena manifesting within pairs of verses. Secondly, semantic clustering (referred to as CLUSTERING) entails the identification of words that can be used in the same sentence without altering their semantic meaning. Lastly, the analysis of subjective bias (referred to as BIAS) provides a framework for discerning and visualizing the ideological stances present within the translations. These tasks stand poised to illuminate multifaceted dimensions within the corpus, facilitating a deeper understanding of linguistic variations and ideological perspectives across the ensemble of translations.

For the purpose of conducting our experiments, we have selected four editions: Shoeckel and Mateos (ESP), Nácar-Colunga (NAC), Junemann (JUN), and the Latin American edition (LAT). These editions are representative of the twentieth century and are divided equally between publications from Spain and Latin America. As Catholic versions, they may exhibit significant variations based on whether they were published before or after the Second Vatican Council (SVC).

The Second Vatican Council (1962–1965) promoted a theological renewal that revitalized the task of translating ancient language texts (Latin, Greek, Hebrew) to

Table 4 Example of an aligned passage from Luke in OSO, JER, LAT, ESP

	OSO	JER	LAT	ESP
LUC017034	Os digo que aquella noche estarán dos en una cama; el uno será tomado, y el otro será dejado	Yo os lo digo: aquella noche estarán dos en un mismo lecho: uno será tomado y el otro dejado	Yo les declaro, que aquella noche, de dos personas que estén durmiendo en una misma cama, una será llevada y la otra dejada	El que pretenda poner su vida al seguro, la perderá; y en cambio, el que la pierda, la recobrará
	I tell you that that night there will be two in one bed; the one will be taken, and the other will be left	I tell you: that night there will be two in the same bed: one will be taken and the other left	I declare to you, that that night, of two people who are sleeping in the same bed, one will be taken and the other left	Whoever tries to put his life in insurance will lose it; and instead, whoever loses it, will recover it
LUC017035	Dos mujeres estarán moliendo juntas; la una será tomada, y la otra será dejada	habrá dos mujeres moliendo juntas: una será tomada y la otra dejada.»	dos mujeres estarán moliendo juntas, pero una será llevada y la otra dejada"	Esto les digo: Aquella noche estarán dos en una cama, a uno se lo llevarán y al otro lo dejarán
	Two women will be grinding together; the one will be taken, and the other will be left	there will be two women grinding together: one will be taken and the other left."	two women will be grinding together, but one will be taken and the other left"	I tell you this: That night there will be two in one bed, one will be taken away and the other will be left
LUC017036	Dos estarán en el campo; el uno será tomado, y el otro será dejado	Y le dijeron: «¿Dónde, Señor?» El les respondió: «Donde esté el cuerpo, allí también se reunirán los buitres.»	Entonces preguntaron a Jesús: "¿Dónde sucederá eso, Señor?"	estarán dos moliendo juntas, a una se la llevarán y a la otra la dejarán'
	Two will be in the field; the one will be taken, and the other will be left	And they said to him: "Where, Lord?" He answered them: "Where the carcass is, there also the vultures will gather."	So they asked Jesus: "Where will that happen, Lord?"	there will be two milling together, one will be taken away and the other will be left'
LUC017037	Y respondiendo, le dicen: ¿Dónde, Señor? Y él les dijo: Donde estuviere el cuerpo, allá se juntarán también las águilas		Y él respondió: "Donde esté el cuerpo, allí se juntarán los buitres"	Ellos le preguntaron: ¿Dónde será, Señor? El contestó: Donde se reúnen los buitres, allí está el cuerpo
	And answering, they say to him: Where, Lord? And he said to them: Where the body is, there the eagles will also gather		And he replied: "Where the body is, there the vultures will gather."	They asked him: Where will it be, Lord? He replied: Where the vultures gather, there is the body

Table 5 Versions of Bible compared in the paper

	Spain	Latin America
Before SVC	NAC	JUN
After SVC	ESP	LAT

make them more accessible to the less educated faithful. Several different versions were produced for European and American Spanish-speaking audiences, with the aim of employing a lexicon that is more easily understood by a wider public.

Additionally, it is noteworthy that NAC and JUN are renowned for their formally equivalent translations, while ESP and LAT tend to adopt a phrase-to-phrase approach (Table 5).

4.1 Paraphrase detection

Paraphrase can be defined as the generation of two distinct sentences that possess semantic equivalence (Das & Smith, 2009). Traditional definitions of paraphrase emphasize the semantic similarity between two texts expressed using different words (Hirst, 2003; Zhou et al., 2006). Typically, paraphrasing involves the utilization of synonyms to convey a meaning very similar to the original statement. However, the resulting sentences may not always be entirely equivalent, leading to what Bhagat and Hovy (2013) refer to as quasi-paraphrase or approximate paraphrase. In this context, Castro et al. (2011) distinguish between low-level and high-level paraphrase. The former relies on the use of synonymous terms, while the latter incorporates syntactic and discursive variations.

To address the challenge of paraphrase detection, machine learning methods have emerged as the most successful approach. These methods rely on the availability of datasets comprising pairs of sentences that are labeled as either paraphrase or non-paraphrase. Various techniques have been employed to construct paraphrase corpora, such as utilizing question-answering systems (Dong et al., 2017), manual creation (Gómez-Adorno et al., 2020), sometimes with the aid of crowdsourcing (Burrows et al., 2013; Xu et al., 2015), back-translation (Creutz, 2018), and multiple translations (Farwell et al., 2009).

Paraphrases can take various forms, encompassing a spectrum of linguistic alterations while preserving the core meaning. Several categories of paraphrases exist, reflecting different ways in which language can be rephrased or reformulated. Identifying these categories is crucial for accurate and comprehensive paraphrase detection. In our investigation of paraphrase in Bible translations, we have incorporated the theoretical categories presented by Barrón-Cedeño et al. (2010) and Mota-Montoya et al. (2016). From these taxonomies, we have carefully selected four specific categories that we believe offer sufficient coverage for our study.

By considering these distinct categories of paraphrase, our aim is to delve into the multifaceted approaches employed by biblical translations in utilizing various linguistic strategies. Through this exploration, we seek to uncover the diverse methods by which these translations effectively convey semantically equivalent

messages. By examining the phenomenon of paraphrase within the context of biblical texts, we can gain valuable insights into the intricate interplay between language, meaning, and the transmission of religious and spiritual concepts. Ultimately, our investigation contributes to a deeper understanding of how linguistic choices impact the interpretation and dissemination of biblical content across different translations.

The first category is substitution, which involves the replacement of lexical units while maintaining the same underlying semantic content. This linguistic phenomenon allows for the use of different words or phrases to convey an equivalent meaning.

The second category we have considered is morphological modification, which occurs when two words sharing the same lemma undergo alterations in their morphology. This type of paraphrase explores modifications within the inflectional or derivational forms of words while preserving their fundamental semantic relationship.

Another category we have focused on is modification in the order of words. This pertains to instances where two sentences exhibit semantic equivalence, but the arrangement of constituents within each sentence differs. This rearrangement can involve changes in the word order or the repositioning of phrases, ultimately resulting in a different syntactic structure while maintaining the overall meaning.

Lastly, we have included the category of omission, which involves the removal of specific words or phrases from a sentence without compromising the overall intended meaning. Omission-based paraphrase enables the reduction of linguistic content while retaining the essential semantic information within the sentence.

This analysis involves a comparative examination of the books NACMAR, ESPMAR, JUNMAR, and LATMAR, specifically focusing on the Gospel of Saint Mark, which is the shortest among the synoptic gospels. However, direct comparisons were conducted only between the NAC-ESP, JUN-LAT, and ESP-LAT pairs. Furthermore, for the sake of simplicity, the analysis is limited to the first chapter of the Gospel of Saint Mark. In this study, we present the results of a meticulous word-by-word analysis aimed at identifying and categorizing the types of paraphrase based on the aforementioned categories.

The initial step involves labeling the Bible verses where paraphrase phenomena occur and assigning them to one of the four categories: substitution, morphological modification, modification in the order of words and omission. The alignment process scrutinizes each lexical unit encompassing the first chapter of the Gospel of Saint Mark.

Table 6 illustrates the analysis of the first verse of Mark in the NAC and ESP versions (NACMAR001001 vs. ESPMAR001001). Each word or multi-word expression is allocated to a separate line in the table and assigned a corresponding label. The different columns represent the pair of versions being studied, facilitating a clear comparison between them. The last column indicates the identified category of difference, if any, between the compared versions.

We give now some examples of the four phenomena that can be distinguished in paraphrase, as stated above.

Table 6 Comparative analysis of the first verse of the Gospel of Mark, with the paraphrasing phenomena detected

	NACMAR	ESPMAR	Phenomenon
MAR001001_1	Principio *Beginning*	Orígenes *Origins*	Substitution
MAR001001_2	del *of the*	de la *of the*	Equal
MAR001001_3	Evangelio *Gospel*	Buena Noticia *Good News*	substitution
MAR001001_4	de *of*	de *of*	equal
MAR001001_5	Jesucristo *Jesus Christ*	Jesús, *Jesus*	Morphological modification
MAR001001_6		Mesías, *Messiah*	Omission
MAR001001_7	Hijo de Dios *Son of God*	Hijo de Dios *Son of God*	Equal

Omission: We use dashes (-) and brackets ([…]) for omitted text:

(1)

NACMAR001020	
Y—los llamó. [Ellos luego,] dejando a su padre Zebdeo en la barca con los jornaleros, se fueron en pos de Él	*And—he called them. [They then,] leaving their father Zebdeo in the boat with the laborers, went after him*
ESPMAR001020	
Y [en seguida] los llamó;—dejaron a su padre, Zebdeo, en la barca con los jornaleros y se marcharon con él	*And [immediately—they left their father, Zebdeo, in the boat with the laborers and went away with him*

Substitution: Both lexical units in the analysis share the same meaning. We use brackets ([…]) for substituted units.

(2)

ESPMAR001042	
[En seguida] se le quitó la lepra y quedó [limpio]	*[Right away] the leprosy was removed and he was [clean]*
LATMAR001042	
[Al instante] se le quitó la lepra y quedó [sano]	*[Instantly] the leprosy was removed and he was [healthy]*

Morphological modification

(3)

NACMAR001022	
Se maravillan de su doctrina, pues la enseñaba como quien [tiene autoridad], y no como los escribas	*They marvel at his doctrine, for he taught it as one who [has authority], and not as the scribes*
NACMAR001022	
Estaban asombrados de su enseñanza, porque enseñaba como quien [está autorizado], y no como los letrados	*They were amazed at his teaching, because he taught as one who [is authorized] and not as the scribes*

Modification in the order of the words: It is denoted by the inclusion of brackets ([…]) within the sentence, which serve to indicate the specific location where the modification occurs. The strategic use of brackets not only highlights the position of the modification but also aids in establishing semantic alignment between the sentences. This alignment is achieved by arranging the words in a manner that enables the correspondence of semantically related terms. For instance:

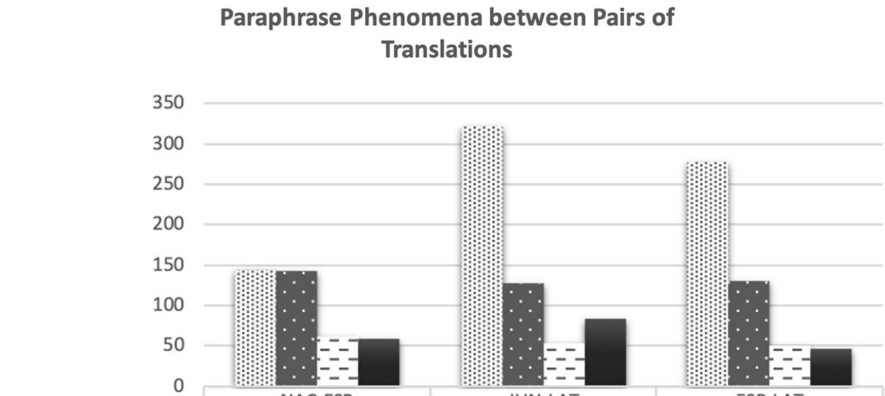
(4)

NACMAR001002	
He aquí que envío delante de ti [mi ángel], que preparará tu camino	*Behold, I send before you [my angel] who will prepare your way*
ESPMAR001002	
Mira, envío [mi mensajero] delante de ti, él preparará tu camino	*See, I send [my messenger] ahead of you, he will prepare your way*

The moved phrase in (4) shows not only modification in the order, but also lexical substitution.

4.1.1 Results

The forty-five verses comprising the first chapter of Mark have been aligned in pairs for comparison: ESP-NAC, JUN-LAT, and ESP-LAT. Figure 1 provides an overview of the most prevalent mechanism observed, which is omission. This indicates that certain phrases or expressions are not conveyed in one of the contrasted pairs. Omissions appear to be more frequent between JUN and LAT, suggesting substantial differences between these versions. Similar patterns of omission are also observed between ESP and LAT. It appears that LAT, being a translation aimed at maximizing comprehensibility, tends to provide additional explanations for certain concepts. As a result, LATMAR001 contains 860 words, while ESPMAR001 has 781, NACMAR001 has 764, and JUNMAR002 has 757. This disparity in word count



	NAC-ESP	JUN-LAT	ESP-LAT
Omission	144	322	277
Substitution	143	128	131
Morphological Mod.	61	54	50
Word Order Mod.	59	84	46

Fig. 1 Number of occurrences for each paraphrase phenomenon within the three pairs of translations analyzes for the first chapter of Marc

highlights the divergent approaches of the Latin American translations in terms of their intended audience. Additionally, they differ in their translation strategies, with JUN adopting a more literal approach and LAT opting for a more explanatory style.

Data in Fig. 1 reveals a total of 743 instances of omissions, 402 instances of substitutions, 165 instances of morphological modifications, and 190 instances of word order modifications. These numbers provide a quantitative overview of the prevalence of each paraphrase category across the translation pairs studied.

Substitution emerges as a consistently observed and widely represented phenomenon within each translation pair, encompassing the diverse choices made in employing distinct words to convey the same message in the four different Bibles that we mentioned before (Shoeckel and Mateos [ESP], Nácar-Colunga [NAC], Junemann [JUN], and the Latin American edition [LAT]).. On the other hand, morphological modification remains a relatively stable category across all versions examined. Generally, alterations in word inflection and tense are more prominently influenced by the narrative structure rather than substantial deviations in translation strategy.

Notably, JUN stands apart from the other translations in terms of word order. While the NAC-ESP and ESP-LAT pairs exhibit greater similarity in this aspect, JUN demonstrates the highest degree of deviation from the expected word order patterns. This disparity suggests a distinctive approach in JUN's translation methodology, potentially driven by its aim to maintain a closer adherence to the original Greek syntax.

By carefully analyzing the patterns of substitution, morphological modification, and word order, we gain valuable insights into the specific characteristics and divergences exhibited by each translation pair. This exploration contributes to a more comprehensive understanding of the nuanced dynamics underlying the translation choices made in rendering the Gospel of Mark across different versions.

4.2 Semantic clustering

Clustering has been applied to almost every discipline. The process of identifying clusters has variously been called cluster analysis, classification, categorization, taxonomy, typology, or clumping, according to the discipline. The primary goal of clustering is to collect together a set of elements associated by some common characteristic, in such a way that it is possible to cluster, and so on.

Semantic clustering has garnered significant attention within the field of Natural Language Processing (NLP) due to its relevance in various applications, including plagiarism detection, question answering, textual entailment, summarization, and automatic machine translation evaluation. The primary objective of semantic clustering is to identify pairs of words that can be interchangeably used in the same sentence without changing the meaning of the verses. Currently, word embeddings have emerged as the dominant approach for measuring word similarity. These embeddings are rooted in the distributional hypothesis (Harris, 1954), which posits that words with similar contextual usage tend to share similar meanings. However, the effectiveness of word embeddings in capturing semantic similarity relies on the availability of substantial volumes of text data.

While considerable research has been conducted on comparing long texts, the comparison of short texts presents a more challenging task. Han et al. (2013) classify methods for detecting and measuring similarity between short texts into three groups: 1) Vector space models, 2) Alignment of segments and computation of word pair similarity, and 3) Machine learning models that combine multiple measures and incorporate lexical, semantic, or syntactic features.

As an example of a segment alignment method, Sierra and McNaught (2000) developed a semantic clustering algorithm specifically designed to identify word pairs where one member can be replaced by the other in a definition without altering the underlying concept. This is achieved through the alignment of definitions expressing the same concept but utilizing different wording. Automatic alignment of parallel texts aims to determine the most probable correspondences between words in the target sentence and words in the source sentence. For instance, consider the following two verses in example (5):

(5)

ESPMAR005034	
Él le dijo: Hija, tu fe te ha curado. Vete en paz y sigue sana de tu tormento	*He said to her: Daughter, your faith has healed you. Go in peace and stay healthy from your torment*
LATMAR005034	
Jesús le dijo: Hija, tu fe te ha salvado. Vete en paz y queda sana de tu enfermedad	*Jesus said to her: Daughter, your faith has saved you. Go in peace and be healed of your illness*

Alignment serves the purpose of identifying word equivalences within the given verses. By observing the sentences, several word pairs can be identified, namely: (Él, Jesús), (curado, salvado), (sigue, queda), and (tormento, enfermedad). These pairs exhibit correspondence, indicating that the words can be substituted for each other without significantly altering the meaning.

Distribution-based clustering methods assume that the similarity of words can be judged by analyzing the similarity of the context in which they occur. Brown et al. (1992) use a mutual information measure in a window of 1,001 words, excluding the two words before and after the keyword, applied to large corpora. However, the alignment algorithm employed is not inherently statistical in nature, which means it does not rely on large amounts of data and can generate clusters even when word alignment is infrequent. The alignment process only compares the words in two verses sequentially, establishing correspondences between words that can replace each other in the verses without producing any major change in meaning. The objective is to identify the minimum cost associated with each operation required to transform one phrase into another. The operations are: substitutions of a word for another, insertion of a word into a string, and deletion of a word from a string. The Levenshtein distance (Levenshtein et al., 1966) is utilized to quantify this cost.

To achieve alignment, a dynamic programming method (Wagner and Fisher, 1974) is employed. This method enables the alignment of elements in two strings based on the Levenshtein distance calculation, resulting in ordered pairs representing the alignment. Each word pair is associated with a Levenshtein distance cost, as depicted in Table 7.

Experimental findings have demonstrated that using stem forms, obtained through the use of Freeling, yields superior matching results compared to utilizing full word forms. By utilizing the dynamic programming approach in conjunction with the Levenshtein distance, the alignment algorithm facilitates the analysis of word correspondences and the computation of associated costs.

The alignment process yields a list of triplets consisting of $(ff_i, ff_j, cost[i][j])$, where ff_i and ff_j represent the full forms of the strings S1 and S2, respectively. The objective of clustering is to establish matches between pairs of different words, such as "curado" and "salvado".

To measure the similarity between a matched couple, the algorithm quantifies the number of surrounding identical pairs above and below the matched pair. This concept is akin to the "longest common subsequence" proposed by Wagner and Fisher (1974) for comparing two strings, defined as the longest common subsequence between the two strings. In this case, the two strings differ only by the matched couple. Consequently, the algorithm introduces the concept of the "longest collocation couple" (lcc), which refers to the maximal sequence of word pairs composed of equal couples surrounding the matched couple.

Table 7 Alignment for MAR005034 in ESP and LAT

ESP	LAT	Cost
Él *He*	Jesús *Jesus*	1
le *to her*	le *to her*	1
dijo *said*	dijo *said*	1
Hija *daughter*	Hija *daughter*	1
tu *your*	tu *your*	1
fe *faith*	fe *faith*	1
te *you*	te *you*	1
ha *has*	ha *has*	1
curado *healed*	salvado *saved*	2
ve *go*	ve *go*	2
te *(you)*	te *(you)*	2
en *in*	en *in*	2
paz *peace*	paz *peace*	2
y *and*	y *and*	2
sigue *stay*	queda *be*	3
sana *healthy*	sana *healed*	3
de *from*	de *of*	3
tu *your*	tu *your*	3
tormento *torment*	enfermedad *illness*	4

The algorithm produces a new triplet (ff_i, ff_j, lcc_{ij}), where (ff_i, ff_j) represents the matched couple and lcc_{ij} denotes the length of the longest collocation couple. It is possible to find multiple *lcc* values for any given pair of strings. By ranking all the triplets based on *lcc* in descending order, it becomes evident that a higher lcc value corresponds to a greater similarity between the words in the matched pair.

The most promising clusters are typically found at higher *lcc* values. Experimental results indicate that a *lcc* length of 5 serves as a reliable threshold. While there may be valid matches with *lcc* values of 4 or 3, the majority of these tend to duplicate matches with higher *lcc* values.

4.2.1 Clustering for Mark

The alignment of Bible verses was conducted by cross-referencing the corresponding verses in each pair of Bibles. To illustrate this process, Table 8 presents the pairs with the highest lcc value achieved by aligning the book of Mark from the New Spanish Bible (ESP) and the Latin American Bible (LAT).

Each matched couple serves as an initial cluster, representing sets of words that are used interchangeably within specific contexts. In a consecutive sequence of matched couples, there may be instances where a stem form appears in multiple distinct bindings. In such cases, it is possible to form clusters by grouping all

the couples that share a common stem form, utilizing the transitive property, i.e., ff1 = ff3 if ff1 = ff2 and ff2 = ff3.

Table 9 presents a selection of semantic clusters derived from grouping the matched couples identified in the book of Mark across the ESP, JNM, NAC, and LAT Bibles. These clusters demonstrate the associations between words that exhibit semantic equivalence within the analyzed texts.

Upon comparison with the *Diccionario de sinónimos y antónimos* (Cortés et al., 2006) via WordReference.com, it becomes evident that certain semantic clusters align with the synonymous terms provided in the dictionary. For instance, nouns like *mente*, *inteligencia*, and *entendimiento*, as well as verbs like *enojar* and *indignar*, coincide with their corresponding synonyms in the dictionary. In other cases, semantic clusters consist of indirect synonyms. For instance, although *departamento* and *pieza* are not synonymous, they both share *habitación* as a synonym. By applying the transitive property, these words are considered to belong to the same cluster.

Furthermore, there are instances where semantic clusters comprise cohyponyms, such as *príncipe*, *jefe*, and *oficial* for nouns, or lexical entailment, such as *decir*, *preguntar*, *responder*, *contestar*, *asegurar*, and *contar* for verbs. These clusters demonstrate relationships of inclusion or entailment, where one word encompasses or entails the meaning of another within the same cluster.

4.3 Subjective bias

Most writings are biased by personal subjectivity, ideological criteria, or deeply rooted social prejudices, such as gender, race, and political tendencies (Pryzant et al., 2020).

Currently, there is a concern to achieve a language with a neutral point of view (NPOV). Bias detection studies also promote mitigation formulas to achieve neutral texts (Recasens et al., 2013).

The Bible, due to its intrinsic characteristics, does not have the goal of neutrality. Translators try to faithfully reflect the ideas of the original, but the versions they generate reveal their theological, political or philosophical ideas. Furthermore, the underlying translation theory also has an influence on the final result.

In our examples, we have taken books from the New Testament, written in Koine Greek. Two of the versions that we analyze, Biblia Latinoamericana (LAT) and Nueva Biblia Española (ESP), were criticized for their perceived bias. The Biblia Latinoamericana (LAT), in particular, was accused of going beyond translation or versioning and incorporating words that were closer to and more understandable for its intended audience. This version of the Bible became associated with the Liberation Theology movement advocated by figures such as Gustavo Gutiérrez, Helder Cámara, Oscar Romero, and Leonardo Boff.

Thomas (2021) highlights that the changes in the texts primarily focused on selected words comprehensible to the target audience. Nevertheless, the censors also targeted the illustrations in the initial version. In Argentina, in particular, it faced

Table 8 Semantic couples with high lcc

Vers	ESP	LAT	lcc
MAR012030	mente	inteligencia	23
MAR012033	entendimiento	inteligencia	22
MAR012014	departamento	pieza	20
MAR012036	diestra	derecha	19
MAR004041	decían	preguntaban	17
MAR016009	mañana	madrugada	16
MAR006007	impuros	malos	14
MAR009017	dijo	respondió	14
MAR010014	enojó	indignó	14
MAR005040	iban	venían	13
MAR006056	permitiera	dejara	13
MAR010021	luego	después	13
MAR014037	vino	volvió	13
MAR001016	lago	orilla	12
MAR004027	germina	brota	12
MAR005010	misericordia	compasión	12
MAR008006	tierra	suelo	12
MAR009004	hablaban	conversaban	12
MAR011027	príncipes	jefes	12
MAR014062	diestra	derecha	12
MAR016015	predicad	anuncien	12

persecution, with an ecclesiastical authority even instructing their congregation to destroy copies of the Bible, claiming that they were an insult to God.

LAT aimed to translate cultisms into a simpler and more accessible register.

In 1975, the Nueva Biblia Española (ESP) was introduced, also in alignment with the spirit of the Second Vatican Council, with the intention of providing a language that is more understandable for the audience. This version was developed under the supervision of Luis Alonso Schökel and Juan Mateos.

ESPreceived acclaim for its literary quality. Although it has not undergone the same level of scrutiny as the Biblia Latinoamericana, it has faced criticism, particularly regarding its New Testament texts. It was noted at the time that Mateos had developed his own exegetical method for studying the New Testament, based on philological and semantic analyses of each word, but that it contained several doctrinal imprecisions.

Detecting bias in natural language processing (NLP) presents a challenging task (Bruce & Wiebe, 1999; Recasens et al., 2013). This can involve removing bias components (Pryzant et al., 2020) or utilizing automatic text generation techniques (Dun et al., 2019).

Table 9 Semantic clusters from Mark

Mente, inteligencia, entendimiento	*Mind, intelligence, understanding*
Departamento, pieza	*Apartment, piece*
Diestro, derecho	*Right, right handed*
Decir, preguntar, responder, contestar, asegurar, contar	*Say, ask, answer, answer, ensure, tell*
Mañana, madrugada	*Morning, early morning*
Impuro, malo, inmundo	*Unclean, bad, unclean*
Enojar, indignar	*Anger, outrage*
Ir, venir, volver, regresar, llegar, bajar, pasar, acercar	*Go, come, come back, get down, pass, zoom in*
Permitir, dejar	*Allow, leave*
Luego, después	*Then, after*
Lago, orilla, mar	*lake, shore, sea*
Germinar, brotar	*germinate, sprout*
Misericordia compasión	*mercy compassion*
Tierra, suelo	*Ground*
Hablar, conversar	*Speak, talk*
Príncipe, jefe, oficial	*Prince, chief, official*
Predicar, anunciar	*Preach, announce*
Hombre, gente, nación, gentío, pueblo, aldea, multitud	*Man, people, nation, crowd, town, village, crowd*

4.3.1 Bias in paraphrase

In this section, we explore how the various phenomena observed in paraphrasing can exhibit bias based on the ideological and, in this case, theological stance of the translator. The act of substitution is particularly revealing of ideological, theological, and even philological divergences. Each translation reflects a distinct perspective, and this becomes evident within this category. For instance, consider the selection of words in contexts like the one illustrated in example (2): *clean* (ESPMAR001042, JUNMAR001042, NACMAR001042) versus *healthy* (LATMAR001042), while also taking into account that *saved* is used in other verses with a similar meaning. Another intriguing example is the use of the word *angel* (JUNMAR001002, NACMAR001002) versus *messenger* (ESPMAR001002, LATMAR001002) in example (4). This choice carries distinct theological and social implications, differentiating the pre-SVC translations from the post-SVC translations. ESP and LAT, in an attempt to narrate Jesus' birth in a less miraculous manner, utilize the fact that the Greek word ἄγγελος also means *messenger* to convey this translation to the reader.

Concerning morphological modifications, JUN stands apart from the other translations, likely due to its composition in the nineteenth century, although published in the twentieth century. As a consequence, JUN exhibits archaic usage of clitics. An example can be found in (6) for MAR001017. It demonstrates how LAT, ESP, and NAC employ the phrase *Jesús les dijo*, while JUN writes *Díjoles Jesús,* with the pronoun *les* in an out-fashioned position. Similarly, JUN employs the form *haréos*

[I will make you], whereas LAT uses *los haré* and ESP and NAC translate it as *os haré*, showcasing the differing placement of the pronoun.

(6)

JUNMAR001017	
Y díjoles Jesús: «Venid en pos de mí, y haréos ser pescadores de hombres»	*And Jesus said to them, "Follow me, and I will make you fishers of men."*
LATMAR001017	
Jesús les dijo: "Síganme y yo los haré pescadores de hombres"	*Jesus told them, "Follow me and I will make you fishers of men."*
ESPMAR001017	
Jesús les dijo:-Veníos conmigo y os haré pescadores de hombres	*Jesus said to them, "Come with me and I will make you fishers of men."*
NACMAR001017	
Y Jesús les dijo: Venid en pos de mí y os haré pescadores de hombres	*And Jesus said to them: Come after me and I will make you fishers of men*

Finally, word order is used to emphasize some part of the verse. Let's compare. MAR001006 in the four translations in example (7):

(7)

NACMAR001006	
Llevaba Juan un vestido de pelos de camello, y un cinturón de cuero ceñía sus lomos, y se alimentaba de langostas y miel silvestre	*Wore John a camel hair dress, and a leather belt girded his loins, and he fed on locusts and wild honey.* (Google translation)
ESPMAR001006	
Juan iba vestido de pelo de camello, con una correa de cuero a la cintura, y comía saltamontes y miel silvestre	*Juan was dressed in camel hair, with a leather belt around his waist, and he ate grasshoppers and wild honey.* (Google translation)
JUNMAR001006	
Y estaba Juan vestido de pelos de camello y ceñidor de cuero en torno de su cintura, y comiendo langostas y miel silvestre	*And there was Juan dressed in camel hair and a leather girdle around his waist, and eating locusts and wild honey.* (Google translation)
LATMAR001006	
Además de la piel que tenía colgada de la cintura, Juan no llevaba más que un manto hecho de pelo de camello. Su comida eran langostas y miel silvestre	*Apart from the skin that hung from his waist, Juan wore nothing but a cloak made of camel hair. His food was locusts and wild honey.* (Google translation)

Only ESP uses the canonical order of words in Spanish: SVO (Subject, Verb, Object). In contrast, NAC and JUN start the sentence with the verb, something that is recurrent in both translations, and gives the story an epic and archaic tone. However both NAC and JUN follow the exact order of the Greek sentence: καὶ ἦν ὁ Ἰωάννης ἐνδεδυμένος τρίχας καμήλου καὶ ζώνην δερματίνην περὶ τὴν ὀσφὺν

αὐτοῦ, καὶ ἐσθίων ἀκρίδας καὶ μέλι ἄγριον (*And was John clothed with camel's hair and a leather belt around his loins, and grasshoppers and wild honey*).

Finally, LAT opens the sentence with the focus over John dressed with *FUR*, which highlights he was poor and austere. This can make him closer to the potential latin american readers to whom the translation is addressed.

4.3.2 Bias in Semantic Clustering

Through the alignment of the corpus, our semantic clustering analysis lets us identify pairs of words that can be used indistinguishably without changing the meaning of the sentence. Now, using straightforward methods, we are equipped to detect bias within the previously identified pairs of words. To accomplish this, we employ distributional semantics, a valuable tool for examining differences in meaning and similarity. Distributional semantics not only aids in studying connotations and meanings but also reveals how words specialize within specific contexts. By comparing the most commonly paired verbs with *gente* (people) and *turba* (crowd), it becomes readily apparent to infer the origin of the negative context.

Figure 2 shows how the word g*ente* [people] has a positive polarity while *turba* [mob] is mainly used when what is going to be said about the people is negative.

The following example (8) shows the word selection in the four versions—ESP, LAT, NAC, JUN—of MAR014043.

(8)

ESPMAR014043	
Aún estaba hablando cuando se presentó Judas, uno de los Doce, acompañado de una *turba* con machetes y palos, de parte de los sumos sacerdotes, los letrados y los senadores	*He was still speaking when Judas, one of the Twelve, appeared, accompanied by a mob with machetes and sticks, made up of high priests, lawyers, and senators.* (Google translation)
LATMAR014043	
En aquel instante, cuando aún estaba Él hablando, llegó Judas, uno de los Doce, y con él un *tropel* con espadas y garrotes, de parte de los escribas y de los ancianos	*At that moment, while He was still speaking, Judas, one of the Twelve, arrived, and with him a troop with swords and clubs, from the scribes and the elders.* (Google translation)
NACMAR014043	
Jesús estaba aún hablando cuando se presentó Judas, uno de los Doce; lo acompañaba *un buen grupo de gente* con espadas y palos, enviados por los jefes de los sacerdotes, los maestros de la Ley y los jefes judíos	*Jesus was still speaking when Judas, one of the Twelve, appeared; He was accompanied by a good group of people with swords and sticks, sent by the chief priests, the teachers of the Law, and the Jewish chiefs.* (Google translation)
JUNMAR014043	
Y al punto, aún hablando él, llégase Judas uno de los doce, y, con él, una *turba* con cuchillas y palos, de los sumos sacerdotes, y los escribas y los ancianos	*And immediately, while he was still speaking, Judas one of the twelve arrived, and, with him, a mob with knives and sticks, of the high priests, and the scribes and the elders*

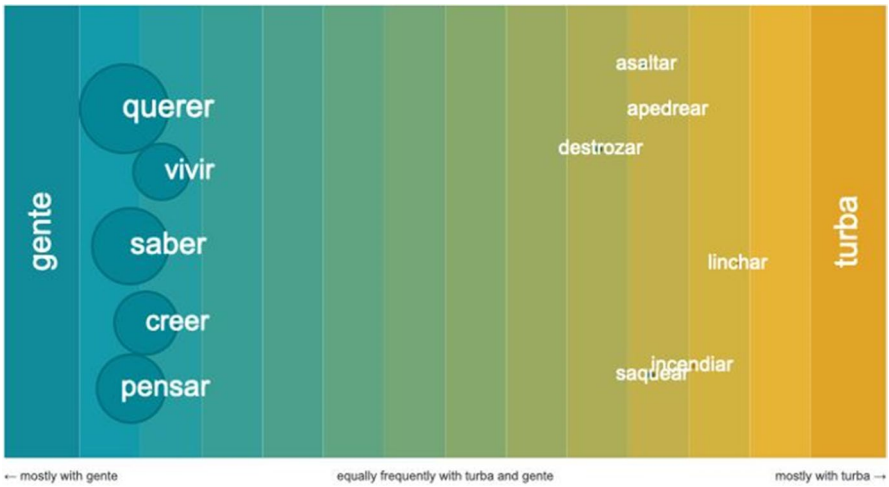


Fig. 2 Words frequently associated with *gente* and *turba*

In the case of NAC, the translation incorporates the word *tropel*, which carries the same meaning as *turba*. However, *tropel* has less negative connotations. The translators aim to depict the negative nature of this group of people while simultaneously maintaining a certain degree of neutrality in the scene.

LAT, on the other hand, employs a periphrasis, a multi-word expression, instead of using the highly negative term *turba*. This choice may be attributed to the historical association of *turba* with the working class and anarchist movements, among others.

In contrast, ESP and JUN both use the word *turba*, likely for different reasons. These two translations exhibit the least similarity. ESP perhaps intends to vividly express the intentions of the individuals depicted in the scene. JUN, on the other hand, provides a literal translation of the Greek word ὄχλος, which refers to the populace or a mob.

5 Conclusions and future work

This article presents an aligned parallel corpus consisting of translated versions of the Bible in Spanish. The corpus comprises eleven Bibles, originating from various centuries (XVI, XIX, XX), different religious denominations (Protestant, Catholic), and geographical regions (Spain, Latin America).

The translations have been meticulously aligned to ensure semantic equivalence across all versions at the verse level. The primary objective of this resource is to facilitate Natural Language Processing tasks that necessitate semantically aligned data, such as paraphrase detection, machine translation, semantic clustering, bias identification, and the study of geographic and chronological variations. The scarcity

of such resources in the Spanish language often limits computational experiments in these areas, which we aim to address through this work.

Moreover, we present analyses conducted using the corpus along with preliminary findings. Specifically, we focus on paraphrase detection, semantic clustering, and bias identification. To illustrate these analyses, we provide examples from the Gospel of Mark in four Bible versions chosen to represent distinct geographical regions and ideological perspectives: Spain versus Latin America, before and after the Second Vatican Council (SVC). The meticulous alignment of the texts facilitates qualitative analysis of the quantitative data extracted.

In future research, our objectives include advancing automatic paraphrase detection and conducting a more in-depth analysis of the ideological stances underlying each translation. Additionally, we plan to do research on some aspects of linguistic change, both from a geographical (dialectal) and diachronic (linguistic change) point of view. Especially with reference to bias and paraphrasing, further research is needed to distinguish whether some of the changes observed are due to ideological/theological or rather dialectal differences.

We consider the Bible a valuable resource for such investigations due to the extensive research on its authorship, the exegetical and philological work conducted on the text, and the extensive documentation available on the translation perspectives adopted in each version.

Annex 1

	AME	EMN	ESP	JER	JNM	LAT	NAC	OSO	PET	REV	SER
1CO	9439	9867	9348	9054	8263	9684	8878	9142	10693	9111	9399
1CR	20034	19969	16723	1847	1895	18906	18713	1974	20996	19265	18999
1JN	2559	2504	2506	2459	238	2505	2393	2533	2661	2503	2441
1MC		2141		20942	20934	2055					
1PD	2503	2564	2419	2355	2142	2579	2268	2507	2740	2408	2350
1RY	23476	23507	19844	21352	24724	21066	21949	23756	24608	23414	22420
1SM	23815	23386	20454	22010	24069	21727	21905	24020	25705	23682	23378
1TL	1974	1920	1798	1848	1794	1892	1799	1885	2043	1858	1954
1TM	2472	2508	2369	2460	2024	2537	2277	2345	2574	2295	2494
2CO	6175	6373	5965	5850	5411	6113	564	5914	703	5896	6013
2CR	24946	24840	21404	23863	24211	23982	23640	24957	26259	24777	24742
2JN	283	316	304	300	284	291	289	301	315	300	301
2MC		15887		15376	14623	15201					
2PD	1534	1616	1490	1498	1373	1575	1486	1599	1714	1535	1479
2RY	22498	22607	18883	20587	21514	20767	20843	22363	23394	22126	21704
2SM	19667	19727	16779	18428	19801	17679	18550	19744	20980	19578	19167
2TL	1068	1070	1043	1054	956	1086	995	1017	1158	1033	1070
2TM	1704	1804	1680	1714	1481	1687	1571	1656	1808	1628	
3JN	309	324	301	275	256	325	276	306	347	311	280

	AME	EMN	ESP	JER	JNM	LAT	NAC	OSO	PET	REV	SER
ABD	576	583	460	579	559	581	573	586	708	587	540
AMS	3804	3824	3265	3677	3727	3854	3663	3766	4435	3744	3551
APC	1165	11300	10814	11173	11096	11142	11276	11965	12220	11665	11232
BAR		3473		48	4542	4613					
CNT		2511	2383	2431	2377	2476	2416	2542	2960	2527	2333
COL	2148	2168	1985	1959	1895	2138	1963	2035	2279	2039	2102
DAN	11118	11113	9876	10490	12394	10017	10268	11084	13346	10940	10407
DET	26125	25748	22885	25084	25003	24556	24606	25795	26779	25350	25413
ECL	5419	5384	4646	5012	5265	4996	5087	5501	6124	5461	4863
ECS		26714		25030	25305	25338					
EFS	3207	3241	3039	3046	2921	3175	2990	3129	3498	3097	3173
ESD	7123	6854	5690	6585	6665	6454	6679	7043	7368	6981	6948
EST	5627	5551	4794	5296	6735	5433	5356	5616	6182	5530	8046
EXD	30805	31019	25923	28562	28064	28655	27831	30294	30427	29618	28871
EZQ	35638	35813	30401	34533	34163	31467	33253	35573	38872	35600	34433
FIL	2251	2346	2226	2194	1966	2276	2103	2150	2470	2184	2310
FLM	460	493	427	448	412	446	393	446	538	425	463
GAL	3193	3231	3191	3144	2765	3351	2982	3111	3519		3190
GEN	37431	36688	31423	34988	34354	36385	33704	37073	37281	36525	35334
HAB	1355	1381	1144	1312	1288	1360	1346	1363	1585	1366	1240
HAG	1038	1070	852	1049	1052	962	994	1071	1144	1038	1011
HBR	7027	6770	6682	6780	6191	7152	6781	6827	7984	6835	7124
HCH	23585	23330	21689	22656	21218	23651	21826	22973	24847	22883	22846
ISA	34063	34262	29688	32057	31537	32755	32389	33414	39701	33156	31064
JCS	17995	17483	15236	16930	17717	16472	16852	17984	18733	17795	17318
JDT		10246		10065	10294	8781					
JER	39382	39788	32843	37972	35478	37123	37199	39173	43959	38630	37776
JOB	17187	17927	15082	16454	17190	16217	16594	17361	20781	17210	15578
JOL	1833	1884	1541	1782	1832	1748	1771	1803	2061	1776	1294
JON	1242	1224	1053	1180	1208	1114	1100	1233	1301	1219	1188
JOS	18119	17921	15009	17054	18001	15412	16812	18178	18072	17663	17276
JUD	616	676	631	620	541	691	585	643	720	600	607
JUN		17913	18635	18011	17375	18455	17546	18367	19592	18259	18565
LAM	3296	3247	2879	3172	2977	2997	3328	3170	3811	3100	3083
LEV	22677	23011	18484	20654	20274	20678	19535	21722	17132	21321	21049
LUC	24665	23734	23018	23141	22041	24774	22630	24148	26036	23986	24360
MAL	1704	1689	1426	1628	1710	1735	1638	1655	1905	1660	1434
MAR	14235	13717	12657	13539	12649	14431	13104	14039	15108	13984	14115
MAT	22692	22119	20863	21802	20639	23207	21098	22445	23954	22232	22310
MIQ	2786	2811	2310	2688	2776	2814	2709	2723	3326	2697	2654
NAH	1140	1158	990	1109	1108	983	1128	115	1377	1108	974
NHM	7763	9857	8297	9339	9560	9176	9462	9954	10703	9880	9737
NUM	30963	30286	25673	27859	28256	28597	28049	30080	29452	29710	29524

	AME	EMN	ESP	JER	JNM	LAT	NAC	OSO	PET	REV	SER
OSE	4721	4803	3977	4450	4454	4852	4474	4564	5728	4552	4288
PRV	14881	14640	12857	14258	14612	14361	14387	14777	17311	14801	13300
ROM	9899	10076	9641	9459	8695	9977	9283	9535	11381	9521	10281
RUT	2416	2313	2116	2299	2281	2230	2243	2429	2551	2381	2470
SAB		10267		9794	8947	9955					
SAL	42172	41704	35988	38167	38385	36348	39582	40767	47187	40652	38714
SNT	2356	2365	2319	2257	2063	2343	2165	2319	2592	2293	2210
SOF	1464	1541	1235	1443	1448	1535	1428	1470	1637	1435	1375
TBS		8484		8388	6172	6733					
TIT	941	1061	993	935	831	1023	935	930	1058	930	997
ZAC	5837	5828	4800	5492	5696	5547	5484	5899	6395	5769	5319

Author contributions Conceptualization: GS, GBE, NB; Methodology: GS, GBE, NB; Formal analysis and investigation: ADV, NGC; Writing—original draft preparation: GS, GBE, NB, ADN, NGC; Writing—review and editing: GS, GBE, NB; Funding acquisition: GS, GB; Resources: GS, GBE, ADV, NGC; Supervision: GS, GBE.

Funding This research has been funded by PASPA-DGAPA-UNAM and Conacyt, grant CF-2023-G-64 and Conacyt CB A1-S-27780.

Data and material availability https://github.com/GIL-UNAM/Alineamiento-Biblias, https://github.com/GIL-UNAM/SpanishParaphraseCorpora/tree/main/Biblias, http://www.geco.unam.mx/geco3/proyecto/CPBE

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Barrón-Cedeño, A., Vila, M., & Rosso, P. (2010). Detección automática de plagio: de la copia exacta a la paráfrasis. Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. Jornadas (in) formativas de lingüística forense, pages 76–96.

Bhagat, R., & Hovy, E. (2013). What is a paraphrase? *Computational Linguistics, 39*(3), 463–472.

Brown, P. F., de Souza, P. V., Mercer, R. L., Delia Pietra, V. J., & Lai, J. C. (1992). Class-based N-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.

Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering, 5*(2), 187–205.

Burrows, S., Potthast, M., & Stein, B. (2013). Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(3), 1–21.

Castro, B., Sierra, G., Torres-Moreno, J. M., & da Cunha, I. (2011). El discurso y la semántica como recursos para la detección de similitud textual. Anais do III Workshop "A RST e os Estudos do Texto", Cuibá, Brasil, Octubre 24–26.

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation, 49*(2), 375–395.

Cortés, P., Gallardo, A., & Grande, P. (2006). *Diccionario de Sinónimos y Antónimos*. Espasa Calpe.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*.

Cysouw, M., & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals, 60*(2), 95–99. https://doi.org/10.1524/stuf.2007.60.2.95

Das, D. & Smith, N.A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 468–476), Suntec, Singapore, August. Association for Computational Linguistics.

de Vries, L. (2007). Some remarks on the use of Bible translations as parallel texts in linguistic research. *Language Typology and Universals, 60*(2), 148–157. https://doi.org/10.1524/stuf.2007.60.2.148

Dong, L., Mallinson, J., Reddy, S., & Lapata, M. (2017). Learning to paraphrase for question answering. ArXiv, abs/1708.06022.

Dun, P., Zhu, L., & Zhao, D. (2019). Extending answer prediction for deep bi-directional transformers. In *32nd Conference on Neural Information Processing Systems (NIPS)*.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O., & Jøhndal, M. (2018). The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources & Evaluation, 52*, 29–65.

Enrique-Arias, A., & Pueyo Mena, F. (2008). Biblia medieval. Retrieved August 26, 2022, from http://corpus.bibliamedieval.es/.

Farwell, D., Dorr, B., Green, R., Habash, N., Helmreich, S., Hovy, E., Levin, L., Miller, K., Mitamura, T., Rambow, O., et al. (2009). Interlingual annotation of multilingual text corpora and framenet. *Multilingual FrameNets in Computational Lexicography: Methods and Applications, 200*, 287.

Gómez-Adorno, H., Bel-Enguix, G., Sierra, G., Torres-Moreno, J.-M., Martínez, R., & Serrano, P. (2020). Evaluation of similarity measures in a benchmark for Spanish paraphrasing detection. In *Advances in Computational Intelligence. MICAI 2020. Lecture Notes in Computer Science, volume 12469*. Springer.

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., & Weese, J. (2013). Umbc ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 44–52).

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162.

Herrera, J., Penas, A., & Verdejo, F. (2007). Paraphrase extraction from validated question answering corpora in Spanish. *Procesamiento del Lenguaje Natural, 39*, 37–44.

Hirst, G. (2003). Paraphrasing paraphrased. In *Keynote address for The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*.

Klein, W. W., Blombert, C. L., & Hubbard, R. L. (2017). *Introduction to Biblical interpretation* (3rd ed.). Zondervan.

Lavidas, N. (2022). *The Diachrony of written language contact. A contrastive approach*. Brill.

Lawrenz, M. (2014). *How to understand the bible: A simple guide* (Kindle). WordWay.

Levenshtein, V. I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Mayer, T. & Cysouw, M. (2014). Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3158–3163). Reykjavik, Iceland, May. European Language Resources Association (ELRA).

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., & Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 2884–2892). Marseille, France, May. European Language Resources Association.

Mota-Montoya, M., Da Cunha, I., & Lopez-Escobedo, F. (2016). Un corpus de paráfrasis en español: metodología, elaboración y análisis. *RLA. Revista de lingüística teórica y aplicada, 54*(2), 85–112.
Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *In Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 480–489.
Queralt, S., Zarauza, M. M., & García, R. G. (2018). Evidencias lingüísticas del plagio en el periodismo español. *Estudios Sobre el Mensaje Periodístico, 24*(2), 1559.
Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers, pp. 1650–1659).
Resnik, P., Olsen, M. B., & Diab, M. (1999). The bible as a parallel corpus: Annotating the "book of 2000 tongues". *Computers and the Humanities, 33*(1/2), 129–153.
Sierra, G., & McNaught, J. (2000). Extracting semantic clusters from MRDs for an onomasiological search dictionary. *International Journal of Lexicography, 13*(4), 264–286.
Sierra, G., Solorzano, J., & Curiel, A. (2017). GECO a web-based collaborative corpus manager. *Linguamatica, 9*(2), 57–72. https://doi.org/10.21814/lm.9.2.256
Thomas, H. (2021). *In the way of the story: Reading biblical narrative*. Wipf and Stock Publishers.
Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *Language Typology and Universals, 60*(2), 118–134. https://doi.org/10.1524/stuf.2007.60.2.118
Xu, W., Callison-Burch, C., & Dolan, B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 1–11).
Zhou, L., Lin, C.-Y., Munteanu, D. S., & Hovy, E. (2006). ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 447–454). New York City, USA, June. Association for Computational Linguistics.

Bibles of the Corpus

Escuela Bíblica de Jerusalén (Trads.) (1956/1967). *La Biblia de Jerusalén*. París: Desclée de Brouwer.
Jünemann, W. (1928). *La Sagrada Biblia*. Editorial Diocesana de Concepción.
La Biblia. (1975). España: Herder Editorial.
La Biblia de las Américas. (1986) EE.UU: Fundación Lockmann.
Reina, C. de (Trad.) *La Biblia, que es, los Sacros Libros del Viejo y Nuevo Testamento.* (1569) Berna.
Ricchiardi, R., Hurault, B., & (Trads.),. (1972). *La Nueva Biblia - Edición pastoral para Latinoamérica*. Editorial San Pablo y Verbo Divino.
Schökel, L., & Mateos, J. (Trads.) (1975) *Nueva Biblia Española*. España: Ediciones Cristiandad.
Martín. E. (Trad). *La Santa Biblia.* (1961). Madrid: San Pablo.
Nácar, E. y Colunga, A. (Trads.) (1944). *La Sagrada Biblia.* España: Editorial Católica.
Torres Amat, F. (Trad). *La Sagrada Biblia.* (1825). Barcelona.
Valera, C. (Trad.) (1602/1960) *La Santa Biblia.* Amsterdam: Editorial Sociedades Bíblicas Unidas.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.