UNIVERSITÀ DEGLI STUDI DI BERGAMO

UNIVERSITÀ DEGLI STUDI DI PAVIA

Scuola di Alta Formazione Dottorale

Corso di Dottorato in Scienze Linguistiche

CICLO XXXVII

The Encoding of Semantic Roles with Transfer and Communication Verbs
Insights from a Multilingual Parallel Corpus

Tutor:
Prof.ssa Silvia Luraghi

Candidato:
Luca Brigada Villa

Table of Contents

List of Figures.....v

List of Tables.....vii

List of Abbreviations.....viii

1 Introduction.....1

1.1 Overview.....1

1.1.1 Defining semantic roles.....4

1.1.2 The importance of a cross-linguistic perspective.....7

1.1.3 Semantic roles encoding in languages.....10

1.2 Semantic roles annotation.....12

1.2.1 Manual annotation efforts.....13

1.2.2 Automatic Semantic Role Labeling (SRL).....16

1.3 Quantitative methods to analyze semantic roles.....19

1.3.1 Semantic maps.....21

1.4 Goal of this work.....24

1.4.1 Transfer verbs.....25

1.4.2 Communication verbs.....26

2 Data and Methods.....28

2.1 Data collection and preprocessing.....29

2.1.1 Data preprocessing.....33

2.1.2 Sampling....35
2.2 Morphological and syntactic annotation....37
2.2.1 Dependency Grammar....38
2.2.1.1 Universal Dependencies....41
2.2.2 Parsing operations....51
2.3 Alignment of the texts....55
2.3.1 Alignment algorithm....58
2.4 Semantic roles extraction....69
2.4.1 Porting of semantic roles annotation....71
2.4.2 Integration of semantic roles into the morphosyntactically annotated resource....73
2.5 Analysis....77
3 Results and Discussion....79
3.1 Construction of the results table....79
3.1.1 Distance matrix....85
3.2 Clustering of the occurrences....87
3.2.1 Analysis of the clustering....95
3.2.2 Concluding notes on the clustering analysis....123
3.3 Semantic maps....127
3.3.1 Determining the optimal number of dimensions....128
3.3.2 Plots and interpretation of dimensions....132

4 Conclusions.....140

4.1 Summary of findings.....140

4.2 Implications of the study.....143

4.3 Limitations of the study.....146

4.4 Directions for future research.....149

References.....152

Appendix A Additional figures.....167

List of Figures

Figure 1. Semantic map from Haspelmath (1997).....30

Figure 2. Dunn Index for different clustering methods and numbers of clusters.....100

Figure 3. Davies-Bouldin Index for different clustering methods and numbers of clusters.....101

Figure 4. Distribution of the roles in three clusters (hierarchical clustering). 104

Figure 5. Distribution of the roles in four clusters (K-Means).....112

Figure 6. Distribution of the roles in four clusters (PAM).....113

Figure 7. Distribution of the roles in five clusters (K-Means).....119

Figure 8. Distribution of the roles in five clusters (PAM).....122

Figure 9. Distribution of the roles in six clusters (K-Means).....124

Figure 10. Distribution of the roles in six clusters (PAM).....125

Figure 11. Clusters' entropy variation (K-Means).....134

Figure 12. Clusters' entropy variation (PAM).....135

Figure 13. Stress values and random stress values for all the possible dimensions.....138

Figure 14. Differences in stress values between consecutive dimensions..139

Figure 15. Percentage difference in stress values.....140

Figure 16. Distribution of the occurrences using V1 and V2 as dimensions. 142

Figure 17. Distribution of the occurrences using V1 and V3 as dimensions. 145

Figure 18. Distribution of the occurrences (Destination and Time) using V1 and V3 as dimensions..... 146

Figure 19. Distribution of the roles in three clusters (K-Means)..... 176

Figure 20. Distribution of the roles in three clusters (PAM)..... 177

Figure 21. Distribution of the roles in four clusters (hierarchical clustering). 178

List of Tables

Table 1. List of languages in the initial sample.....39

Table 2. Final language sample.....46

Table 3. CoNLL-U fields and their description.....52

Table 4. Stanza's performances by task.....61

Table 5. Example of optimal path.....73

Table 6. Example of an alternative path.....73

Table 7. Example of another alternative path.....74

Table 8. Example of execution of the algorithm to convert the movements list into a sentence mapping.....76

Table 9. Results table.....89

Table 10. Normalized entropies of the three clusterings (3 clusters).....111

Table 11. Normalized entropies of the two clusterings (4 clusters).....118

Table 12. Normalized entropies of the two clusterings (5 clusters).....123

Table 13. Occurrences of communication and transfer verbs in the clusters. 129

List of Abbreviations

1	first person
2	second person
3	third person
ABS	absolutive
ACC	accusative
ACT	active
ALL	allative
AOR	aorist
AUX	auxiliary
COM	comitative
COND	conditional
DAT	dative
F	feminine
FUT	future
GEN	genitive
IMP	imperative
INF	infinitive
INS	instrumental
IPFV	imperfective
LOC	locative

M	masculine
MID	middle
MOD	modal
NEG	negation
NOM	nominative
PASS	passive
PRF	perfect
PFV	perfective
PL	plural
POSS	possessive
POT	potential mood
PRS	present
PTCP	participle
REFL	reflexive
REL	relative
SG	singular
SBJ	subject
TOP	topic
VOC	vocative

1 Introduction

1.1 Overview

Semantic role theory traces its origins to the mid-20th century, evolving significantly through various linguistic paradigms. The early conceptualization of semantic roles, or thematic roles, grew out of efforts to understand the deeper meaning conveyed by syntactic structures in sentences. This development has provided linguists with tools to describe how participants in relate to the main action or event described by the verb. The formal study of semantic roles is often attributed to Charles Fillmore, whose introduction of Case Grammar in the 1960s marked a seminal moment in linguistic theory. Fillmore (1968) proposed that the syntactic structure of a sentence reflects underlying semantic relationships, which he termed “deep cases”. These cases were intended to capture the roles that sentence constituents play in the context of the action or state denoted by the verb. Fillmore identified several universal semantic roles, including Agent,

Instrument and Experiencer, arguing that these roles are central to understanding meaning across languages. Fillmore's case roles were directly tied to deep semantic functions, suggesting that languages universally utilize a set of semantic templates to structure meaning. His theory posited that understanding these templates could reveal much about the cognitive processes that underpin language use.

Following Fillmore's foundational work, semantic role theory was further elaborated by other linguists who introduced additional frameworks and refined the concept of semantic roles. Among these developments was the theory of "thematic roles" developed in the late 1970s and 1980s, which provided a slightly different perspective, more focused on the verb-centered organization of roles within sentences. These thematic roles, often including Agent, Theme, and Goal, were seen as crucial for processing the syntactic and semantic information of verbs.

David Dowty's argument structure theory in the 1980s further expanded on these ideas by linking semantic roles to syntactic configurations. Dowty introduced the concept of "proto-roles", which blend semantic and syntactic features to describe the typical properties of agents and patients (Dowty, 1991). This approach provided a more nuanced understanding of how roles could be dynamically assigned based on the

cumulative weight of semantic features associated with a particular linguistic expression.

Another significant advancement in semantic role theory came with the development of Role and Reference Grammar (RRG) by Robert Van Valin in the early 1990s. RRG offered a comprehensive framework that integrated the syntax, semantics, and pragmatics, placing a strong emphasis on the role of semantics in syntactic structure. RRG distinguished between “actor” and “undergoer” roles as part of its layered structure of the clause, aiming to universally account for the syntax-semantics interface across languages (Van Valin, 2005).

Further enriching the landscape of semantic role theory, Fillmore's own later work on Frame Semantics expanded the scope of semantic roles within a broader cognitive and experiential framework. Introduced in the 1980s, Frame Semantics revolves around the idea that words evoke certain “frames” or schemata that structure understanding. In this framework, semantic roles are defined relative to these frames, which encompass a range of events (Fillmore, 1982). This approach has significantly influenced computational linguistics and the development of resources like FrameNet, which maps lexical units to semantic frames, demonstrating the enduring

impact of semantic role theory on practical applications in language technology.

1.1.1 Defining semantic roles

Semantic roles are abstract concepts that encapsulate the relationship of the verb to its participants, specifying the function each participant plays in relation to the verb's action or state. These roles, along with adverbial semantic roles, shape not only the grammatical structure but also the semantic interpretation of the sentence. In the sentence (1):

(1) Alice gave Mark a book.

we can say that 'Alice' is the 'Agent' (the giver), 'Mark' the 'Recipient' (the receiver) and the 'book' the 'Theme' (the item being given). Each of these roles serves specific functions that are crucial for the complete understanding of the sentence's meaning.

A persistent challenge in the study of semantic roles is establishing a universally accepted set of roles. The diversity of strategies of semantic roles encoding across languages often reflects deep grammatical and conceptual differences, complicating efforts to create a one-size-fits-all approach. The debate over the universality of semantic roles is deeply tied to the distinction between microroles and macroroles, concepts that offer

different levels of abstraction and generalization in the representation of participants' functions in the events indicated by verbs.

Microroles are highly specific to particular verbs and closely tied to the semantic nuances of those verbs. These roles reflect the fine-grained semantic distinctions that individual verbs can encode about their participants. For instance, the verb “to give”, as in example (1) involves roles such as ‘giver’, ‘recipient’, and ‘the item being gived’, which are specific to the action of giving. In contrast, the verb “to promise” involves different roles such as ‘promiser’, ‘promisee’, and ‘promise content’. These roles are verb-specific, because they arise from the inherent lexical semantics of the verb, dictating the types of arguments that the verb can logically take and the relationships among those arguments.

The specificity of microroles means that they can vary significantly even among closely related languages, as they are sensitive to the subtle semantic distinctions encoded by different verbs within the lexicon of each language. This variability presents challenges for establishing a standardized set of roles, as microroles by their nature resist generalization across verbs and languages.

In contrast to the high specificity of microroles, macroroles are broader categories designed to generalize across different verbs and, ideally,

across languages. Macroroles aim to reduce the complexity of microroles into more manageable and universal categories. Two commonly referenced macroroles in linguistic theory are ‘Actor’ and ‘Undergoer’ (Van Valin, 2005). Typically, the Actor macrorole is assigned to the participant who initiates or controls the action. This role includes various roles like agent, causer, and experiencer, which share the common feature of being in control of the action, regardless of the specific action described. The Undergoer macrorole refers to the participant who undergoes the effect of the action. This macrorole can include roles such as patient, theme, recipient, and target, which, despite their differences in specific verb contexts, all share the characteristic of being affected by the action.

By abstracting away from the fine details of microroles, macroroles offer a more universal framework that can accommodate the diversity of grammatical encoding across languages (Daniel, 2014, p. 207; Luraghi & Narrog, 2014, p. 2). Despite their utility, their application is not without challenges. One major issue is the extent to which macroroles can capture the full range of participant roles in all languages without oversimplification. Additionally, the imposition of a macrorole framework may sometimes force a fit where the natural linguistic data resist such

categorization, especially in languages with highly idiosyncratic case or argument structures.

1.1.2 The importance of a cross-linguistic perspective

After exploring the utility of both microroles and macroroles, it becomes evident that a comprehensive understanding of semantic roles necessitates examining how these roles are represented and function across different languages. The cross-linguistic perspective in linguistic research is invaluable for developing a comprehensive understanding of language, transcending the boundaries of individual languages to explore universal patterns and unique linguistic features across different language families. In the context of this work, which focuses on how semantic roles are encoded across languages, adopting a cross-linguistic perspective is crucial for several reasons: enhancing the generalizability of linguistic theories, identifying language-specific peculiarities, and contributing to the advancement of computational linguistic tools.

One of the primary benefits of a cross-linguistic perspective is the ability to test the universality of linguistic theories. By examining how different languages encode semantic roles, researchers can determine whether certain linguistic phenomena are universally applicable or if they

are specific to certain language environments. This broadens the scope of existing linguistic theories, providing a more robust framework that accounts for a diverse range of linguistic structures.

Cross-linguistic analysis also plays a pivotal role in uncovering language-specific features that may not be apparent when studying a single language. This aspect is particularly relevant in typological studies, where languages are compared to find typological universals and features. By statistically analyzing how different languages handle semantic roles, this work can uncover patterns of similarities and deviations that highlight the strategies languages employ to communicate similar semantic information. Such findings are essential for linguistic typology, as they contribute to our understanding of language diversity and complexity.

From a practical perspective, the cross-linguistic study of semantic roles is integral to the development of advanced computational linguistic tools, such as machine translation systems and multilingual semantic parsers. Most computational tools are developed and trained primarily on data from widely spoken languages, such as English. This often leads to less accurate tools for languages with fewer resources. By incorporating findings from a cross-linguistic analysis, developers can define more linguistic aware tagsets for the annotation of semantic roles and create more nuanced

algorithms that better handle the linguistic variability found across languages.

It is clear that a cross-linguistic perspective not only enhances the theoretical robustness of linguistic studies but also significantly impacts broader linguistic research. The necessity to define semantic roles based on empirical evidence from corpus data becomes increasingly important as we seek to refine the accuracy and utility of linguistic models, especially in determining the appropriate set and granularity of semantic roles. Corpus data from multiple languages provide a rich source of empirical evidence that is critical for the accurate definition and analysis of semantic roles. By examining authentic language use across diverse linguistic contexts, we can observe how semantic roles are naturally encoded in syntax and discourse. This evidence is invaluable for validating theoretical constructs and for adjusting linguistic models to better reflect the actual usage of language, particularly in defining the specific set and level of detail needed in semantic role classifications.

The use of corpus data allows for comprehensive cross-linguistic studies that can statistically analyze the occurrence and variation of semantic roles encoding across languages. Such studies are instrumental in identifying not only the commonalities across languages but also the unique

ways in which speakers structure information. For instance, corpus data can reveal how the role of 'Agent' might be explicitly marked in one language through case markings, while another might use verb conjugations or specific syntactic positions to convey the same role. These insights are crucial for developing a nuanced understanding of semantic roles that transcends linguistic boundaries and aids in defining a universally applicable yet sufficiently detailed set of semantic roles.

Corpus-based analyses can also contribute significantly to linguistic typology by providing data-driven insights into how languages vary in their semantic roles encoding. Additionally, these studies help in identifying rare or unique patterns that may have been overlooked in theoretical models, prompting reevaluations or refinements of existing linguistic classifications, and informing the granularity and composition of semantic role sets.

1.1.3 Semantic roles encoding in languages

Languages employ a range of strategies to encode semantic roles, including morphological, syntactic, and lexical methods (Luraghi & Narrog, 2014), each providing unique mechanisms for indicating the relationships between verbs and their arguments.

Many languages use morphological markers, such as case inflections, to indicate semantic roles. Morphological marking involves the use of specific endings or affixes on nouns to denote their roles within a sentence. For example, case systems often distinguish subjects from objects or agents from patients through distinct inflections. This method offers clear and unambiguous cues about the roles of sentence participants, contributing to the overall syntactic and semantic clarity of the language .

In contrast, languages with relatively fixed word orders often rely on the syntactic position of noun phrases to encode semantic roles. For example, the subject-verb-object (SVO) order prevalent in many languages places the subject before the verb and the object after it, which helps to identify their respective roles. This syntactic strategy leverages word order to convey the relationships between sentence elements without needing extensive morphological marking.

Some languages employ lexical items and prepositions to indicate semantic roles. Prepositions can specify the roles of noun phrases in relation to the action described by the verb, as seen in constructions like “to the store” indicating a goal or “with a hammer” indicating an instrument. These lexical strategies provide flexibility and can be particularly useful in languages with limited inflectional morphology.

Voice alternations, such as active and passive constructions, also play a crucial role in the encoding of semantic roles. In passive constructions, the patient (or theme) typically becomes the subject, while the agent may be demoted or introduced with a prepositional phrase. Additionally, verb agreement or cross-referencing systems in some languages encode semantic roles through verb morphology. For example, in certain languages, verbs agree with their subjects and objects in grammatical features, providing information about the roles of sentence participants through verb forms.

The strategies employed by languages to encode semantic roles reflect deep grammatical and conceptual differences across linguistic systems. Some languages may utilize a combination of these methods, balancing morphological, syntactic, and lexical strategies.

1.2 Semantic roles annotation

Semantic roles annotation is a critical task in natural language processing (NLP) that involves labeling words or phrases in a sentence with their corresponding semantic roles. This task is essential for various NLP applications, including machine translation, information extraction, and

question answering, as it provides a deeper understanding of sentence structure and meaning by identifying the relationships between verbs and their arguments.

1.2.1 Manual annotation efforts

Manual annotation of semantic roles is a labor-intensive process requiring linguistic expertise and meticulous attention to detail. The process typically begins with the development of comprehensive annotation guidelines, which are crucial to ensure consistency and accuracy. These guidelines define the set of semantic roles, criteria for their assignment, and provide illustrative examples.

Training annotators is another critical step, involving extensive sessions to familiarize them with the guidelines and the semantic roles to be annotated. Annotators undergo practice sessions, receive feedback, and participate in calibration exercises to achieve high inter-annotator agreement.

During the annotation process, annotators label sentences by identifying verbs and their arguments and assigning the appropriate semantic roles to each argument based on the guidelines. This step is often followed by quality control measures, including multiple rounds of review

and correction. Cross-checking by other annotators and supervisors ensures the reliability of the annotations.

Significant manual annotation efforts have led to the creation of valuable resources such as the PropBank (Kingsbury & Palmer, 2002; Palmer et al., 2005), FrameNet (Baker et al., 1998; Fillmore & Baker, 2012), and VerbNet (Kipper Schuler, 2005) corpora. PropBank provides a corpus of texts annotated with information about basic semantic propositions or roles, essentially linking verbs with their arguments. It focuses on adding a layer of predicate-argument information, or semantic roles, to the syntactic trees of the Penn Treebank. This helps in understanding who did what to whom, when, where, and how, by labeling the arguments of each verb in a sentence according to their roles. FrameNet, on the other hand, is built on the theory of frame semantics and involves annotating texts with semantic frames, which are conceptual structures describing various types of events, relationships, or objects along with their participants. FrameNet provides detailed information about how words are used in actual texts by mapping out the various contexts in which a word can appear, helping in the creation of more nuanced and context-aware language models. VerbNet is the largest on-line network of English verbs that groups verbs according to their shared syntactic and semantic behavior. It integrates both syntactic and semantic

information by defining verb classes and specifying the syntactic frames and semantic roles associated with each class. VerbNet enhances the PropBank and FrameNet resources by providing a hierarchical organization of verbs, detailing how different verbs can fit into different frames and argument structures, thus supporting more sophisticated natural language understanding and generation tasks.

Additionally, the Valency Patterns Leipzig (ValPaL) database (Hartmann et al., 2013) offers a comprehensive cross-linguistic comparison of valency classes. It includes annotations on major argument alternations for 80 verbs, selected as representative of the verbal lexicon. For each role, the database details a set of microroles and their syntactic behavior in these alternations, accompanied by annotations and examples from the languages in the database. Furthermore, the Pavia Verbs Database (PaVeDa; Luraghi et al., 2024) extends the work of ValPaL by including a number of ancient languages not covered by ValPaL and introducing new features that enable direct comparison, both diachronic and synchronic. PaVeDa builds on the ValPaL database of verbs' valency patterns and alternations, enhancing its scope and utility for linguistic research.

These annotated corpora are essential for training and evaluating automatic semantic role labeling systems, providing a foundation for further research and development in the field.

1.2.2 Automatic Semantic Role Labeling (SRL)

Automatic Semantic Role Labeling aims to replicate the manual annotation process using computational methods, thereby enhancing scalability and efficiency. SRL systems are designed to automatically identify and label semantic roles in text, utilizing a range of machine learning techniques.

Supervised learning approaches in SRL rely heavily on annotated corpora for training machine learning models. These models are typically trained using features extracted from the text, such as syntactic parse trees, part-of-speech tags, and word embeddings. The training process involves applying algorithms like support vector machines, conditional random fields, or neural networks to learn from the annotated data. Once trained, these models can be used to predict semantic roles in new, unseen text.

Unsupervised and semi-supervised learning approaches seek to reduce the dependence on large annotated datasets. These methods include clustering techniques, which group similar instances in the text to infer semantic roles based on patterns and similarities, and weak supervision,

which combines smaller annotated datasets with large amounts of unlabeled data to enhance model performance (Fürstenau & Lapata, 2012; Lang & Lapata, 2010, 2011).

Recent advancements in deep learning have significantly improved the performance of SRL systems. Neural network architectures, such as recurrent neural networks (RNNs; Zhou & Xu, 2015), convolutional neural networks (CNNs), and transformer models, have been effectively applied to SRL tasks. These models can capture complex patterns and dependencies in the data, leading to more accurate and robust semantic role labeling (He et al., 2017, 2018).

For instance, BERT-based models have been widely adopted in SRL due to their ability to leverage contextual embeddings for improved performance. Studies have shown that using BERT and its variants (e.g., RoBERTa, ALBERT) enhances the accuracy of SRL systems, particularly in capturing long-range dependencies and contextual nuances (Shi & Lin, 2019).

Another promising direction is the use of multitask learning frameworks that combine SRL with other related tasks, such as syntactic parsing or dependency parsing. This approach allows the model to learn

shared representations that benefit multiple tasks, leading to better overall performance (Strubell et al., 2018).

Despite these advancements, challenges remain in automatic SRL, including handling diverse linguistic phenomena, addressing low-resource languages, and improving the generalization of models across different domains and text genres. Nevertheless, automatic SRL offers significant advantages for research by dramatically reducing the time and effort required for manual annotation. It allows researchers to process large volumes of text quickly and consistently, facilitating more extensive and varied studies. Moreover, manual annotation does not avoid errors, as human annotators can introduce inconsistencies and mistakes. Automatic SRL systems, while not perfect, provide a level of consistency and scalability that manual methods cannot match. These systems can be iteratively improved and fine-tuned, leveraging large datasets and advanced machine learning techniques to enhance their accuracy and applicability across different languages and contexts. This streamlines many NLP tasks, making it easier to extract meaningful insights from vast amounts of textual data, ultimately accelerating the pace of research.

1.3 Quantitative methods to analyze semantic roles

Quantitative analysis is essential in the study of semantic roles, offering a systematic and empirical framework to uncover patterns, tendencies, and alignments within linguistic data. This section discusses the various methods applied in quantitative studies to investigate semantic roles.

Quantitative methods in semantic role analysis encompass a range of statistical and computational techniques designed to analyze large linguistic datasets. These methods facilitate the identification of patterns and trends that may not be apparent through qualitative analysis alone. One approach involves clustering algorithms, such as K-Means and Partitioning Around Medoids, which can be used to identify groups of semantic roles that frequently co-occur across different languages. This clustering helps in understanding the distribution and alignment types of semantic roles, revealing underlying structures that are consistent within and across languages.

Quantitative analysis is particularly valuable in cross-linguistic studies, where the goal is to compare and contrast the encoding of semantic roles across different languages. By analyzing large corpora from multiple

languages, researchers can identify universal patterns and language-specific peculiarities. This is especially important for linguistic typology.

The findings from quantitative studies of semantic roles have also important implications for computational linguistics. Enhanced machine translation systems, more accurate semantic parsers, and other natural language processing (NLP) tools benefit from a deep understanding of how semantic roles are encoded and used in different languages. Quantitative methods provide the data-driven foundation necessary to develop and refine these technologies.

Quantitative analysis also enhances the robustness and reproducibility of linguistic research. By employing standardized statistical methods and large, well-annotated datasets, researchers can ensure that their findings are reliable and can be independently verified. Moreover, the use of quantitative methods allows researchers to handle the vast amounts of data generated by modern linguistic research. Techniques such as corpus analysis, statistical modeling, and machine learning enable the efficient processing and analysis of this data, facilitating more comprehensive and detailed studies of semantic roles.

1.3.1 Semantic maps

In addition to the methods outlined in Section 1.3, the concept of semantic maps has emerged as a powerful tool in linguistic analysis. Semantic maps visually represent the relationships between different semantic roles and concepts, providing a clear and intuitive way to explore and compare these relationships across languages. By mapping out semantic roles, researchers can gain insights into how meanings are structured and how they vary from one language to another. This visual representation complements quantitative methods, offering a holistic approach to the study of semantics.

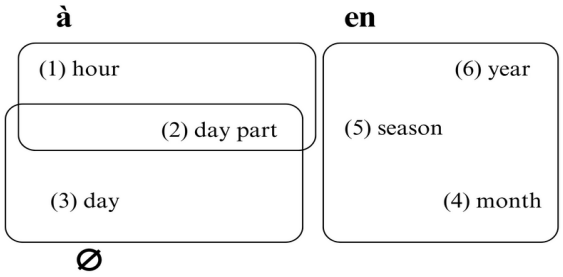


Figure 1. Semantic map from Haspelmath (1997).

Semantic maps were first introduced in linguistic research by Martin Haspelmath in the late 1990s. His seminal work (Haspelmath, 1997) utilized semantic maps (as the one in Figure 1) to illustrate the cross-linguistic similarities and differences in the expression of temporal adverbials. Haspelmath's innovative approach demonstrated the potential of

semantic maps to reveal underlying semantic structures and provided a foundation for subsequent studies in various linguistic domains.

Following Haspelmath's pioneering work, other researchers have expanded the application of semantic maps to different areas of semantics. Haspelmath (2003) further refined the technique and applied it to grammatical categories such as case, tense, and aspect. This study showcased the versatility of semantic maps in capturing the nuances of grammatical meanings across languages.

Semantic maps have also been employed to analyze the distribution and alignment of ditransitive constructions in a wide range of languages (Malchukov et al., 2010). The use of semantic maps in this context highlighted the intricate patterns of variation and provided a clear visual representation of cross-linguistic trends.

Considering diachronic analysis, semantic maps have also been used to identify patterns of polysemy among semantic roles from a cross-linguistic perspective. Luraghi (2014), for instance, explores how such polysemies develop through the meaning extension of morphemes, such as cases and adpositions, that encode semantic roles. Georgakopoulos and Polis (2021), instead, expand the application of the semantic map model to

diachronic lexical semantics by integrating a quantitative analysis of extensive synchronic polysemy data with a qualitative assessment of diachronic material from ancient Egyptian and ancient Greek texts.

The development and application of semantic maps have significantly enhanced our understanding of semantic roles and their interactions. By providing a visual framework, semantic maps allow researchers to identify and analyze patterns that might be difficult to discern through purely quantitative or qualitative methods. As a result, semantic maps have become an indispensable tool in the study of linguistic typology, aiding in the exploration of universal and language-specific semantic structures.

In recent years, the integration of semantic maps with advanced computational techniques has further expanded their utility. Computational models can now generate and analyze semantic maps from large linguistic datasets, enabling more sophisticated and comprehensive analyses. This synergy between traditional linguistic methods and modern computational approaches continues to drive the field forward, offering new insights into the complex web of semantic relationships that underpin human language.

One such advancement is the use of quantitative semantic maps based on multidimensional scaling (MDS). Multidimensional scaling is a

statistical technique that represents data in a low-dimensional space, preserving the distances between data points as accurately as possible. In the context of semantic maps, MDS can be used to visualize the semantic distances between different roles and concepts, allowing for a more precise and quantitative representation of their relationships. have demonstrated the effectiveness of this approach in providing a detailed and quantifiable perspective on semantic roles (Cysouw, 2014; Wälchli, 2010; Wälchli & Cysouw, 2012).

Quantitative semantic maps based on MDS offer several advantages. They enable the identification of subtle semantic distinctions and similarities that may not be evident through traditional methods. Additionally, they provide a robust framework for testing hypotheses about semantic structures and their cross-linguistic variation. By combining the strengths of semantic maps and multidimensional scaling, researchers can achieve a deeper and more nuanced understanding of the complex interplay between language and meaning.

1.4 Goal of this work

The primary goal of this work is to analyze how semantic roles are encoded by examining data from a multilingual parallel corpus, specifically focusing

on transfer verbs and communication verbs. By investigating these two categories of verbs across different languages, this study aims to uncover patterns and variations in the encoding of semantic roles.

1.4.1 Transfer verbs

Transfer verbs are a category of verbs that describe the action of moving an object from one entity to another. Common examples of transfer verbs include “give” “send”, “bring” and “take”. These verbs typically involve three main participants: the giver (agent), the receiver (recipient), and the object being transferred (theme). According to Levin (1993), transfer verbs are crucial in understanding how different languages handle the syntax-semantics interface, especially regarding argument structure and case marking. For instance, in English, the verb “give” in the sentence 13 clearly delineates the agent (Alice), the recipient (Mark), and the theme (a book). In contrast, other languages might use different syntactic structures or morphological markers to encode these roles.

Malchukov et al. (2010) have shown that the encoding of transfer verbs can vary significantly across languages, with some employing double-object constructions and others using prepositional phrases to indicate the

recipient. This variation highlights the importance of a comparative approach in understanding the semantic roles associated with transfer verbs.

1.4.2 Communication verbs

Communication verbs encompass actions related to conveying information, such as “say”, “tell”, “ask”, “inform”, and “declare”. These verbs typically involve a speaker (agent), an addressee (recipient), and the content of the communication (message). For example, in the sentence (2):

(2) John told Mary the news.

“John” is the agent, “Mary” is the recipient, and “the news” is the message.

Research by Goldberg (1995) highlights the importance of communication verbs in examining how languages represent speech acts and information flow. Different languages may vary significantly in how they encode the participants of communication events, whether through direct object marking, word order, or the use of specific verb forms.

Languages employ various syntactic and morphological strategies to encode the semantic roles of communication verbs. In some languages, the roles are marked explicitly through case marking or prepositional phrases, while in others, they might be inferred from context or indicated by word order. For example, in English, the recipient of a communication verb can

be introduced by the preposition “to” (e.g., “He said to her”) or placed directly after the verb (e.g., “He told her”). In contrast, languages like Japanese often use particles to mark the different roles, such as "に" (ni) for the recipient.

This work is structured as follows: Section 2 outlines the data and methods used, detailing the data collection (2.1), data annotation (2.2), text alignment (2.3), and semantic roles extraction (2.4). Section 3 presents the statistical analysis results, offering interpretations: it includes the processing of raw data (3.1), results from cluster analysis (3.2), and the interpretation of semantic maps based on the distances among instances of semantic roles in the multilingual parallel corpus (3.3). Finally, Section 4 discusses the conclusions and potential future developments.

2 Data and Methods

This chapter serves as a comprehensive guide to the methods and techniques utilized in this thesis, which investigates the morphosyntactic structures and semantic roles across multiple languages using the Multilingual Parallel Bible Corpus (Christodouloupoulos & Steedman, 2015; henceforth MPBC). Given the complexity and scope of this study, a detailed methodology is essential to ensure replicability and clarity in how the research questions are addressed.

The primary aim of this research is to understand how different languages structure sentences and assign semantic roles. The specific objectives of the methodology include:

- Data Collection and Preprocessing: to select and prepare a suitable corpus that facilitates cross-linguistic comparison.
- Morphosyntactic Annotation: to annotate the corpus using Dependency Grammar within the Universal Dependencies framework to maintain consistency and comparability across languages.

- Semantic Role Extraction: to integrate semantic roles with morphosyntactic annotations, focusing on the identification of verbs and their arguments across different languages.
- Data Analysis: to employ statistical methods to analyze the data, aiming to reveal patterns and insights into language-specific and general features.

2.1 Data collection and preprocessing

The data employed in this study was extracted from the Multilingual Parallel Bible Corpus (MPBC), an expansive digital corpus created by Christodouloupoulos and Steedman in 2015. This corpus is a significant academic resource, hosted online at the Christos-C Bible Corpus website (https://christos-c.com/bible/), comprising 108 translations of the Bible formatted in XML. This digital format is particularly advantageous for linguistic research because each XML file contains comprehensive metadata that allows for precise retrieval of text according to the book and verse, a feature crucial for any detailed textual analysis across multiple languages.

The selection of languages from the MPBC was determined by a set of strategic criteria designed to ensure the feasibility of conducting high-

quality morphosyntactic annotations and semantic role extractions. First, only languages that were included in the MPBC were considered, focusing on those with a complete translation of the Gospels and Acts. These texts were chosen due to their ubiquitous presence across all Bible translations in the corpus, providing a consistent comparative base.

Furthermore, the availability of robust natural language processing (NLP) tools was a determining factor. Languages for which advanced NLP models for morphosyntactic annotation and semantic role extraction were available were prioritized. This approach streamlined the selection process to 38 languages, each representing a unique linguistic family and structure, which ensured a rich and diverse linguistic sample for the study. This selection was not only pragmatic—given the technological requirements of the study—but also methodologically significant, as it aimed to maximize both linguistic diversity and analytical depth.

This resulted in a set of 38 languages, listed in Table 1.

Table 1. *List of languages in the initial sample*

Language code	Language	Family	Genus
afr	Afrikaans	IE	Germanic
ara	Arabic	Afro-Asiatic	Semitic
bul	Bulgarian	IE	Slavic
ces	Czech	IE	Slavic

dan	Danish	IE	Germanic
deu	German	IE	Germanic
ell	Modern Greek	IE	Greek
eng	English	IE	Germanic
est	Estonian	Uralic	Finno-Ugric
eus	Basque	Isolated	-
fas	Persian	IE	Indo-Iranian
fin	Finnish	Uralic	Finno-Ugric
fra	French	IE	Italic
heb	Modern Hebrew	Afro-Asiatic	Semitic
hrv	Croatian	IE	Slavic
hun	Hungarian	Uralic	Finno-Ugric
hye	Armenian	IE	Armenian
ind	Indonesian	Austronesian	Malayo-Sumbawan
ita	Italian	IE	Italic
jpn	Japanese	Japonic	-
lav	Latvian	IE	Baltic
lit	Lituanian	IE	Baltic
mar	Marathi	IE	Indo-Iranian
nld	Dutch	IE	Germanic
nor	Norwegian	IE	Germanic
pol	Polish	IE	Slavic
por	Portuguese	IE	Italic
ron	Romanian	IE	Italic
rus	Russian	IE	Slavic
slk	Slovak	IE	Slavic
slv	Slovenian	IE	Slavic

spa	Spanish	IE	Italic
srp	Serbian	IE	Slavic
swe	Swedish	IE	Germanic
tel	Telugu	Dravidian	South-Central
tur	Turkish	Altaic	Turkic
ukr	Ukranian	IE	Slavic
vie	Vietnamese	Austro-Asiatic	Mon-Khmer
zho	Chinese	Sino-Tiberan	Sinitic

As illustrated in Table 1, he majority of the languages selected for this study—26 out of 38—belong to the Indo-European family. Furthermore, 27 of these languages are predominantly spoken in Europe. This concentration of Indo-European languages, particularly within European regions, raised concerns about potential genealogical and areal biases in the study. Genealogical bias occurs when the languages studied are overly representative of certain language families, potentially skewing results towards those family-specific linguistic features. Similarly, areal bias can occur when languages from a specific geographical region are overrepresented, which might lead to an overemphasis on regional linguistic phenomena that are not necessarily representative of other regions or language families.

To address these concerns and ensure a more balanced and representative linguistic analysis, the selection set was further refined. This refinement aimed to achieve a broader and more equitable representation of linguistic diversity. The approach involved not just narrowing the focus to non-Indo-European languages but also including languages from more diverse geographical areas beyond Europe. This strategic adjustment was crucial for reducing biases and enhancing the validity and generalizability of the study's findings.

This process of further restricting the sample set is detailed in Section 2.1.2, where methods for mitigating bias in linguistic studies are discussed extensively. The section outlines the criteria and methodologies employed to ensure a balanced representation of linguistic data, thereby facilitating a more accurate and comprehensive analysis of morphosyntactic structures and semantic roles across different language families.

2.1.1 Data preprocessing

The preprocessing stage was critical to adapt the XML formatted texts for the linguistic analyses planned in this study. Each text segment in the MPBC is contained within "seg" tags, which include attributes specifying the book, chapter, and verse. This structured format greatly aids in the accurate

extraction of required texts. An example of a typical XML segment looks like this:

Code 1. Portion of the XML file with the Italian translation.

```
<seg id=”b.MAT.23.6” type=”verse”>
      amano posti d'onore nei conviti, i primi seggi
nelle sinagoghe
</seg>
```

In the example in Code 1, the id attribute “b.MAT.23.6” clearly indicates that the verse is from the Gospel of Matthew, chapter 23, verse 6. This level of detail in the corpus structure facilitates targeted text retrieval, which is essential for consistent and accurate analysis.

A significant aspect of the preprocessing involved reorganizing the verse-based text into sentences. Given that linguistic analysis, particularly syntactic parsing, is more effectively conducted on full sentences rather than on verses, the texts needed to be segmented accordingly. This segmentation involved parsing the text data to delineate sentences clearly, a task that requires precision to ensure that the syntactic structure inherent in the original text is maintained.

Moreover, special HTML characters embedded in the original XML files were converted into their corresponding Unicode characters to prevent any encoding issues during subsequent processing steps. This

standardization is crucial when dealing with texts in multiple languages, where non-standard characters or symbols might otherwise lead to data corruption or processing errors.

Once these preprocessing tasks were completed, the texts were meticulously organized into well-structured files categorized by language, book, and chapter. This organization not only facilitated efficient access during the analysis phase but also ensured that each text segment was readily available in a format conducive to detailed linguistic analysis.

2.1.2 Sampling

The process of selecting languages for this study involved refining an initial set of 38 languages derived from the MPBC. Given the original set's heavy skew towards Indo-European languages, the sampling strategy was designed to ensure a broader linguistic diversity, both in terms of language families and geographical distribution.

The sampling strategy was structured around several key principles. First, it was essential to include at least one language from each language family present in the initial 38-language pool to ensure that the study captured a wide array of linguistic structures and typological features. This

decision aimed to facilitate a comprehensive analysis across diverse linguistic systems.

In addressing the overrepresentation of Indo-European languages, the strategy included a limitation of no more than two languages from each genus within this family. This approach aimed to prevent any single linguistic genus from dominating the dataset, providing a more balanced view across this extensively represented family.

Moreover, for the two languages selected from the same Indo-European genus, a geographical criterion was applied. Languages that were geographically distant from each other were preferred to minimize areal biases—tendencies for languages in close geographical proximity to influence each other's linguistic features due to historical language contact or shared regional influences.

This thorough selection process culminated in a final set of 19 languages. This reduction not only streamlined the focus of the study but also enhanced its methodological rigor by ensuring a balanced representation across different linguistic families and geographic areas. The chosen languages include representatives from unique families such as Uralic and Afro-Asiatic, as well as diverse branches of the Indo-European family, as shown in Table 2.

Table 2. Final language sample.

Language code	Language	Family	Genus
afr	Afrikaans	IE	Germanic
ara	Arabic	Afro-Asiatic	Semitic
ell	Modern Greek	IE	Greek
eng	English	IE	Germanic
eus	Basque	Isolate	-
fas	Persian	IE	Indo-Iranian
fin	Finnish	Uralic	Finno-Ugric
heb	Modern Hebrew	Afro-Asiatic	Semitic
hun	Hungarian	Uralic	Finno-Ugric
ind	Indonesian	Austronesian	Malayo-Sumbawan
ita	Italian	IE	Italic
jpn	Japanese	Japonic	-
lav	Latvian	IE	Baltic
lit	Lithuanian	IE	Baltic
mar	Marathi	IE	Indo-Iranian
pol	Polish	IE	Slavic
rus	Russian	IE	Slavic
spa	Spanish	IE	Italic
tur	Turkish	Altaic	Turkic

2.2 Morphological and syntactic annotation

Before discussing the specifics of the morphological and syntactic annotation processes, it is essential to understand the foundational framework employed in this study—Dependency Grammar (DG). Dependency Grammar is a robust linguistic theory that focuses on the relationships between words within a sentence, positing that the structure of a sentence is derived from the dependencies between words rather than a hierarchy of phrases.

2.2.1 Dependency Grammar

Dependency Grammar, as a theory, originated from the work of Lucien Tesnière (1959) in the mid-20th century, whose seminal ideas have since profoundly influenced modern syntactic theories. The essence of DG is to depict linguistic structures not as isolated units but rather as interconnected nodes within a network. This network is visually and analytically represented in the form of trees—specifically, syntax trees where words of a sentence are interconnected through directed links that establish hierarchical relationships.

Each word in a sentence is treated as a node in this tree, with connections—or edges—defining the syntactic dependencies between them. These dependencies determine how words combine to form phrases and

sentences, with one word typically governing the connection and others depending on it. The primary advantage of using trees in DG is their ability to clearly define and display these relationships. The trees are directed and rooted, meaning they have a clear origin or 'root' that dominates the structure, and they are acyclic, preventing any loops within the hierarchy which ensures clarity and precision in syntactic parsing.

The decision to employ Dependency Grammar in this study was driven by practical considerations aligned with our analytical goals. Specifically, the framework's capacity to straightforwardly represent sentence structure makes it ideal for automated parsing, which is crucial given the multilingual scope of the corpus and the use of Natural Language Processing (NLP) tools. DG's structure is particularly conducive to the tasks of syntactic and semantic role annotation. It allows for precise control over the parsing process, ensuring that each word's role and relationships are clearly defined and consistently interpreted across different languages.

Moreover, Dependency Grammar's focus on word-level connections offers distinct advantages over other syntactic frameworks such as Phrase Structure Grammar (PSG). While PSG also uses tree-based representations, it incorporates both words and phrases as nodes within the same tree, which can complicate the parsing and annotation process. This complexity arises

because PSG's hierarchical structures need to accommodate varying levels of linguistic units, making the trees more difficult to navigate and analyze, especially when attempting to reconcile syntactic structures with semantic roles.

The decision to adopt Dependency Grammar, particularly under the Universal Dependencies (de Marneffe et al., 2021) framework, was influenced significantly by the availability of NLP tools that support automatic annotation in multiple languages. These tools, developed specifically for Dependency Grammar, provide robust mechanisms for parsing complex syntactic structures and control for the semantic roles annotation based on the dependencies that originate from the verb. This is especially critical for a multilingual study, where consistent parsing across languages is necessary to ensure comparative analysis remains valid.

By aligning the study with the Universal Dependencies framework, which is widely supported in the NLP community, it ensures that the tools used are not only capable of handling the linguistic complexity of multiple languages but also adhere to a standardized approach, facilitating reliability and cross-linguistic comparability in the annotations.

2.2.1.1 Universal Dependencies

Universal Dependencies (UD) is a comprehensive, open-source framework designed for the consistent annotation of grammatical information across languages. It utilizes the principles of Dependency Grammar (DG) to represent syntactic relationships as networks of pairwise dependencies between words. By providing a uniform set of labels and guidelines, UD enables the comparison of syntactic structures globally, facilitating a deeper understanding of language patterns and structures. Since its inception in 2014, the UD project has grown significantly, encompassing 259 annotated treebanks covering 148 languages as of its latest release (version 2.13 on November 15, 2023). This extensive coverage and its widespread adoption have made UD a standard in the field of dependency syntax.

The UD annotations are stored in the CoNLL-U format, an evolution of the CoNLL-X format developed by Buchholz & Marsi (2006). This format is meticulously designed to capture detailed linguistic information, organizing data in a way that supports efficient parsing and analysis. Each token in a sentence is recorded on a separate line, followed by annotations that describe its linguistic properties. Sentences themselves are separated by blank lines, ensuring clear demarcation within the text.

The CoNLL-U format is particularly notable for including two preliminary comment lines at the beginning of each sentence. These lines, which start with a hash ('#'), contain the sentence's original text and a unique sentence identifier. This structure not only maintains the integrity of the original data but also allows for precise tracking and referencing within and across treebanks.

Code 2. *Example of metadata of a sentence.*

```
# sent_id = 39
# text = Os digo que entre los nacidos de mujer, no hay
ninguno mayor que Juan.
# alignment = 37
```

Code 2 shows the metadata of the sentence 39 of the Spanish translation of the chapter 7 of the Gospel of Luke: as described earlier, it is possible to see that each line starts with a hash and keys (‘sent_id’, ‘text’ and ‘alignment’) and values (‘39’, ‘Os digo que entre los nacidos de mujer, no hay ninguno mayor que Juan.’ and ‘37’) are separated by a ‘=’ character.

Then, ‘syntactic words’ appear in a one-word-per-line format. UD guidelines define syntactic words as the basic unit of syntactic annotation. They differ from orthographic or phonological words since the division of a sentence into tokens does not take into account as the most important criterion white spaces or prosody. The tokenization of sentences splits off

systematically clitics producing, in these cases, more than one token from one orthographic word. Examples of this procedure can be found in the Spanish word dámelo (‘give it to me’), which would be split into three tokens, namely ‘da’, ‘me’ and ‘lo’.

Table 3. CoNLL-U fields and their description.

Column name	Description
id	The identifier of the token
form	The word form as it appears in the sentence
lemma	The lemma or stem of the word form
upos	The universal part-of-speech tag
xpos	The language-specific part-of-speech tag
feats	The morphological features
head	The identifier of the token which is the syntactic head
deprel	The dependency relation between the head and the token
deps	Enhanced dependencies
misc	Any other annotation

Each line storing a token and its annotation consists of ten fields, separated by a tab character. The first field stores the identifier of the token. The identifier is unique for each token within the sentences and it also gives an information about the position occupied by the token in the sentence. In fact, identifiers are integer progressive numbers starting from 1. In the case

of a ‘multitoken word’, i.e. an orthographic word consisting of two or more tokens, the id will be formatted as follows. Let us suppose that in a French sentence there are 6 tokens before the multitoken word au (‘to the’), consisting of two tokens (‘á’ and ‘le’); in the CoNLL-U file we will find, after the lines storing the sentence’s metadata, 6 lines storing the first 6 tokens of the sentence and their annotations and then three lines as exemplified in Code 3.

Code 3. Example of the annotation of a multitoken word.

```
7-8  au  _  _  _  _  _  _  _  _
7    á   [annotation of the token]
8    le  [annotation of the token]
```

This way of formatting multitokens words allows to avoid losing the original text in the CoNLL-U annotation of the sentence and to say, for each token, if it belonged to a multitoken word or not.

In the CoNLL-U file format, the column following the token identifier holds the token's form as it appears in the text. For a file to be considered well-formed, this sequence of forms within each sentence must align with the contents stored in the metadata text field. Another key layer in this format is the lemma, or the canonical form of a word, often defined as the dictionary entry of a token. The selection of the lemma is arbitrary and varies by language. For instance, Italian treebanks adopt as lemma for nouns

and adjectives the masculine singular form, and for verbs the present infinitive. Other languages may adopt different choices for lemmas, also according to the features that a certain part-of-speech has in such languages. In languages with case marking, nouns are typically lemmatized to the nominative case or to their stem. Other lemmatization choices may trace back to traditional grammars' choices: Latin verbs, for instance, are lemmatized using the first person of the present indicative, Arabic verbs with the third person. Other lemmatization choices are more practically oriented: Italian definite articles 'il', 'lo', 'la', 'i', 'gli' and 'le' are consistently lemmatized with the form 'il', without a particular reason to prefer it to the other forms. One may argue that 'il' is the masculine singular form of the Italian definite article and the choice of this form to lemmatize the other elements in this class is consistent with what is done with noun and adjectives. But 'il' is not the only masculine form of the definite article: the preference given to 'il' instead of 'lo' might be due to the frequencies of these to forms in the corpora, but it is an arbitrary choice. Similarly, for the indefinite articles 'un', 'uno' and 'una', the choice made by the maintainers of the treebanks is to lemmatize all of them with the 'uno' form. This, for the case of Italian treebanks, is due to the fact that the part-of-speech tag used for articles is 'DET' and such label is also used to tag other types of

determiners such as demonstratives and quantifiers. Here, the choice to use these practice to lemmatize Italian definite and indefinite articles it is useful to isolate these elements from the rest of determiners when querying the treebank to extract them. In fact, to extract all Italian articles, it would be enough to specify the set of values that the lemma field should match, i.e. ‘il’ or ‘uno’; adding the value of the part-of-speech in the query would be redundant, since all occurrences of definite and indefinite articles are tagged as ‘DET’, while looking for all the tokens having ‘DET’ ad part-of-speech tag would result in some noise in the results.

The fourth and fifth columns of the CoNLL-U format store the annotation of the part-of-speech tag. I will refer to them as ‘upos’, that stands for universal part-of-speech and ‘xpos’, that is the language specific part-of-speech. When starting the Universal Dependencies project, some resources annotated with the parts-of-speech were already available. They adopted a tagset specifically designed for the purpose of annotating a specific language and often annotators from different projects adopted different tagset to annotate the same language data. Universal Dependencies, instead, decided to use a fixed list of tags to label the parts-of-speech in its treebanks. Doing so, when converting already existing resources to the UD guidelines, annotators and maintainers decided to keep,

along with the converted upos, also the language-specific (or corpus-specific) tags. This annotation was kept separated from the upos and it is stored in the fifth column. The UD guidelines specify 17 upos tags to cover open class words (6 tags), closed class words (8 tags), and other words that do not fit these categories (3 tags).

When available, the use of dual tagging systems (upos and xpos) highlights the flexibility in accommodating both universal and language-specific grammatical classifications. This approach benefits linguistic research by facilitating cross-linguistic analysis and comparison while preserving the detail of individual languages.

The sixth field of the CoNLL-U format is dedicated to the annotation of features. This column stores additional information about the word, including its part of speech and morphosyntactic properties. Each feature in the CoNLL-U file is formatted as `Name=Value`, and any word can have multiple features, separated by a pipe character (`|`). Similar to the treatment of parts of speech, the Universal Dependencies (UD) guidelines define an inventory of features and associated values to ensure uniform encoding. However, users can extend these sets and add language-specific features and values when necessary. The UD guidelines also establish general principles

for the inventory of features and their values, which user-defined features should follow.

Apart from formatting principles, two key points are noteworthy: first, if a feature is not mentioned in the data, it implies an empty value, indicating that the feature is either irrelevant for this part of speech or its value cannot be determined for this word form due to language-specific reasons. Second, it is possible to assign multiple values to a feature for a given word (e.g., `Case=Acc,Dat`), meaning the word may have one of these values, but it is not possible to determine which one specifically.

The seventh and eighth columns of the CoNLL-U format are used to store the syntactic annotation of the sentence. The first of these two fields contains integer values that point to other tokens in the sentence, allowing for the identification of each token's syntactic head and the reconstruction of the syntactic tree structure. There are several restrictions on the values allowed in this field. Firstly, the numbers must correspond to the identifier of a token in the sentence, meaning that negative numbers and numbers exceeding the total number of tokens in the sentence are not permitted. Additionally, a token cannot have itself as its head, adhering to Dependency Grammar rules that prohibit circular syntactic links. By convention, the root

of the syntactic tree is assigned a head value of `0`, and it is the only token permitted to have this value.

Following the column with the head information, there is a column dedicated to annotating the dependency relation. This field specifies the type of syntactic link between the token and its head. For the token that serves as the root of the syntactic tree, the label `root` is used, and this is the only token assigned this specific dependency relation. The other permissible labels are defined by the Universal Dependencies guidelines and constitute a closed set. This set includes 37 labels (de Marneffe et al., 2014) that cover a wide range of syntactic relations, as well as some relations that are not dependency relations in the strict sense. These predefined labels ensure consistency and uniformity in syntactic annotation across different languages and datasets.

The ninth column of the CoNLL-U format is also used to store information necessary for reconstructing a type of dependency structure. According to the UD guidelines, this field is called `deps`, and it stores the "enhanced dependencies". UD trees feature many direct dependencies between content words, with numerous dependency labels offering detailed information about the nature of these relationships. However, in some constructions, the dependency path between two content words can be quite

lengthy, complicating the process of determining their relationship. To address these challenges, the UD guidelines include provisions for an enhanced representation. This enhanced representation makes some of the implicit relations between words more explicit and augments certain dependency labels to facilitate the disambiguation of various types of arguments and modifiers. The possible enhancements include:

- Empty (null) nodes for elided predicates;
- Propagation of incoming dependencies to conjuncts;
- Propagation of outgoing dependencies from conjuncts;
- Additional subject relations for control and raising constructions;
- Coreference in relative clause constructions;
- Modifier labels that contain the preposition, other case marker, or conjunction.

These enhancements provide a more detailed and explicit representation of the syntactic and semantic relationships in a sentence, making it easier to perform accurate and nuanced natural language understanding tasks. Not all the treebanks in the Universal Dependencies project include the annotation of this field. If a corpus does not annotate any of the enhancements defined in the guidelines, an underscore character in the `deps` column is used to indicate the absence of enhanced annotations.

The tenth and final column is called `misc` (short for "miscellaneous") and is used to store any additional information that annotators wish to keep at the token level. It follows the same format as the features field, using pairs of attributes and values separated by 'pipe' characters, and its annotation is optional. The attributes and values in this column are not restricted to a closed set, although some attributes are consistently used across multiple treebanks. This field was particularly useful for my study, as it allowed me to store the annotations of frames and roles obtained by InVeRo-XL. (see Section 2.4.2 where the integration of the annotation of the roles into the treebanks is thoroughly discussed).

2.2.2 Parsing operations

The transformation of texts into CoNLL-U formatted annotations was adeptly handled using Stanza, a versatile and open-source natural language processing toolkit developed by Qi et al. (2020). Supporting 66 human languages, Stanza is designed to perform multiple core NLP tasks—including tokenization, sentence segmentation, syntactic parsing, and named entity recognition—according to the Universal Dependencies (UD) guidelines. Its capability to output fully annotated CoNLL-U files makes it

especially valuable for linguistic research that requires high precision in syntactic and semantic analysis.

Stanza was chosen for this study due to its comprehensive language support and robust performance across essential NLP tasks such as sentence segmentation, tokenization, part-of-speech (POS) tagging, morphological feature annotation, and syntactic parsing. These functionalities are critical for converting unstructured text into reliable linguistic data. The performance metrics for Stanza's pre-trained models, detailed in Table 4, highlight its effectiveness across these varied tasks, assessed through evaluations on dependency treebanks.

Table 4. *Stanza's performances by task.*

Language	Sentence segmentation	Tokenization	POS tagging	Features	Syntax[1]
Afrikaans	99.65	99.99	98.60	95.66	89.54
Arabic	83.56	99.98	95.18	92.18	83.99
Armenian	99.56	99.95	96.42	92.04	87.73
Basque	100.00	99.98	96.07	91.16	86.65
Bulgarian	96.50	99.96	99.16	96.62	93.98
Chinese	99.10	92.14	88.82	87.95	73.41
Croatian	97.08	99.96	98.21	94.60	90.29
Czech	94.62	99.97	98.88	94.65	92.22
Danish	91.81	99.97	98.30	97.14	88.61
Dutch	90.08	99.87	96.75	94.97	91.42
English	95.34	99.78	97.59	96.52	90.49
Estonian	92.88	99.97	97.19	94.39	86.60

1 As performace index for syntax was considered the unlabeled attachment score (UAS).

Finnish	90.70	99.73	97.51	95.39	90.37
French	95.03	99.71	97.63	97.05	93.27
German	83.84	99.62	95.61	87.19	85.80
Hungarian	97.21	99.86	95.80	92.74	83.57
Indonesian	93.04	99.88	94.16	89.38	87.05
Italian	100.00	99.81	98.42	97.51	93.53
Japanese	99.72	97.37	96.38	95.45	90.01
Latvian	98.15	99.81	96.70	89.29	88.91
Lithuanian	89.93	99.94	93.71	85.54	77.82
Marathi	89.36	97.99	85.78	71.57	68.87
Modern Greek	93.66	99.90	97.71	94.01	91.22
Modern Hebrew	96.78	99.64	91.69	86.14	83.23
Norwegian	97.70	99.88	98.38	96.38	93.26
Persian	99.73	99.96	97.17	95.35	92.90
Polish	98.19	99.85	98.63	93.89	93.73
Portuguese	93.10	99.85	97.45	95.33	90.59
Romanian	95.72	99.73	97.50	96.74	90.70
Russian	98.43	99.74	98.06	92.44	93.26
Serbian	98.46	99.99	98.44	93.76	91.44
Slovak	84.31	99.96	96.34	87.35	89.13
Slovenian	99.26	99.93	98.41	95.37	92.45
Spanish	98.49	99.97	98.96	95.65	93.09
Swedish	97.49	99.97	98.28	96.24	90.32
Telugu	98.62	99.79	93.97	93.97	90.09
Turkish	98.16	99.86	93.44	89.07	73.39
Ukrainian	97.21	99.79	97.52	92.07	88.60
Vietnamese	99.25	88.02	81.48	80.56	58.82

The process begins with sentence segmentation, where Stanza identifies sentence boundaries primarily by detecting punctuation marks. To enhance the tool's automatic segmentation accuracy, the texts were manually

reviewed, and any missing punctuation marks were added to ensure correct sentence demarcation.

Following segmentation, the tokenization phase involves splitting the text into individual tokens. Stanza's sophisticated approach to tokenization goes beyond simple whitespace separation; it accurately determines boundaries between words, punctuation, and other elements, considering the linguistic structure of the text.

After tokenization, each token is annotated with a part-of-speech (POS) tag that categorizes its grammatical function within the sentence. Stanza uses contextual information to ensure that each word is labeled according to its syntactic role, facilitating accurate analysis in subsequent stages.

The annotation process also extends to morphological features, where each token is assigned detailed grammatical attributes such as tense, number, person, and gender. These features provide additional linguistic information that supports deeper syntactic and semantic analyses.

The culmination of the process is syntactic parsing, where Stanza constructs a syntactic tree for each sentence. In this structure, tokens are linked to their heads, forming clear dependency relationships.

The accuracy of Stanza's syntactic parsing is evaluated using two primary metrics: the Unlabeled Attachment Score (UAS) and the Labeled Attachment Score (LAS). UAS measures the proportion of tokens in a sentence that are correctly attached to their syntactic heads, focusing solely on the structural correctness of the syntactic trees. Although only UAS is reported in Table 4 due to its emphasis on structural aspects, LAS provides a more nuanced measure by also considering the accuracy of the dependency labels.

These metrics serve as benchmarks to show Stanza's performance in handling the complex demands of syntactic parsing, confirming the toolkit's capability to deliver high-quality, reliable annotations that are consistent across various languages and texts.

2.3 Alignment of the texts

The MPBC, as detailed in the previous sections, has been meticulously structured according to a hierarchical organization based on language, book, and chapter. For efficient analysis, I divided the text within each language into chapters and stored these divisions into separate files. Leveraging the capabilities of the Stanza pipeline for sentence segmentation, I then

formatted the text in each file to follow a one-sentence-per-line structure. This organization was crucial for supporting subsequent analyses, particularly for sentence alignment across different languages.

To conduct a thorough analysis of semantic roles across various languages, it was essential to align the texts at the sentence level. The roles associated with verbs—key components in semantic structure—are typically confined within single sentences.

Sentence alignment involves matching sentences in two parallel documents, which are translations of the same text in different languages. The objective is to identify corresponding sentences across these translations, a task complicated by variations in sentence structure and length across languages. The alignment process must account for scenarios where a sentence in one text corresponds to multiple sentences in another, reflecting the complex syntactic and semantic transformations that occur in translation.

Historically, sentence alignment techniques, such as those developed by Gale & Church (1993), were language agnostic and primarily based on sentence length, positing that longer sentences in one language tend to translate into longer sentences in another. Subsequent advancements incorporated lexical features to enhance precision and speed (Li et al., 2010;

Ma, 2006). With the advent of more sophisticated technologies, techniques like Bleualign (Sennrich & Volk, 2010), which utilize automatic translations to score and adjust alignments, and Coverage-Based methods (Gomes & Lopes, 2016) that employ phrase tables from Moses (Koehn et al., 2007), have further refined the accuracy and efficiency of sentence alignment.

In recent years, neural network-based models have revolutionized sentence alignment. These models leverage their capacity to learn complex patterns from large datasets and handle noisy or incomplete data efficiently, delivering state-of-the-art results in text alignment tasks. Central to the effectiveness of these neural networks are word and sentence embeddings.

Embeddings are numerical representations of words or sentences that encapsulate semantic meanings. In the context of neural networks, embeddings are dense vectors where each dimension represents a latent feature of the word or sentence, learned from the data. Words or sentences with similar meanings are represented by vectors that are closer together in the vector space, enabling the model to capture and utilize semantic relationships effectively.

These embeddings are typically generated through models like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or

FastText (Joulin et al., 2016) for words, and more complex architectures like BERT (Devlin et al., 2019) or GPT for sentences. The vectors are trained to predict aspects of the language, such as the context in which a word appears (contextual embeddings), which provide to the embeddings rich semantic properties. For sentence embeddings specifically, techniques often involve encoding the aggregated or weighted properties of word embeddings within the sentence, or directly learning sentence representations using the context of the surrounding text.

By incorporating embeddings into alignment models, researchers can vastly improve the accuracy of aligning texts across languages by focusing on semantic consistency rather than just lexical or syntactic similarity. This methodological advancement has significant implications for automatic machine translation systems and other NLP applications, offering a more nuanced understanding of language transfer in multilingual contexts.

2.3.1 Alignment algorithm

In the process of aligning sentences within the MPBC, advanced neural network-based techniques were utilized, significantly enhancing the accuracy and efficiency of the alignment methodology. A critical component of this strategy involved employing a state-of-the-art model, the

intfloat/multilingual-e5-large (Wang et al., 2024). designed specifically for handling multilingual sentence representations. This model was instrumental in achieving high-quality alignments across the diverse languages represented in the MPBC.

The intfloat/multilingual-e5-large model, a powerful neural network designed for multilingual contexts, was chosen for its robust performance in processing and aligning sentences across different languages. This model is adept at generating embeddings that capture deep semantic nuances, making it ideal for the task of sentence alignment where semantic consistency across languages is crucial. By leveraging this model, I were able to significantly improve the semantic coherence and accuracy of the aligned sentences.

In the alignment process, English was utilized as a pivot language. This approach means that all other languages in the corpus were aligned to English, and then English served as a reference point to establish alignments between each of the other languages. This strategy is commonly employed in multilingual studies due to the extensive availability of English language resources and its status as a lingua franca in much of the global linguistic research. Using English as a pivot language simplifies the alignment process, as it reduces the complexity of directly aligning each language pair and ensures consistency and comparability across all language alignments.

The culmination of the sentence alignment process using the intfloat/multilingual-e5-large model and English as the pivot language was the creation of a similarity matrix. This matrix is a crucial tool in computational linguistics for representing and quantifying the semantic relationships between sentences across different languages.

The similarity matrix was constructed by calculating the semantic similarity between the embeddings of sentences from the English text and those of the corresponding sentences in each of the other languages. Each cell in the matrix represented a numerical value that quantified the degree of similarity between a pair of sentences, one from the source language and one from the English text.

The values in this matrix ranged from zero to one, where zero indicated no semantic similarity and one indicated perfect semantic congruence. These values were derived from the cosine similarity between the sentence embeddings generated by the intfloat/multilingual-e5-large model. Cosine similarity is a common measure in natural language processing used to calculate the cosine of the angle between two vectors, in this case, the sentence embeddings in the high-dimensional space. A higher cosine value indicates a smaller angle, suggesting a greater similarity.

To extract the optimal alignment from the similarity matrix, an algorithm was employed to traverse this matrix and identify the pairings of sentences that maximized the overall semantic similarity across the corpus. This step is critical as it determines the quality of the subsequent analyses and the reliability of the findings regarding syntactic and semantic features across languages.

The alignment algorithm looked for the highest similarity scores in the matrix to pair sentences from different languages. This approach ensured that each sentence in a non-English language was matched with its most semantically similar counterpart in English. Additionally, the algorithm was designed to handle instances where a sentence in the source language might align with multiple sentences in English or vice versa, a common occurrence due to structural and idiomatic differences between languages.

The advantage of working with structured texts as the Gospels and Acts is that there are some constraints for the research of the target sentence when aligning texts. For instance, a sentence in chapter 5 of the Gospel of Luke cannot be aligned to a sentence in a different book, even if semantic comparison between the two sentences gives a high result. Furthermore, given the versions of the same chapter in a certain language and in English, the mapping should follow the original order and reduce as much as

possible the shuffling among sentences. Thus, for instance, the second sentence of a chapter cannot be aligned to the 32nd sentence of its translation in English. In order to align the sentences in my corpus, I used the English translation in the MPBC as a pivot and I obtained a mapping of every text to the English version. To do so, I designed an algorithm that, given a sentence from a chapter of a certain language X, found its translation in the same chapter in the English version.

The procedure consists in taking the two translations of the same chapter (the English and the one in the target language) and, relying on the similarity matrix of the sentences in the two texts, find the best alignment. The similarity matrix is a table listing the similarity scores of each combination of sentence pairs consisting of a sentence of the first text and a sentence of the other text. The similarity score ranges from 0 (lowest similarity) to 1 (identical sentences) and corresponds to the cosine similarity between the vectors representing the two sentences.

Code 4. General scheme of the alignment algorithm.

```
# given a n x m similarity matrix

1: start ← (0;0)
2: end ← (0;0)
3: WHILE end NOT EQUAL (n-1;m-1)
4:     GET the candidate cells # the 4x4 square whose
topleft cell is start
5:     end ← the cell with the highest value among the
candidate cells
6:     FIND the best path joining start and end
7:     RECORD movements in best path
8:     start ← end
9: ENDWHILE
```

Code 4 shows in pseudocode the general scheme of the algorithm. I will now break it down and dive into the most relevant operations. Given a similarity matrix of n rows and m columns, corresponding to the sentences of a chapter in a certain language and in English, respectively, (0;0), i.e. the coordinates of the top left cell in the matrix, is assigned to the variables start and end. Then, a loop starts and goes on until the value of the variable end, is not equal to (n-1;m-1), i.e. the coordinates of the bottom right cell in the matrix. Within the loop, the first operation is the one that gets the list of relevant cells where to look for the new coordinates to assign to the end variable. It focuses on the square of dimension 4, whose top left cell is the one pointed out by the coordinates in the start variable and assigns to the variable end the coordinates of the cell with the highest value. Then, it finds

the best path to reach the end cell from the start cell considering only movements which increment the row index or the column index (or both) by 1. The best path is the one which crosses the cells with the highest average similarity score.

Table 5. *Example of optimal path.*

.909	.757	.749	.758
.831	.886	.829	.835
.763	.838	.793	.796
.770	.850	.912	.890

Table 6. *Example of an alternative path.*

.909	.757	.749	.758
.831	.886	.829	.835
.763	.838	.793	.796
.770	.850	.912	.890

Table 7. *Example of another alternative path.*

.909	.757	.749	.758
.831	.886	.829	.835
.763	.838	.793	.796
.770	.850	.912	.890

In Table 5, Table 6 and Table 7, I showed three examples of possible paths to join the starting cell (in green) to the ending cell (in red). As described earlier, when considering a 4x4 square, the cell containing the highest value (excluding the starting cell from the comparison) is the one that will be considered as end of the path. The paths in Table 5, Table 6 and Table 7 all follow the rules for the movements: the movements in the first path are SE, S, SE, the movements in the second path are S, SE, SE, while those in the third are SE, E, S, S[2].Considering the average similarities in the paths, we can see that the best path among the three (and among all the possible paths) is the first one, whose average similarity is 0.879 (vs. 0.860 in the second and 0.855 in the third).

2 S, E and SE refer to 'South', 'East' and 'South-East', respectively.

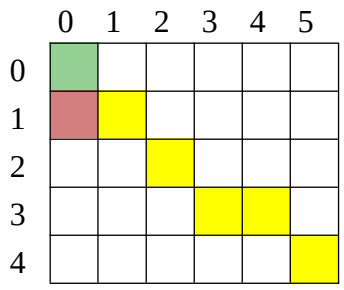
Once found the best path from the starting to the ending cell (line 6 in Code 4) and recorded all the movements (line 7 in Code 4), the ending cell becomes the new starting cell (line 8 in Code 4) and the instructions in the loop are executed again, until the matrix is completely traversed. When the movements have been all recorded, then they can be converted into a mapping between the sentences of the two translations.

Code 5. *Scheme to convert the list of movements into a mapping between the translations.*

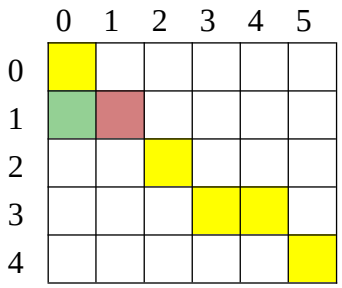
```
1:  movements ← list of movements recorded traversing
the similarity matrix
2:  coordinates ← (0;0)
3:  mapping ← {}
4:  sent1 ← []
5:  sent2 ← []
6:  FOR movement in movements
7:      IF movement CONTAINS “E” THEN
8:          ADD coordinates[column] TO sent2
9:      ENDIF
10:     IF movement CONTAINS “S” THEN
11:         ADD coordinates[row] TO sent1
12:     ENDIF
13:     IF movement EQUAL TO “SE” THEN
14:         ADD (sent1: sent2) TO mapping
15:         sent1 ← []
16:         sent2 ← []
17:     ENDIF
18:     coordinates ← GET FROM coordinates and
movements
19: ENDFOR
20: ADD coordinates[row] TO sent1
21: ADD coordinates[column] TO sent2
22: ADD (sent1: sent2) TO mapping
```

Code 5 shows the sequence of operations executed in order to obtain an alignment of two translations from the list of movements done to traverse the similarity matrix following the path that maximize the average similarity score. The mapping consists of pairs of list whose length may vary. We can find 1:1 mappings, when a sentence in text 1 corresponds to a sentence in text 2, 1:N mappings, when a sentence in text 1 corresponds to more than one sentence in text 2 (N:1 if the opposite happens), and N:M mappings, when a group of sentences in text 1 corresponds to more than one sentence in text 2.

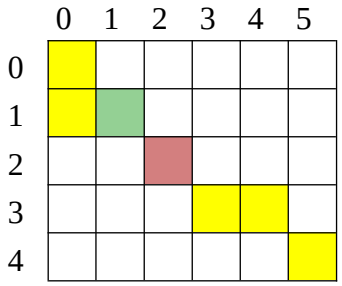
Table 8. *Example of execution of the algorithm to convert the movements list into a sentence mapping.*



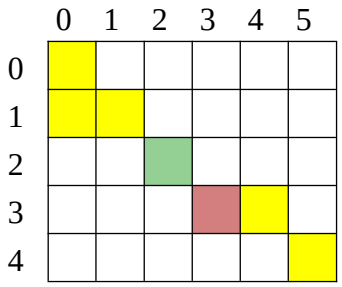

```
movements = [S, E, SE, SE, E,
SE]
coordinates = (0;0)
mapping = {}
sent1 = []
sent2 = []
after executing the commands in
the loop
sent1 = [0]
sent2 = []
mapping = {}
coordinates = (1;0)
```



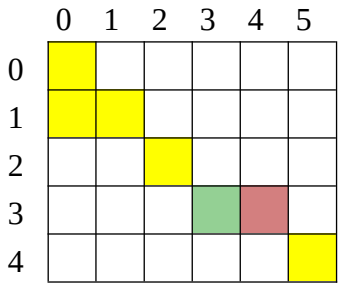

```
movements = [S, E, SE, SE, E,
SE]
coordinates = (1;0)
mapping = {}
sent1 = [0]
sent2 = []
after executing the commands in
the loop
sent1 = [0]
sent2 = [0]
mapping = {}
coordinates = (1;1)
```



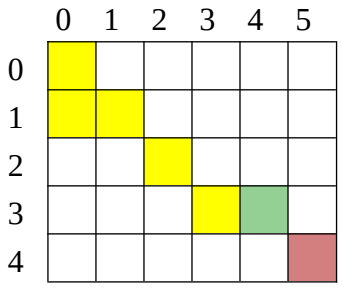

```
movements = [S, E, SE, SE, E,
SE]
coordinates = (1;1)
mapping = {}
sent1 = [0]
sent2 = [0]
after executing the commands in
the loop
sent1 = [0,1]
sent2 = [0,1]
mapping = {[0,1]: [0,1]}
sent1 = []
sent2 = []
coordinates = (2;2)
```




```
movements = [S, E, SE, SE, E,
SE]
coordinates = (2;2)
mapping = {[0,1]: [0,1]}
sent1 = []
sent2 = []
after executing the commands in
the loop
sent1 = [2]
sent2 = [2]
mapping = {[0,1]: [0,1], [2]:
[2]}
sent1 = []
sent2 = []
coordinates = (2;2)
```




```
movements = [S, E, SE, SE, E,
SE]
coordinates = (3;3)
mapping = {[0,1]: [0,1], [2]:
[2]}
sent1 = []
sent2 = []
after executing the commands in
the loop
sent1 = []
sent2 = [3]
mapping = {[0,1]: [0,1], [2]:
[2]}
coordinates = (3;4)
```




```
movements = [S, E, SE, SE, E,
SE]
coordinates = (3;4)
mapping = {[0,1]: [0,1], [2]:
[2]}
sent1 = []
sent2 = [3]
after executing the commands in
the loop
sent1 = [3]
sent2 = [3,4]
mapping = {[0,1]: [0,1], [2]:
[2], [3]: [3,4]}
sent1 = []
sent2 = []
coordinates = (4;5)
the loop ends because all the
movements have been processed
sent1 = [4]
sent2 = [5]
mapping = {[0,1]: [0,1], [2]:
[2], [3]: [3,4], [4]: [5]}
```

2.4 Semantic roles extraction

In this study, semantic roles were extracted specifically from the English text using InVeRo-XL (Di Fabio et al., 2019), a robust instrument designed

to identify and categorize the roles played by different words in a sentence, such as agents, patients, and instruments. These roles provide deeper insights into the syntactic functions and relationships inherent within the text, essential for subsequent linguistic analysis.

While InVeRo-XL is a powerful tool for semantic role labeling, its application in this study faced certain technical limitations that restricted its use to English texts only. This limitation necessitated a focused approach where semantic roles were initially extracted exclusively from the English translations.

Once the semantic roles were extracted from the English texts, they were carefully integrated into the corresponding treebanks. This integration was facilitated by adding the roles to the 'misc' field in the CoNLL-U formatted files. The 'misc' field, typically used for miscellaneous annotations not covered by the standard fields in the format, served as an ideal place to record these semantic roles, allowing for additional semantic information to be preserved without altering the core syntactic annotations. This process will be discuss more in detail in Section 2.4.2.

2.4.1 Porting of semantic roles annotation

The next challenge was to port these annotations onto the treebanks of other languages. This task was accomplished by leveraging word-to-word alignments created using the microsoft/xlm-align-base (Chi et al., 2021) model. This model, known for its effectiveness in aligning words across languages by generating robust cross-lingual embeddings, allowed for precise mapping of semantic roles from the English text to the other languages in the corpus. The alignment process ensured that the semantic roles assigned to words in English were accurately transferred to their corresponding aligned words in other languages, maintaining the integrity and relevance of the semantic role annotations across linguistic boundaries.

The microsoft/xlm-align-base model played a crucial role in this process by providing high-quality word-to-word alignments. The model's capability to understand and map linguistic features across different languages ensured that the semantic roles extracted from English could be effectively aligned with the exact counterparts in the parallel texts. This alignment was vital not only for transferring semantic information but also for maintaining the consistency of semantic analysis across the corpus. It allowed the research to extend the insights gained from the English semantic

role labeling to other languages, enriching the comparative linguistic study with a broader multilingual perspective.

One significant advantage of this approach is its potential to increase the number of languages that can be included in the sample. By using a robust alignment model like microsoft/xlm-align-base, it is possible to extend semantic role labeling to a wide range of languages, even those with limited resources or linguistic data. This inclusivity enhances the comparative analysis, providing a more comprehensive understanding of linguistic phenomena across different languages. The broader the linguistic diversity included in the study, the richer the insights that can be drawn about universal and language-specific semantic structures. Several studies have successfully utilized transfer annotation techniques to enhance multilingual corpora via projection and alignment applying these methods to universal part-of-speech tagging (Agić et al., 2015; Das & Petrov, 2011), named entity recognition (Ni et al., 2017; Schäfer et al., 2022) and semantic roles annotation (Padó & Lapata, 2009).

These methods leverage the strengths of the alignment model to generate accurate mappings, which can significantly reduce the manual effort required in annotating multiple languages. Automating the alignment

process not only saves time but also enables the handling of larger datasets, which is crucial for large-scale linguistic research.

However, these approaches also have potential drawbacks. One major concern is the introduction of alignment errors, which can propagate through the dataset and affect the quality of the annotations. Misalignments can occur due to various factors, such as syntactic differences, idiomatic expressions, or cultural context that the model may not fully capture. These errors can lead to incorrect semantic role assignments, thereby reducing the reliability of the data and potentially skewing the analysis results.

Despite these challenges, the benefits of transfer annotation based on alignments—such as increased linguistic inclusivity and reduced manual annotation effort—make it a valuable method to advance in this study.

2.4.2 Integration of semantic roles into the morphosyntactically annotated resource

After the extraction of semantic roles for which I employed the InVeRo-XL tool, as detailed in Section 2.4, I integrated the annotation into the treebanks. The output from InVeRo-XL (see Code 6) required significant post-processing to ensure compatibility with the syntactic annotations provided in the CoNLL-U format.

Several adjustments were necessary to align the InVeRo-XL output with the Universal Dependencies (UD) guidelines and the CoNLL-U format. For instance, the IDs for arguments in the InVeRo-XL output started from 0, whereas CoNLL-U format IDs begin from 1. This discrepancy required modifying the IDs to match the CoNLL-U specifications. Additionally, the tokenization used by InVeRo-XL sometimes differed from the UD guidelines, necessitating a retokenization process to ensure consistency with the syntactic annotations. Furthermore, some punctuation marks were not considered by InVeRo-XL, which could lead to misalignment issues in the annotations, so these punctuation marks had to be manually integrated into the semantic role annotations.

Code 6. Part of the annotation of a frame with its roles in JSON format.

```
{
                "tokenIndex": 8,
                "verbatlas": {
                    "frameName": "GIVE_GIFT",
                    "roles": [
                        {
                            "role": "Agent",
                            "score": 1.0,
                            "span": [
                                3,
                                4
                            ]
                        },
                        {
                            "role": "Theme",
                            "score": 1.0,
                            "span": [
                                9,
```

```
                                14
                            ]
                        }
                    ]
                }
```

Once the InVeRo-XL output was corrected, the next step was to map the annotated roles onto the treebank. The integration of semantic roles into the treebank involved assigning the frame label to the "content word" of the predicate as identified by InVeRo-XL. Each frame was assigned a unique ID to allow for direct referencing by each argument within the sentence. Tokens that were part of an argument were annotated with the corresponding role label, ensuring that every element of the argument structure was captured and labeled accurately. All this information was added in the last column of the CoNLL-U file (the misc field) using the format `Frame|Role=id`, facilitating the integration of semantic annotations with existing syntactic annotations.

Code 7. *Part of a sentence enhanced with the annotation from InVeRo-XL.*

```
13    i     i     CCONJ [...] _
14    oddał oddać VERB  [...] Frame=69:GIVE_GIFT
15    go    on    PRON  [...] Role=69:Theme
16    matce matka NOUN  [...] Role=69:Recipient
17    jego  on    PRON  [...] _
```

To further refine the semantic annotations, roles assigned to tokens corresponding to punctuation marks were removed, as these did not contribute to the semantic structure of the sentence. Additionally, it was crucial to verify that all elements belonging to the same argument formed a subtree dependent on the predicate, maintaining the structural integrity of the annotations.

Once did this, the result could be processed and the relevant information for the analysis could be retrieved. Relying on the annotation of the frames, I extracted the semantic roles from the parallel treebanks. This extraction process involved systematically identifying and recording the semantic roles associated with each frame across the different languages represented in the treebanks. To facilitate further analysis and ensure easy access to the data, all the information was stored in JSON files.

For each instance of each parallel occurrence, I stored the relevant information that was annotated in the treebank. This included details such as the frame ID, the roles and their labels, the corresponding tokens, and their annotation. By organizing the data in JSON format, I ensured that the information was both structured and accessible, allowing for efficient querying and analysis in subsequent stages of the research.

2.5 Analysis

After extracting the occurrences, I proceeded to analyze the parallel occurrences of the semantic roles. To do so, I first restricted the set to the roles of the verbs under analysis. Focusing on communication and transfer verbs, I utilized the labels assigned by InVeRo-XL to the frames and limited the selection to the roles associated with the following labels: `ASK_REQUEST`, `AFFIRM`, `ORDER`, `SPEAK`, `DECREE_DECLARE`, `REFER`, `ANSWER`, and `INFORM` for communication verbs, and `BRING`, `GIVE_GIFT`, `OFFER`, `SEND`, `TAKE`, and `THROW` for transfer verbs. Additionally, the set of parallel occurrences was filtered to include only those present in all the languages in the sample.

Although the annotation of frames and roles was transferred from the English section of the parallel treebank to other languages (as described in Section 2.4.1), not all sections of the parallel treebank had all the roles and frames annotated. This discrepancy was due to occasional misalignment, where some annotations were applied to tokens not part of the subtree of the token with the frame label. After restricting the selection to the target roles, I stored the results in a table (detailed in Section 3.1), which formed the basis for the statistical analysis.

The analysis of the results consisted of two parts: the first part was a clustering analysis aimed at grouping the occurrences to identify meaningful clusters representing the semantic roles associated with the verbs under analysis. The second part was a multidimensional scaling (MDS) technique, which allowed visualization of the set of occurrences in a bidimensional space while maintaining the distances among the elements. The goal of this visualization was to interpret the dimensions by assigning them linguistic meanings and identifying patterns that might otherwise remain undiscovered.

Both analyses relied on a distance matrix, which collected the distances between all possible pairs of elements in the set of results. As a measure of distance between the elements, I used the Hamming distance, which counts the differing elements in strings of equal length. Detailed descriptions of these analyses are provided in Section 3.

3 Results and Discussion

In this chapter, I present the findings derived from the statistical analysis conducted on the occurrences extracted from the parallel corpus. The collection of parallel occurrences comprises elements tagged with the same role across all languages included in the corpus. This uniformity is crucial in constructing the results table and the distance matrix constructed upon the results, as elaborated in Section 3.1. Subsequently, in Section 3.2 I discuss the clustering methodologies employed to group the roles of the analyzed verbs. Finally, in Section 3.3, I show the outcomes of the multidimensional scaling (MDS) technique and elaborate on the significance of the relevant dimensions .

3.1 Construction of the results table

Before analyzing the results, I constructed a table storing the relevant data for my analysis of each of the parallel occurrences extracted from the

corpus. This result table was used to build the distance matrix whose construction process will be described later in Section 3.1.1.

As shown in Section 2.4.2, the role annotation generated by InVeRo-XL was refined to ensure its reliability. Initially, I excluded role annotations for all tokens lying outside the subtree dependent on the verb annotated with the frame label associated with the role. Subsequently, I identified the root of the subtree and retained the role annotation exclusively on that token. This was made to facilitate the transfer of annotation from the English section to the other sections of the parallel corpus, as discussed in Section 2.4.1.

Table 9. *Results table.*

Role id	Lang1.adp	Lang1.case	Lang1.position	Lang2.adp	Lang2.case	Lang2.position	...
0	None	None	before	None	Nom	before	...
1	None	None	before	None	Nom	after	...
2	to	None	after	None	Acc	after	...
...	...	...	...	...	...	...	...

The extraction of the parallel occurrences of the roles resulted in 363 elements. In order to analyze them, I constructed a comprehensive table containing data for each occurrence across the languages in the corpus. This data included the features retrievable from a treebank that languages employ to encode semantic roles, as discussed by Luraghi and Narrog (2014, pp. 3–

7) and discussed in Section 1.1.3. For each instance of a role in a language, I stored:

- the presence or absence of an adposition (along with its specific form when applicable) dependent from the token annotated with the role label;
- the grammatical case of the token annotated with the role label (where available);
- its positional relation to the verb.

With a sample size of 19 languages, my table encompassed a total of 58 columns (3 for each language and a column to store the id of the parallel occurrence).

Before proceeding with the quantitative analysis, I manually examined each of the 363 parallel occurrences to verify the accuracy of the alignment and the subsequent transfer of annotations to other sections of the parallel treebank. During this process, I also corrected any role annotations provided by the InVeRo-XL tool as needed. Following this manual review, I excluded 65 occurrences from my sample due to erroneous alignments. Most of these excluded occurrences involved roles tagged as agents. This inconsistency can be attributed to the linguistic characteristics of certain languages in the

sample, which, unlike English, are pro-drop languages. An example of the type of instances excluded from the sample is provided in (3):

(3) Luke 14:22

(a) English:

"Lord, it is done as you commanded, and there is still room."

(b) Italian:

"Signore,	*è*	*stato*	*fatto*	*come*	
Lord.VOC	be.3SG	be.PTCP.PASS	do.PTCP.PASS	as	
hai	*ordinato,*	*ma*	*c'è*	*ancora*	
have.2SG	order.PTCP	but	exist.3SG	still	
posto."					
place					

(c) Spanish:

"Señor,	*se*	*ha*	*hecho*	*lo*	*que*
Lord.VOC	it.REFL	have.3SG	do.PTCP	what	that
mandaste,	*y*	*aún*	*queda*	*lugar."*	
order.2SG.PST	and	still	remain.3SG	place	

(d) Finnish:

"Herra,	*on*	*tehty,*	*minkä*	**käskit**,	*ja*
Lord.VOC	be.3SG	do.PTCP	what	order.2SG.PST	and
vielä	*on*	*tilaa."*			
still	be.3SG	place			

In the four parallel occurrences shown in (3), I have formatted in bold the token tagged with the role, whose annotation was transferred to the corresponding parallel tokens. It can be observed that the annotation of the token “you”, tagged as Agent in the English sentence, was incorrectly transferred to the parallel sentences and applied to tokens that do not accurately align with “you”. This issue arises because, unlike English, languages such as Italian, Spanish and Finnish, which have verbal morphology that marks the person of the subject, can omit explicit subjects.

In these cases, the annotation was transferred to the auxiliary “hai” (the second person singular form of the verb “to have” in Italian) and to the verbs “mandaste” and “käskit” (the second person singular forms of the verb “to command” in Spanish and Finnish, respectively). Although the alignment links elements that all refer to the second person singular—namely, the pronoun “you,” the auxiliary “hai”, and the verbs “mandaste” and “käskit”—this result is not entirely satisfactory for the purposes of this analysis. Consequently, occurrences like the one in (3) were excluded from the analysis and were not considered in the calculation of the distance matrix.

3.1.1 Distance matrix

To facilitate the grouping of results into clusters, I transformed the results matrix into a distance matrix, which essentially serves as a table storing the distances between each pair of records. This conversion required selecting a suitable measure capable of capturing the differences between the rows in the table. Given that each row contains an identical number of fields, the Hamming Distance (Hamming, 1950) is an appropriate metric to measure the dissimilarity among the parallel occurrences.

The Hamming Distance between two strings of equal length is defined as the count of positions where corresponding symbols differ. In the context of this study, the index quantifies discrepancies between two parallel occurrences, considering the instances where the adposition used, the retrieved case, or the role's position relative to the verb varies within the same language. For instance, if two languages in the corpus use different adpositions or if the role's position relative to the verb is distinct, these differences will increase the Hamming Distance, highlighting variations that are crucial to identify semantic role patterns across languages.

By using the Hamming Distance, I ensured that even minor but significant differences within languages were captured. Each row in the table represented the linguistic features of a parallel occurrence of a role, so

the distance between two rows indicated the number of differences among the corresponding features of two parallel occurrences. For instance, when comparing rows 1 and 2 in Table 9, the Hamming Distance between these records would amount to 3, reflecting distinct values for “lang1.adp”, “lang1.position” and “lang2.case”.

The aim of clustering these occurrences was not only to group similar items together but also about gaining insights into the underlying structure of the data. By transforming the results matrix into a distance matrix and applying clustering techniques (Section 3.2), I aimed to reveal the hidden relationships and patterns that characterize the roles of verbs in different languages. This method allowed for a deeper understanding of cross-linguistic similarities and differences, providing a foundation for further analysis and interpretation.

After computing the Hamming distance between all possible pairs of rows in the results matrix, I generated the distance matrix, a fundamental component in many clustering and similarity analysis tasks. The distance matrix is a square matrix that captures the pairwise dissimilarities or distances between elements in a dataset. One key feature of the distance matrix is its symmetry: the distance from point A to point B is the same as

the distance from point B to point A. This symmetry arises from the nature of the distances being symmetric measures of dissimilarity.

Furthermore, the diagonal of the distance matrix consists entirely of zeros. This feature indicates that the distance from an element to itself is zero, which is intuitive since an element is identical to itself and thus has no dissimilarity. In practical terms, this means that the diagonal of the distance matrix represents the "distance" from an element to itself, which is always zero.

Overall, the distance matrix provides a comprehensive and systematic representation of the dissimilarities between elements in a dataset. Its symmetry and zero-diagonal properties make it a valuable tool for various analytical tasks, including clustering (see Section 3.2), multidimensional scaling (see Section 3.3), and similarity analysis.

3.2 Clustering of the occurrences

Clustering is a fundamental technique in data analysis that involves grouping a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. In the context of this study, clustering was performed based on the distance matrix,

which quantifies the dissimilarity between pairs of parallel occurrences. This distance matrix serves as the foundation for various clustering algorithms, enabling the identification of groupings within the data.

In this work, I tested three different clustering techniques to determine the most effective method for grouping the occurrences: hierarchical clustering (Kaufman & Rousseeuw, 1990), K-means clustering (Lloyd, 1982; MacQueen, 1967), and Partitioning Around Medoids (Kaufman & Rousseeuw, 1990; PAM).

Hierarchical clustering method builds a hierarchy of clusters either by repeatedly merging smaller clusters into larger ones (agglomerative approach) or by splitting larger clusters into smaller ones (divisive approach). In this study, the agglomerative hierarchical clustering technique was used, where each observation starts as its own cluster, and pairs of clusters are merged step by step based on their distance until all observations belong to a single cluster.

K-means clustering algorithm partitions the data into a certain number of clusters, where each data point belongs to the cluster with the nearest mean. The process involves initializing the desired number of centroids, assigning each data point to the nearest centroid, and then

updating the centroids based on the mean of the assigned points. This process is repeated iteratively until convergence.

Partitioning Around Medoids (PAM) is similar to K-means, and aims to partition the data into K clusters, but instead of using means, it selects actual data points as medoids to represent the center of each cluster.

To apply the three types of clustering techniques, I employed the R package `cluster` (Maechler et al., 2023).

The number of clusters was not predefined in this study. Instead, a range of possible cluster counts, from 2 to 10, was tested for each clustering technique. This approach allows for the identification of the most suitable number of clusters that best represent the underlying structure of the data without any preconceived biases.

To evaluate the quality of the clustering results, I used two indices: the Dunn Index (Dunn, 1974) and the Davies–Bouldin Index (Davies & Bouldin, 1979) that I calculated using the R packages `cluster` (Maechler et al., 2023) and `fpc` (Hennig, 2024). The Dunn Index is used to identify clusters that are compact and well-separated. It is defined as the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance (Formula 1). A higher Dunn Index indicates better clustering quality, with well-defined, distinct clusters.

$$Dunn\ Index = \frac{min(inter-cluster\ distances)}{max(intra-cluster\ distances)}$$

Formula 1: *Formula of the Dunn Index.*

The Davies–Bouldin Index measures the average similarity ratio of each cluster with the cluster that is most similar to it. It is calculated as the average ratio of intra-cluster distances to inter-cluster distances, where lower values indicate better clustering quality.

$$Davies-Bouldin\ Index = \frac{1}{N}\sum_{i=1}^{N} \underset{j \neq i}{max}\left(\frac{\sigma_i + \sigma_j}{d_{ij}}\right)$$

Formula 2: *Davies-Bouldin Index formula.*

In Formula 2, σ_i is the average distance between each point in cluster ii and the centroid of cluster i, and d_{ij} is the distance between the centroids of clusters i and j. Lower values of the Davies–Bouldin Index indicate that clusters are compact and well-separated.

By testing a range of cluster numbers and evaluating the clustering results with both the Dunn Index and the Davies–Bouldin Index, I tried to identify the optimal clustering configuration. The indices provide complementary perspectives on clustering quality, with the Dunn Index focusing on the separation and compactness of clusters, and the Davies–Bouldin Index providing a measure of the overall clustering performance

relative to intra-cluster and inter-cluster distances. This comprehensive evaluation ensures that the chosen clustering method and the number of clusters best capture the structure of the parallel occurrences.

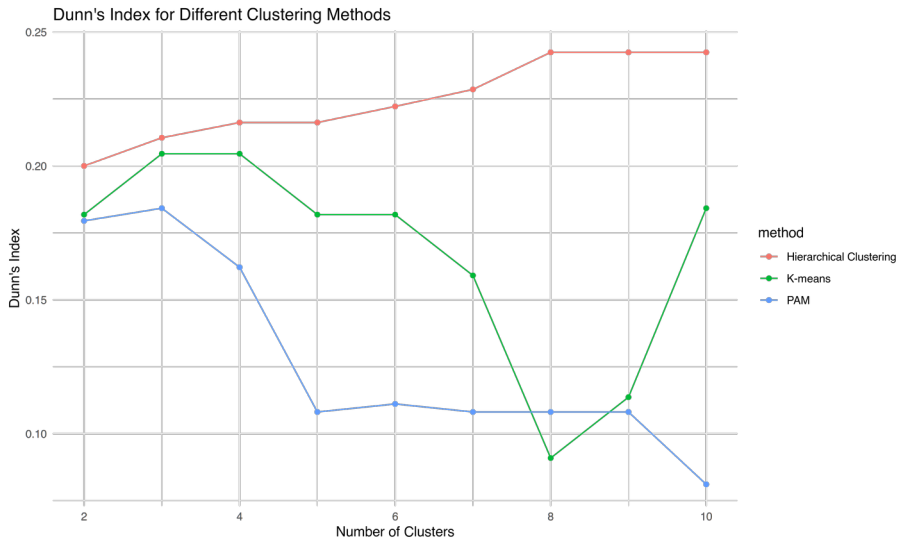


Figure 2. *Dunn Index for different clustering methods and numbers of clusters.*

The plot in Figure 2 illustrates how the Dunn Index varies with the number of clusters when using three clustering methods: hierarchical clustering, K-Means, and PAM. The Dunn Index is calculated by dividing the minimum inter-cluster distance by the maximum intra-cluster distance, so the highest value of this index is achieved when inter-cluster distance is maximized and intra-cluster distance is minimized. Because this metric

considers the extreme distances, a single poorly separated or less cohesive cluster can significantly affect the result.

The analysis of this metric suggests that K-Means clustering performs best with 3 or 4 clusters, while PAM is optimal with 3 clusters. However, the hierarchical clustering results are less clear: the Dunn Index increases monotonically, indicating that the optimal number of clusters for the sample might be 10 or even more. This increasing trend suggests that hierarchical clustering might continuously find more refined groupings as the number of clusters increases.

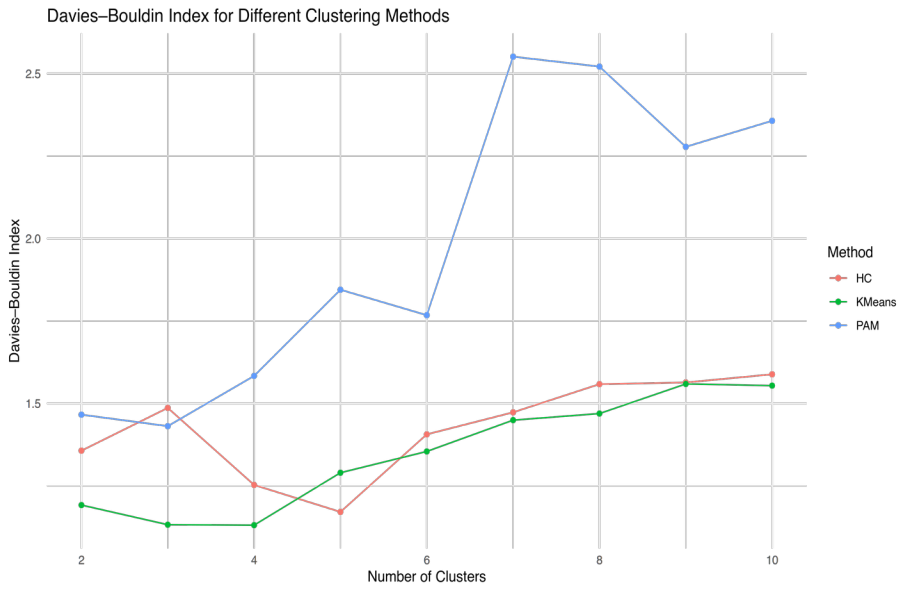


Figure 3. *Davies-Bouldin Index for different clustering methods and numbers of clusters.*

Examining the behavior of the other index used to evaluate clustering quality, the Davies-Bouldin Index, we observe in Figure 3 that the optimal number of clusters suggested is 3 and 4 for K-Means, 3 for PAM—confirming the results indicated by the Dunn Index—and 5 for hierarchical clustering. The combination of these two indices provides a strong guide for determining which clustering configurations to examine in more detail.

Specifically, I will focus on the clustering results with 3, 4, and 5 clusters across the three methods considered, analyzing their differences and evaluating their performance in effectively separating the roles of the verbs under analysis. The evaluations provided by the Dunn and Davies-Bouldin indices are internal evaluations of the clustering quality: they assess the goodness of separation based on metrics inherent to the clustering algorithm, such as the distance matrix.

By focusing on these specific cluster counts, we can gain deeper insights into which method and cluster configuration most accurately represent the underlying structure of the data and effectively differentiate the semantic roles of the verbs.

Another type of evaluation that can be made is an external evaluation of the clustering: this would involve knowing the optimal number of clusters

beforehand, based on the inherent nature of the elements in the sample. In this case, in order to evaluate the clustering results, one would compare the clusters obtained from the algorithm against a predefined set of labels or categories known to be accurate. Such an evaluation would provide additional insights beyond the internal metrics, helping to validate the effectiveness of the clustering algorithms in real-world scenarios. It would also highlight any discrepancies between the clustering results and the expected outcomes, guiding further refinement of the clustering process.

In this case, the clustering solutions with 3, 4, and 5 clusters are supported by the inherent nature of the elements in the sample. These elements represent parallel occurrences of the roles associated with transfer and saying verbs, which typically involve several core arguments: an agent, a theme and a recipient. Additionally, these verbs may include other roles that express temporal, directional, locational or other adverbial modifiers. The presence of these consistent roles across the data suggests that clustering solutions with 3, 4, and 5 clusters align well with the underlying semantic structure, capturing both the primary arguments and the potential supplementary modifiers.

In Section 3.2.1, I analyze more in detail the clusters obtained using the three methods and these three clustering solutions.

3.2.1 Analysis of the clustering

In this section, I will analyze in detail the clusters obtained using hierarchical agglomerative clustering, K-Means, and Partitioning Around Medoids (PAM). Specifically, I will focus on the results obtained with 3, 4 and 5 clusters, as these were identified as the optimal clustering solutions, as described in Section 3.2.

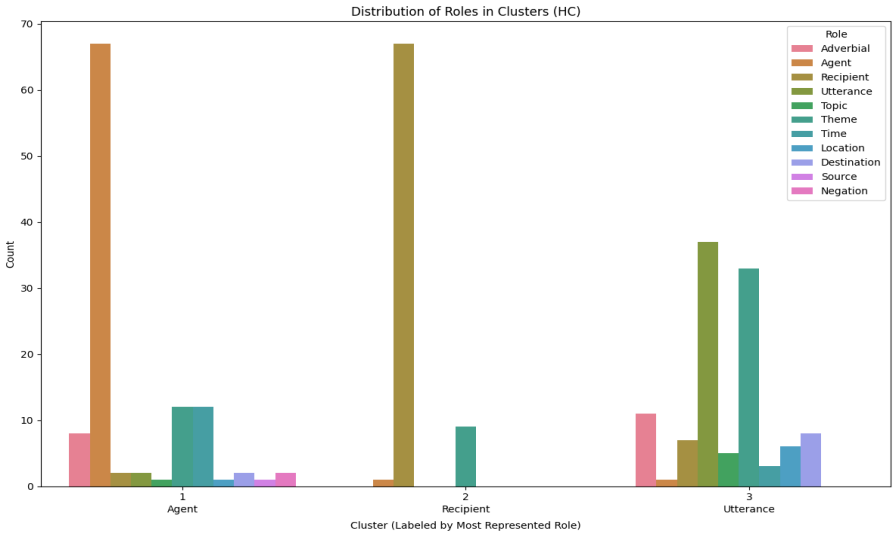


Figure 4. Distribution of the roles in three clusters (hierarchical clustering).

Figure 4 shows the distribution of roles in the clusters obtained using hierarchical clustering with three groups as the desired output. On the

horizontal axis, each colored bar represents a role, and each group of bars corresponds to a cluster. Clusters are numbered, and to aid in visualization and interpretation, each cluster number is accompanied by the label of the most represented role within that cluster. Comparing the results, we observe that the three methods—hierarchical clustering, K-Means, and PAM—split the sample of parallel occurrences into groups where the most represented roles are the same (see also Figure 19 and Figure 20 in Appendix A). The roles first separated by the algorithms are the agents (as in (4)), recipients (as in (5)), and utterances of the verbs of saying (as in (6)).

(4) Luke, 9:10-11

(a) English:

*The **apostles**, when they had returned, told him what things they had done.*

(b) Greek:

Καὶ ὑποστρέψαντες οἱ ἀπόστολοι, διηγήσαντο πρὸς αὐτὸν ὅσα ἐπραξαν.

Kaì	*hypostrépsantes*	*hoi*	*apóstoloi,*
And	return.AOR.PTCP.M.NOM.PL	the	apostle.PL
diēgḗsanto	*pròs*	*autòn*	*hósa*
relate.AOR.MID.3PL	to	he.ACC	what.ACC.PL
épraxan.			
do.AOR.ACT.3PL			

(c) Basque:

Eta	*itzuliric*	*Apostoluéc*	*conta*
And	return.PTCP	apostle.PL	tell.INF
cietzoten		*Iesusi,*	*eguin*
tell-him.3PL.ABS.3SG.DAT		Jesus.DAT	do.PTCP
cituzten	*gauça*	*guciac.*	
do.3PL.ABS	thing.PL	all.PL	

(d) Turkish:

Elçiler *geri* *dönünce,* *yaptıkları* *her*

Apostle.PL back return.COND do.PTCP.3PL every

şeyi *İsa'ya* *anlattılar.*

thing.ACC Jesus.DAT tell.3PL.PST

(5) John, chapter 19:10-11

(a) English:

"You would have no power at all against me, unless it were given to ***you*** *from above."*

(b) Polish:

"Nie *miałbyś* *żadnej* *mocy* *nade*

Not have.2SG.COND any.F.GEN.SG power over

mną, *jeźliby* *ci* *nie* *była*

me if you.2SG.DAT not be.3SG.PST

dana *z* *góry."*

give.PTCP.PASS.SG from above.GEN.SG

(c) Hebrew:

לא	*היתה*	*לך*	*רשות*	*עלי*
Lo	*hayetah*	*lekha*	*reshut*	*alay*
Not	be.3SG.PST	you.SG.DAT	authority	me.DAT
luleh	*natan*	*lekha*	*mil'ma'alah*	
if.NEG	give.3SG.PST	you.SG.DAT	from above	

(d) Indonesian:

Engkau	*tidak*	*mempunyai*	*kuasa*	*apapun*
You.NOM	not	have.2SG	power	any
terhadap	*Aku,*	*jikalau kuasa*	*itu*	*tidak*
Against	me	if power	that	not
diberikan		*kepadamu*	*dari*	*atas.*
give.PTCP.PASS		you.DAT	from	above

(6) Mark, 12:13

(a) English:

*But he, knowing their hypocrisy, said to them: "**Why do you test me?**"*

(b) Italian:

Ma	*egli,*	*conoscendo*	*la*	*loro*	*ipocrisia,*
but	he	know.PTCP	the.F.SG	their	hypocrisy
disse:		*“Perché*	*mi*	*tentate?”*	
say.3SG.PST		why	me.ACC	tempt.2PL.PRS	

(c) Lithuanian:

Žinodamas	*jų*	*veidmainystę,*	*Jis*	*tarė:*
know.PTCP	their	hypocrisy	he	say.3SG.PST
“Kam	*spendžiate*	*man*	*pinkles?”*	
why	set.2PL.PRS	me.DAT	trap.ACC.PL	

(d) Hungarian:

Õ	*pedig*	*ismervén*	*az*	*õ*	*képmutatásukat,*
he	but	know.PTCP	the	his/her	hypocrisy.ACC.SG
monda		*nékik:*	*“Mit*	*kísértetek*	*engem?”*
say.3SG.PST		they.DAT	why	tempt.2PL.PST	me.ACC

At first glance, the bar charts reveal that the occurrences of these three roles—agents, recipients, and utterances—are almost absent from the other groups, indicating that these roles are well separated from the others. However, the clusters containing these roles also include occurrences of other roles. This suggests that setting the number of clusters to three may

not be the optimal solution for any of the clustering algorithms tested. This observation is evident in the cluster with the most utterances: in all three plots (Figure 4, and Figure 19 and Figure 20 in the Appendix A), there are also around 35 occurrences of themes in these groups.

The optimal situation would be where each cluster contains occurrences of only one role, ensuring a clear and distinct separation of roles. To assess how far we are from this ideal scenario, we will measure the entropy of the clusters. Entropy (Shannon, 1948) quantifies the uncertainty or randomness within the clusters; lower entropy values indicate that the clusters are more homogeneous and well-separated. By calculating the entropy for each cluster configuration, we can determine the degree of mixing of different roles within the clusters and identify which clustering method and number of clusters provide the most distinct and well-defined groupings. This analysis will help us evaluate the effectiveness of the clustering algorithms in achieving the desired separation of roles. To facilitate comparison across clusters of different sizes, I employ normalized entropy, which scales the entropy values to a range between 0 and 1. Normalized entropy measures the degree of randomness or disorder within a cluster, with lower values indicating more homogeneity (i.e., the cluster contains instances that are more similar to each other).

To calculate the normalized entropy for each cluster, we first need to determine the theoretical maximum entropy (H_{max}), which occurs when the occurrences of the roles are uniformly distributed, meaning each role is equally represented in the cluster. Next, we calculate the entropy for a cluster (H_C) and divide it by the maximum entropy. A normalized entropy of 0 indicates perfect homogeneity (all occurrences in the cluster belong to the same role), while a normalized entropy of 1, which occurs when H_C equals H_{max}, indicates maximum disorder (the occurrences of the roles are uniformly distributed within the cluster).

Table 10. *Normalized entropies of the three clusterings (3 clusters).*

	Hierarchical	K-Means	PAM
Agent	0.5820	0.5744	0.5734
Recipient	0.3899	0.6022	0.3609
Utterance	0.7997	0.7286	0.8492

Table 10 shows the normalized entropies of the three clusters obtained using the three algorithms (hierarchical clustering, K-Means and PAM). In all the three tests, the cluster showing the highest value of entropy is the one with most utterances. This result is not unexpected: as I mentioned earlier, in the three settings, we see around 35 occurrences of themes in the cluster with more utterances. The two roles that are better

separated are the agents and the recipients, in particular when using hierarchical clustering and PAM algorithms.

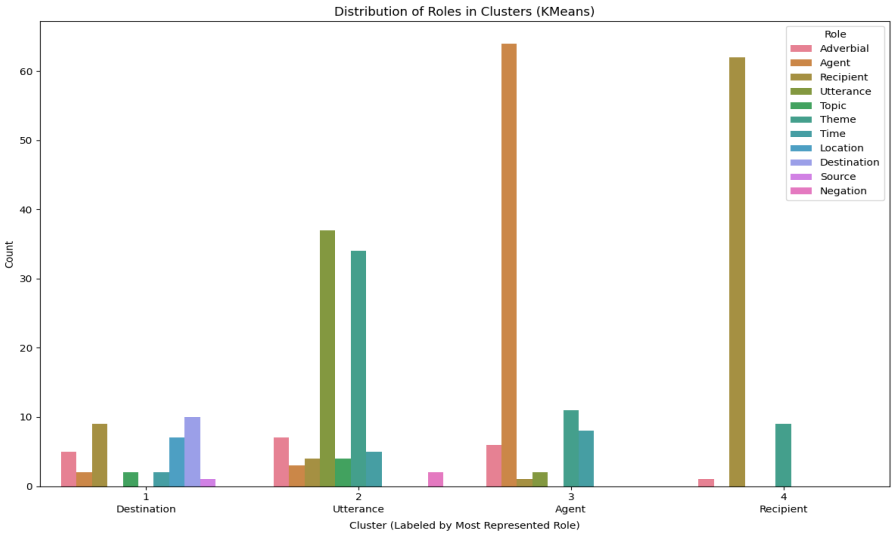


Figure 5. *Distribution of the roles in four clusters (K-Means).*

Setting a higher number of desired clusters appears to improve the performance of the K-Means and PAM algorithms. Both algorithms introduce a cluster with a new predominant role, and as shown in Figure 5 and Figure 6, the clusters identified in the previous analysis appear more distinct and better defined. Regarding the clusters obtained from hierarchical clustering, an additional cluster is formed by extracting occurrences from other groups. However, this new cluster contains only two elements, and the other clusters remain largely unchanged from the previous

configuration (see Figure 21 in Appendix A). Given the unsatisfactory results of the hierarchical clustering with 4 clusters, I will focus only on the results obtained using the other two methods.

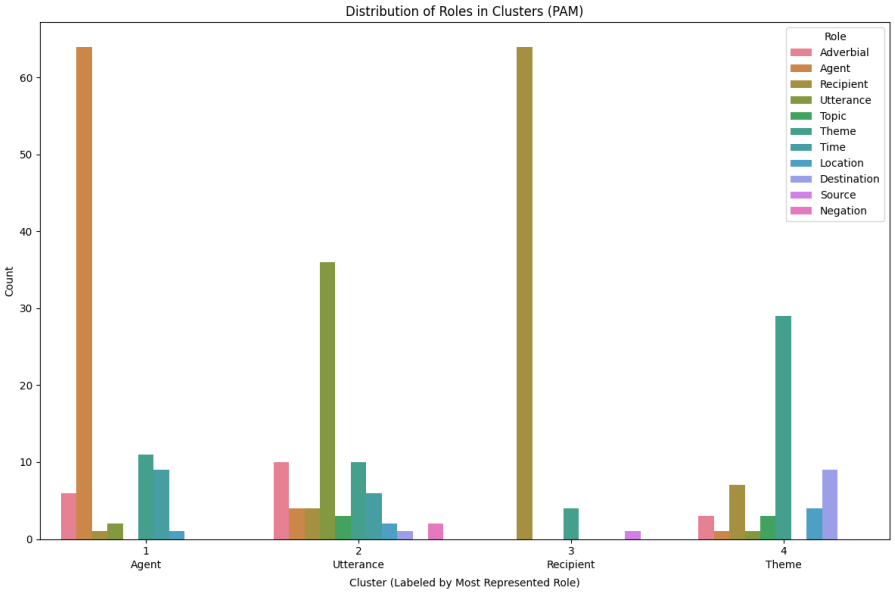


Figure 6. *Distribution of the roles in four clusters (PAM).*

(7) Acts, 9:8

(a) English:

They led him by the hand, and brought him ***into Damascus****.*

(b) Japanese:

そこで人々は、彼の手を引いてダマスコへ連れて行った。

Soko de	*hitobito*	*wa,*	*kare no*	*te*	*o*
There	people.PL	TOP	he	POSS hand	ACC
hiite	*Damasuko*	*e*	*tsurete itta.*		
pull.PTCP	Damascus	to.ALL	take.3SG.PST		

(c) Russian:

И	*повели*	*его*	*за*	*руки,*	*и*	*привели*	*в* *Дамаск.*
I	*poveli*	*ego*	*za*	*ruki,*	*i*		
And	lead.3PL.PST	he.ACC	by	hand.PL.ACC	and		
priveli	*v*	*Damask.*					
bring.3PL.PST	in	Damascus					

(d) Turkish:

Sonra	*kendisini*	*elinden*	*tutup*
Then	he.REFL.ACC	hand.ABL	hold.PTCP
Şam'a	*götürdüler.*		
Damascus.all	take.3PL.PST		

The new clusters obtained by the two algorithms predominantly feature different roles. The K-Means algorithm introduces a new cluster primarily consisting of destinations (see example (7)), while the PAM algorithm's new cluster has a majority of themes (as in example (8)).

(8) Acts, 11:1

(a) English:

Now the apostles and the brothers who were in Judea heard that the Gentiles had also received ***the word of God****.*

(b) Farsi:

پس رسولان و برادرانی که در یهودیه بودند، شنیدند که امت‌ها نیز کلام خدارا پذیرفته‌اند.

Pas	*rasulān*	*va*	*barādarāni*	*ke*	*dar*
So	apostle.PL	and	brother.PL	who	in
Yahudiyye	*budand,*		*shenidand*	*ke*	*ummat-hā*
Judea	be.3PL.PST		hear.3PL.PST	that	nation.PL
niz	*kalām-e*	*khodā-rā*	*pazīrfte-and.*		
also	word	God	accept.3PL.PRF		

(c) Spanish:

Los	*apóstoles*	*y*	*los*	*hermanos*	*que*
The	apostle.PL	and	the	brother.PL	who.REL
estaban	*en*	*Judea*	*oyeron*	*que*	*también*
be.3PL.IPFV	in	Judea	hear.3PL.PST	that	also
los	*gentiles*	*habían*	*recibido*	*la*	
the	gentile.PL	have.AUX.3PL	receive.PTCP	the	
palabra	*de*	*Dios.*			
word	of	God			

(d) Polish:

I	*usłyszeli*	*Apostołowie*	*I*	*bracia,*	
And	hear.3PL.PST	apostle.PL	and	brother.PL	
którzy	*byli*	*w*	*Judzkiej*	*ziemi,*	
who.REL	be.3PL.PST	in	Judean.LOC	land.LOC	
że	*I*	*poganie*	*przyjęli*	*słowo*	*Boże.*
that	also	gentile.PL	accept.3PL.PSTword		God.GEN

For the case of K-Means, this can be explained by examining the entropies of the clusters in the previous configuration. Although the three algorithms identified the same types of clusters, their compositions varied.

The recipient cluster of the K-Means algorithm was more diverse compared to the corresponding groups of the other two algorithms. As shown in Table 11, with the addition of the new cluster and the redistribution of destinations from other groups, the normalized entropy of the recipient cluster drops significantly (0.4079 vs. 0.6022).

Table 11. *Normalized entropies of the two clusterings (4 clusters).*

	K-Means	PAM
Agent	0.5744 (≈)	0.5610 (≈)
Recipient	0.4079 (↓)	0.2696 (↓)
Utterance	0.7376 (≈)	0.7620 (↓)
Destination	0.8808 (new)	Not applicable
Theme	Not applicable	0.7363 (new)

The same observation applies to the introduction of the new cluster in the PAM algorithm, which predominantly consists of themes. By comparing the normalized entropy values in Table 10 and Table 11, we see that the composition of this new group positively impacts the entropies of the recipient and utterance clusters, while leaving the entropy of the agent cluster unchanged. This suggests that the new theme cluster gathers elements from the recipient and utterance groups, thereby increasing their purity.

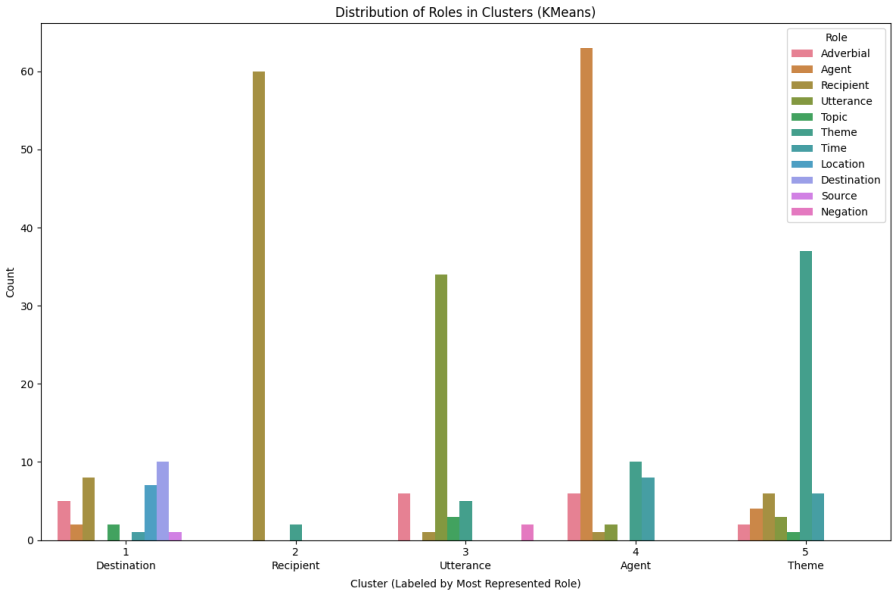


Figure 7. *Distribution of the roles in five clusters (K-Means).*

When adding another cluster, both K-Means and PAM continue to provide meaningful distributions. As shown in Figure 7 and Figure 8, the previously existing clusters become more refined, with reduced noise in the data. The newly added clusters contribute valuable information to the analysis. Specifically, the two new clusters each predominantly gather occurrences of different roles: the K-Means algorithm creates a new group mainly consisting of themes, while the PAM algorithm forms a new group primarily composed of time adverbial roles.

(9) Matthew, 17:9

(a) English:

As they were coming down from the mountain, Jesus commanded them, saying: "Don't tell anyone what you saw, until the Son of Man has risen from the dead".

(b) Italian:

E	*mentre*	*discendevano*		*dal*	*monte,*	
And	while	come-down.3PL.IPFV		from the	mountain	
Gesù	*ordinò*	*loro:*	*"Non*	*parlate*		*a*
Jesus	order.3SG.PST	them	NEG	speak.2PL.IMP		to
nessuno	*di*	*questa*	*visione,*	*finché*	*il*	*Figlio*
no one	of	this	vision	until	the	Son
dell'	*uomo*	*non*	*sia*	*risorto*		
of the	man	NEG	be.3SG.SBJ.PRS	rise.PTCP.PST		
dai	*morti".*					
from the	dead.PL					

(c) Greek:

Καὶ ἐνῶ κατέβαινον ἀπὸ τοῦ ὄρους, παρήγγειλεν εἰς αὐτοὺς ὁ Ἰησοῦς, λέγων: "Μὴ εἴπητε πρὸς μηδένα τὸ ὅραμα, ἕως οὗ ὁ Υἱὸς τοῦ ἀνθρώπου ἀναστῇ ἐκ νεκρῶν".

Kai	*enō*	*katebainon*	*apo*	*tou*	
And	while	come-down.3PL.IPFV	from	the.GEN.SG	
orous,	*parēngeilen*	*eis*	*autous*	*ho*	
mountain.GEN.SG	order.3SG.AOR	to	they.ACC	the	
Iēsous,	*legōn:*	*"Mē eipēte*	*pros*		
Jesus.NOM	say.PTCP.PRS	neg	say.2PL.AOR	to	
mēdena	*to*	*horama,*	*heōs*	*hou*	*ho*
anyone.ACC	the	vision.ACC	until	REL	the
Huios	*tou*	*anthrōpou*	*anastēi*		
Son.NOM.SG	the.GEN	man.GEN.SG	rise.3SG.AOR.PASS		
ek	*nekrōn."*				
from	dead.GEN.PL				

(d) Lithuanian:

Besileidžiant	*nuo*	*kalno,*	*Jėzus*
Descend.PTCP.PRS	from	mountain.GEN.SG	Jesus.NOM
jiems	*įsakė:*	*“Niekam*	*nepasakokite*
they.DAT	order.3SG.PST	no one.DAT	tell.2PL.IMP.NEG
apie	*regėjimą,*	*kol*	*Žmogaus*	*Sūnus*
about	vision.ACC.SG	until	Man.GEN.SG	Son.NOM.SG
prisikels	*iš*	*numirusių”.*		
rise.3SG.FUT	from	dead.GEN.PL		

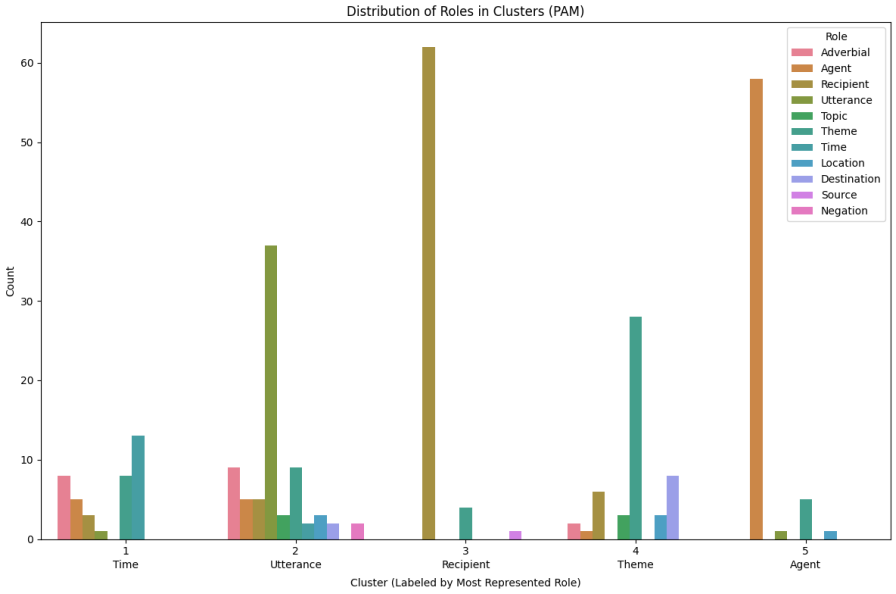


Figure 8. Distribution of the roles in five clusters (PAM).

Table 12 shows that the addition of a new cluster brings benefits in terms of reduced entropy for the already identified groups. Notably, there is a significant drop in the entropy value of the clusters where the majority of elements are recipients and utterances when applying the K-Means algorithm, which groups together occurrences of themes. A similar improvement is observed with the PAM algorithm, where the entropy value of the cluster containing agents decreases when a new cluster is introduced to collect time adverbial roles.

Table 12. Normalized entropies of the two clusterings (5 clusters).

	K-Means	PAM
Agent	0.5715 (≈)	0.3083 (↓)
Recipient	0.2056 (↓)	0.2756 (≈)
Utterance	0.6254 (↓)	0.7585 (≈)
Destination	0.8670 (≈)	Not applicable
Theme	0.6554 (new)	0.7241 (≈)
Time	Not applicable	0.8852 (new)

After examining the clustering results for lusters 3, 4 and 5 c, I also analyzed the clusters obtained by setting the desired number of groups to 6. This was done primarily to determine whether the groups identified by the two algorithms converged to the same set of roles.

As shown in the bar charts in Figure 9 and Figure 10, only the K-Means algorithm successfully gathered occurrences and created a group where the most prominent role was a newly identified one.

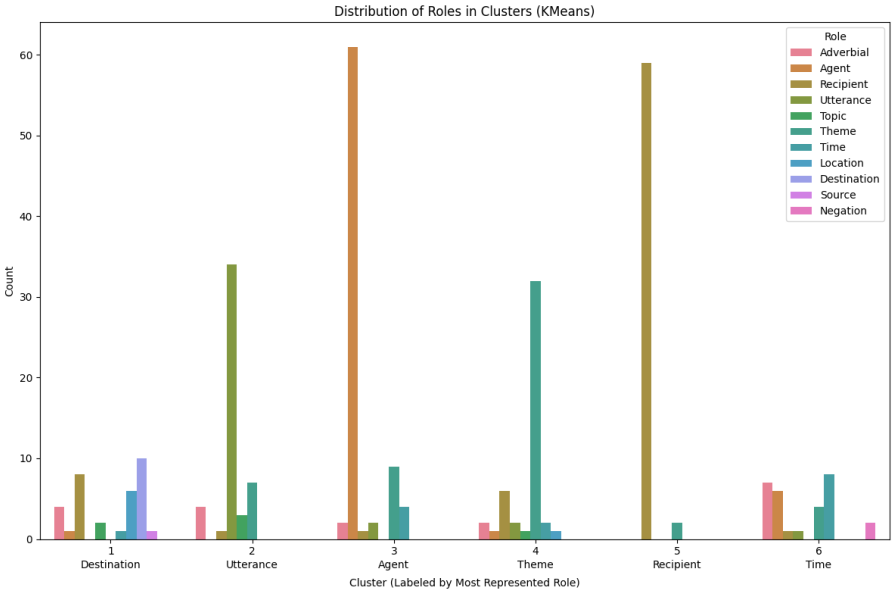


Figure 9. *Distribution of the roles in six clusters (K-Means).*

A cluster containing time adverbial roles, identified by PAM in the previous step, was also recognized by the K-Means algorithm. Conversely, the PAM algorithm grouped recipients (the most represented role in the new cluster) and destinations (the second most represented role in the new cluster) together. Although this result is not entirely satisfactory, it provides valuable insight into the nature of the data. The fact that both K-Means and

PAM cluster occurrences of recipients and destinations together suggests that these two roles are closely related. For some occurrences of recipients, it is unclear whether they should be grouped with the main recipients cluster or with destinations.

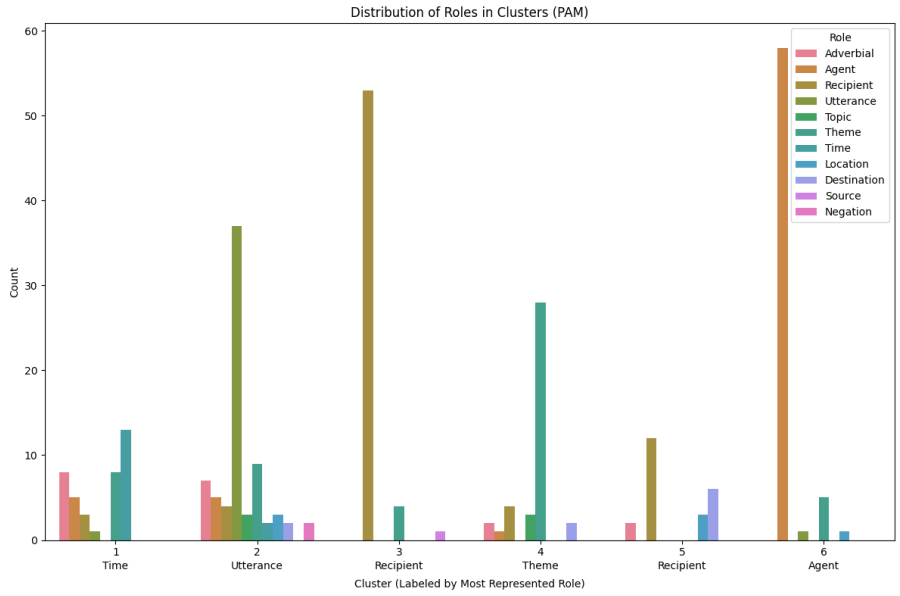


Figure 10. *Distribution of the roles in six clusters (PAM).*

To better understand this, I will now examine some examples of recipients that fall into cluster 1 in the K-Means 6-cluster test and cluster 5 in the PAM 6-cluster test.

(10) Acts, 16:13

(a) English:

and we sat down, and spoke ***to the women*** *who had come together.*

(b) Latvian:

un	*apsēdušies*	*mēs*	*runājām*	*ar*
And	sit down.PTCP.PST	we	speak.1PL.PST	with
sievietēm,	*kas*	*tur*	*bija*	
woman.DAT.PL	who	there	be.AUX.PST	
sapulcējušās.				
gather.PTCP.PST.F.PL				

(c) Spanish:

nos	*sentamos*		*allí*	*y*
we.REFL	sit down.1PL.PST		there	and
hablábamos	*a*	*las*	*mujeres*	*que*
speak.1PL.IPFV	to	the.PL	woman.PL	who
se	*habían*		*reunido.*	
they.REFL	have.AUX.3PL.PST		gather.PTCP.PST	

(d) Hungarian:

és *leülvén,* *beszélgeténk* *az*

And sit down.PTCP.PST speak.1PL.PST the

egybegyûlt *asszonyokkal.*

gather.PTCP.PST woman.PL.INS

(11)Matthew, 26:9

(a) English:

For this ointment might have been sold for much, and given ***to the poor****.*

(b) Afrikaans:

Want	*hierdie*	*salf*	*kon*	*duur*		
Because	this	ointment	have.MOD	expensively		
verkoop	*en*	*die*	*geld*	*aan*	*die*	*armes*
sell.PTCP.PST	and	the	money	to	the	poor.PL
gegee	*geword*				*het.*	
give.PTCP.PST	become.AUX.PTCP.PST				have.AUX.PST	

(c) Hebrew:

כי השמן הזה היה ראוי להמכר במחיר רב ולתתו לעניים:

Ki *ha-shemen* *ha-ze* *haya*

Because oil this be.AUX.3SG.PST

ra'ui *le-himacher* *be-mechir* *rav* *ve-le-tato*

worthy sold.INF price high give.INF

la-aniyim.

poor

(d) Turkish:

Bu *yağ* *pahalıya* *satılabilir,* *parası* *yoksullara*

This oil high price sell.3SG.POT money

verilebilirdi.

give.3SG.POT.PST

At first glance, the examples in (10) and (11) do not reveal any particular pattern: they are two instances of recipients, one from a communication verb and one from a transfer verb, which are the two classes of verbs considered in my analysis. However, upon examining the verbs for which these occurrences are recipients, we observe that, although the total number of occurrences is relatively low, communication verbs are the most represented class (see Table 13).

Table 13. Occurrences of communication and transfer verbs in the clusters.

	Communication	Transfer	Total
K-Means	8	0	8
PAM	8	4	12
Total	16	4	20

Examining the sentences where recipients of communication verbs appear, it becomes apparent that they frequently refer to groups of people (such as ‘the women’ in (10), ‘the disciples’, ‘them’). In instances where only one person is involved, it is typically someone who does not respond, as seen in (12)[3]:

3 The following sentence is: “He stretched it out, and his hand was restored as healthy as the other.”

(12) Mark, 3:5

(a) English:

When he had looked around at them with anger, being grieved at the hardening of their hearts, he said to the man, "Stretch out your hand".

(b) Finnish:

Silloin	*hän*	*katsahtaen*	*ympärilleen*	*loi*
Then	he	glance.PTCP.PRS	around	cast.3SG.PST
vihassa	*silmänsä*	*heihin,*	*murheellisena*	
in anger	his eye	at them	saddened	
heidän	*sydämensä*	*paatumuksesta,*	*ja*	*sanoi*
their	heart	hardness	and	say.3SG.PST
sille	*miehelle:*	*"Ojenna*	*kätesi".*	
to the	man	stretch.IMP.2SG	your hand	

(c) Russian:

И, воззрев на них с гневом, скорбя об ожесточении сердец их, говорит тому человеку: “протяни руку твою”.

I, vozzrev na nikh s gnevom,

And look.PTCP.PFV.PST at them with anger

skorbyá ob ozhestochenii serdets ikh,

grieve.PTCP.IPFV.PRS about hardening.LOC heart.PL.GEN their

govorit tomu cheloveku: “protiani

say.3SG.PRS that.DAT man.DAT stretch.IMP.2SG

ruku tvoiu”.

hand.ACC your.ACC

(d) Basque:

Eta hetarát inguru behaturic asserrerequin, eta

And they.ALL around look.PTCP anger.COM and

hayén bihotzeco obstinationeaz contristaturic, diotsa

their heart.GEN hardness.INS grieve.PTCP say.3SG

guiçonari: “Heda eçac eure escua”.

man.DAT stretch.IMP your hand

The fact that most of the recipients included in this cluster involve roles of communication verbs seems to support Daniel’s hypothesis (2014).

According to Daniel, from the observation of data from East Caucasian languages, these recipients can be considered instances of addressees, and for this reason they should be separated from the main cluster of proper recipients as it happens in this test. It is important to note, however, that such recipients also occur in the main recipient cluster, and the size of this subgroup is too small to draw definitive conclusions. Nevertheless, this observation is intriguing and warrants further investigation in future studies. Expanding the dataset and examining a larger number of occurrences could provide more robust evidence to validate or refute this hypothesis, potentially leading to a deeper understanding of the distinctions between different types of recipients in communication contexts.

3.2.2 Concluding notes on the clustering analysis

In this section, I summarize the key findings from the clustering analysis conducted using hierarchical agglomerative clustering, K-Means, and Partitioning Around Medoids (PAM). The primary aim was to identify the optimal number of clusters and evaluate the quality of the clustering results to effectively categorize the roles of verbs under analysis.

The clustering analysis revealed that the performance of the K-Means and PAM algorithms improved with an increased number of clusters.

Specifically, when the number of clusters was increased to 6, both algorithms produced more meaningful distributions. As demonstrated in Figure 7 and Figure 8, the newly formed clusters were more distinct, with clearer separations between the different roles. For instance, K-Means introduced a new cluster predominantly consisting of themes, while PAM created a cluster primarily composed of time adverbial roles. This refinement suggests that these additional clusters help in better capturing the nuances of semantic roles.

In contrast, the hierarchical clustering algorithm did not show significant improvements with the introduction of a fourth cluster. Although an additional cluster was formed by extracting occurrences from other groups, it contained a negligible number of elements (only 2), and the overall composition of the other clusters remained largely unchanged. This indicates that hierarchical clustering may not be as effective in identifying distinct semantic roles when compared to K-Means and PAM.

To further evaluate the effectiveness of these clustering methods, I calculated the normalized entropy for each cluster. This metric helped quantify the degree of homogeneity within each cluster. Lower entropy values indicate higher purity, meaning the cluster predominantly contains occurrences of a single role. For example, the K-Means algorithm showed a

significant reduction in the normalized entropy of the recipient cluster from 0.6022 to 0.2056 (as shown in Figure 11), highlighting an improvement in cluster purity. Similarly, PAM exhibited positive impacts on the recipient and utterance clusters' entropies (see Figure 12), though the agent cluster's entropy remained unchanged.

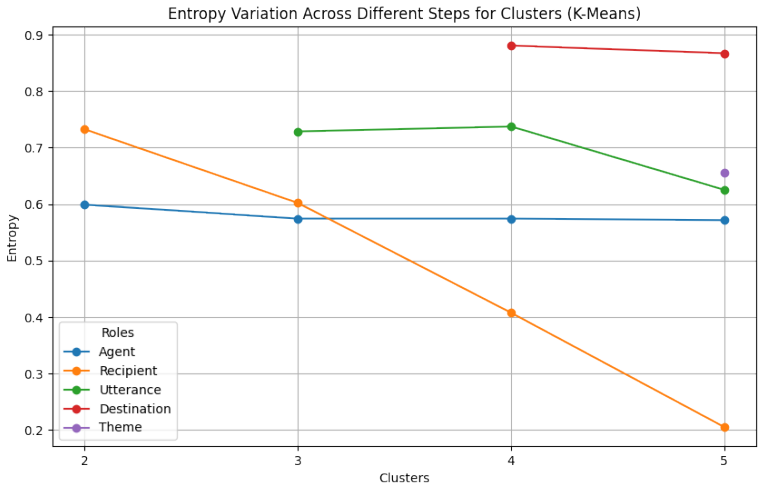


Figure 11. Clusters' entropy variation (K-Means).

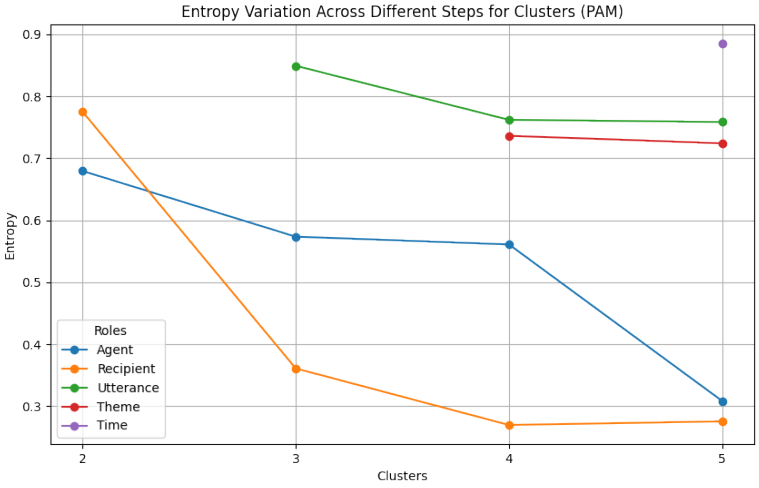


Figure 12. Clusters' entropy variation (PAM).

The clustering results suggest that setting a higher number of clusters, specifically 4 or 5, tends to provide a more accurate representation of the underlying data. This is particularly evident in the case of K-Means and PAM, where the clusters formed are more distinct and representative of specific semantic roles. Additionally, the introduction of new clusters that gather occurrences of different roles, such as themes and time adverbial roles, further validates the clustering methods' ability to capture the complexity of semantic role distribution.

In conclusion, the analysis highlights the strengths and limitations of each clustering method. While K-Means and PAM show significant

improvements with an increased number of clusters, hierarchical clustering appears less effective in creating distinct and meaningful groups.

3.3 Semantic maps

In this section, I present the semantic maps derived from performing multidimensional scaling on the distance matrix, which contains the parallel occurrences of semantic roles extracted from the parallel corpus. I will show how the elements are distributed over a two-dimensional space and whether the dimensions can be interpreted as linguistic parameters. The first step involved determining the optimal number of dimensions based on the distance matrix, which encapsulated the parallel occurrences of semantic roles. This was followed by the calculation of the coordinates for each point within the chosen dimensional space. Finally, I plotted the semantic maps, considering all possible pairs of relevant dimensions, to visualize the distribution and relationships of the semantic roles within the bidimensional space.

3.3.1 Determining the optimal number of dimensions

To determine the optimal number of dimensions for the multidimensional scaling (MDS) analysis, I employed a systematic methodology based on established practices in the field. This process involved several critical steps to ensure both the statistical robustness and interpretability of the resulting dimensions.

Initially, I calculated the stress values for various dimensional configurations. To calculate the stress and then perform the MDS, I employed the R package `smacof` (Leeuw & Mair, 2009; Mair et al., 2022). Stress value, a measure of how well the MDS configuration replicates the original distance matrix, serves as a key indicator of model fit, with lower values indicating superior representation (Kruskal, 1964).

Subsequently, I utilized the elbow method to identify the optimal number of dimensions. This technique involves plotting the stress values against the number of dimensions to generate a scree plot. The “elbow” point in this plot, where the decrease in stress values significantly slows, suggests the optimal number of dimensions (Cattell, 1966). This point represents a compromise between model complexity and fit quality.

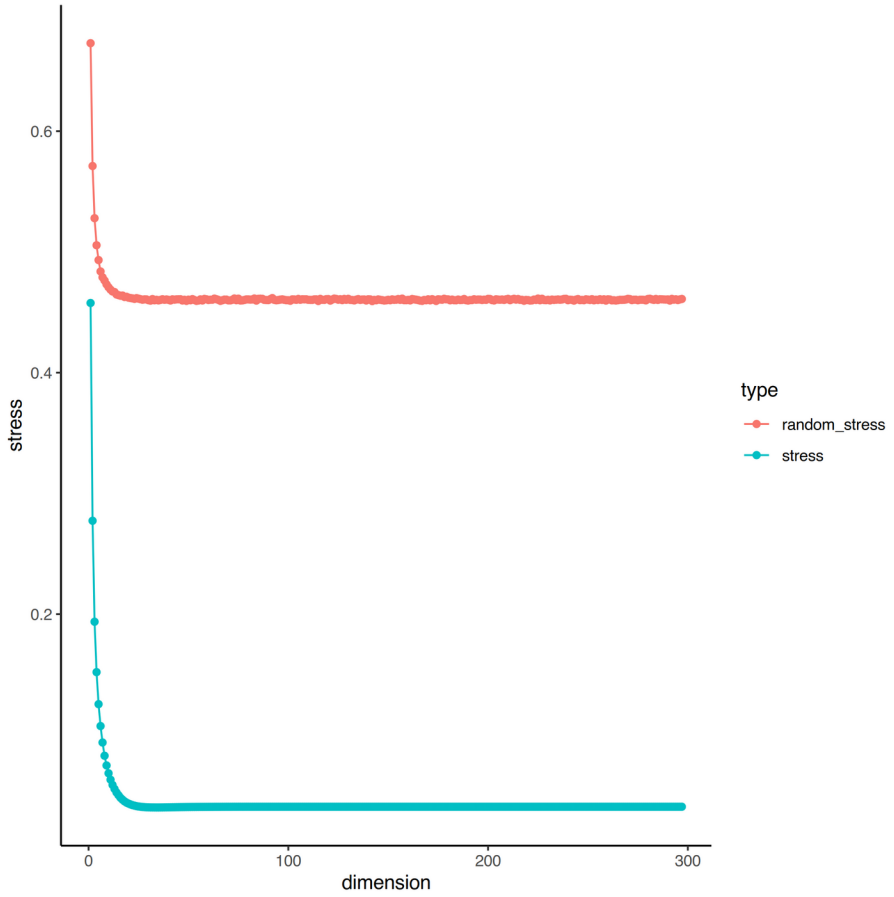


Figure 13. *Stress values and random stress values for all the possible dimensions.*

Figure 13 shows the stress values for all possible dimensions that the MDS can have, ranging from two to one dimension less than the number of parallel occurrences (i.e., 296). Although the elbow is apparent, the resolution of the plot does not allow for precise identification of the number of dimensions corresponding to the elbow. To better pinpoint this point, I

also plotted the difference in stress values between each point and the previous one. This helps identify the point at which the difference approaches zero, indicating that there is no significant difference in stress between a certain number of dimensions and the previous configuration.

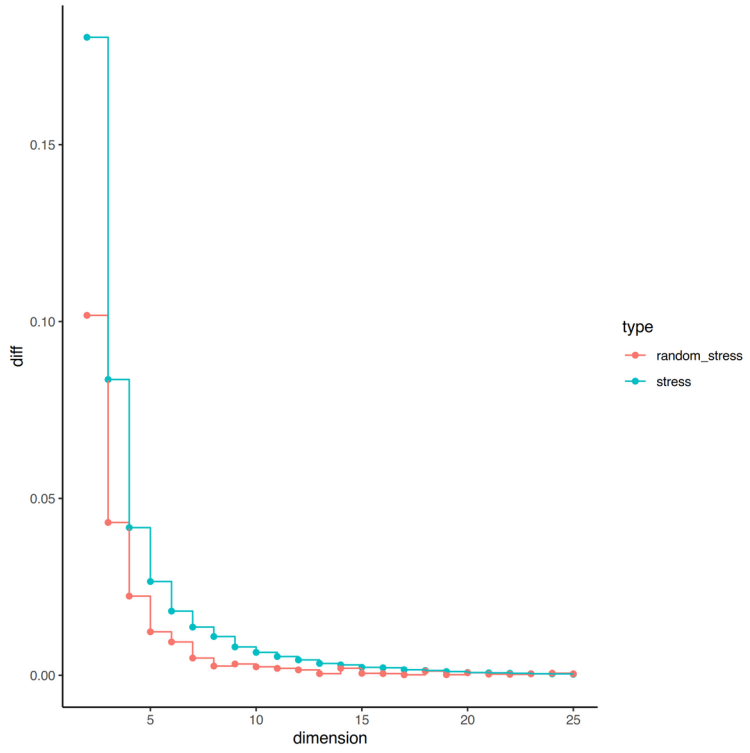


Figure 14. *Differences in stress values between consecutive dimensions*

Figure 14 shows that the stress reduction approaches zero around ten dimensions. To select the number of dimensions, I chose the point where the decrease in stress falls below 10% of the previous stress value. This

threshold of 10% is somewhat arbitrary, and a lower percentage could also have been chosen, but 10% offers a reasonable compromise, capturing significant improvements without adding unnecessary complexity.

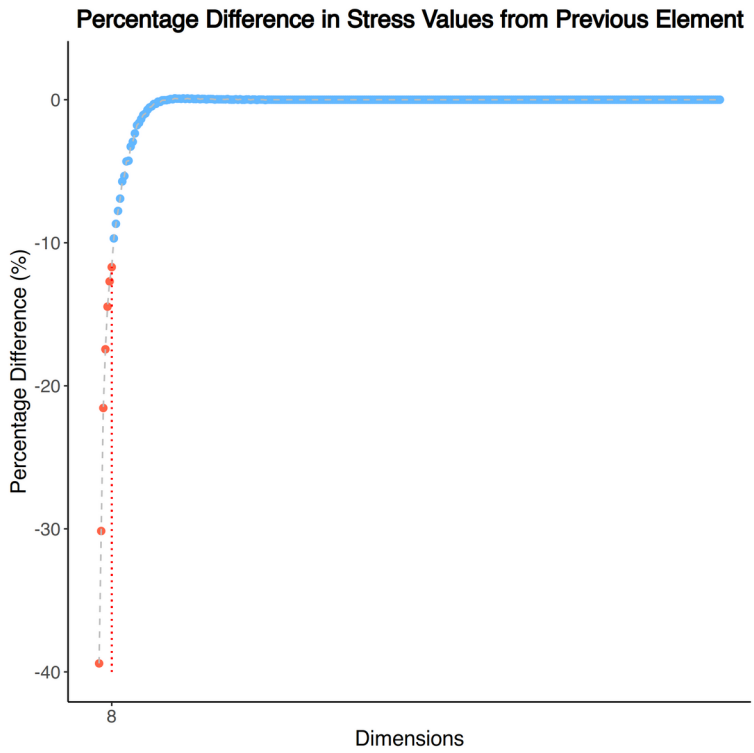


Figure 15. *Percentage difference in stress values.*

As illustrated in Figure 15, the optimal number of dimensions, based on the threshold I established, is eight. Consequently, I conducted the multidimensional scaling using this value as the target number of dimensions. To visualize the distribution of occurrences, I plotted the points

using all possible pairs of dimensions as coordinates, thereby examining their arrangement in a two-dimensional space. In Section 3.3.2, I present the relevant plots and provide a linguistic interpretation of the dimensions, explaining how they relate to the underlying semantic roles and their relationships within the parallel corpus.

3.3.2 Plots and interpretation of dimensions

In this section, I present the plots and attempt to assign linguistic interpretations to the dimensions identified as relevant in the previous steps. It's important to note that not all dimensions proved to be linguistically meaningful. Therefore, I will focus on discussing the relevant plots, while the others are provided in Appendix A for reference. The plots were generated by pairing all possible combinations of the 8 relevant dimensions, resulting in 28 distinct plots. Each plot was carefully examined to analyze the distribution of points representing the parallel occurrences of semantic roles within a two-dimensional space.

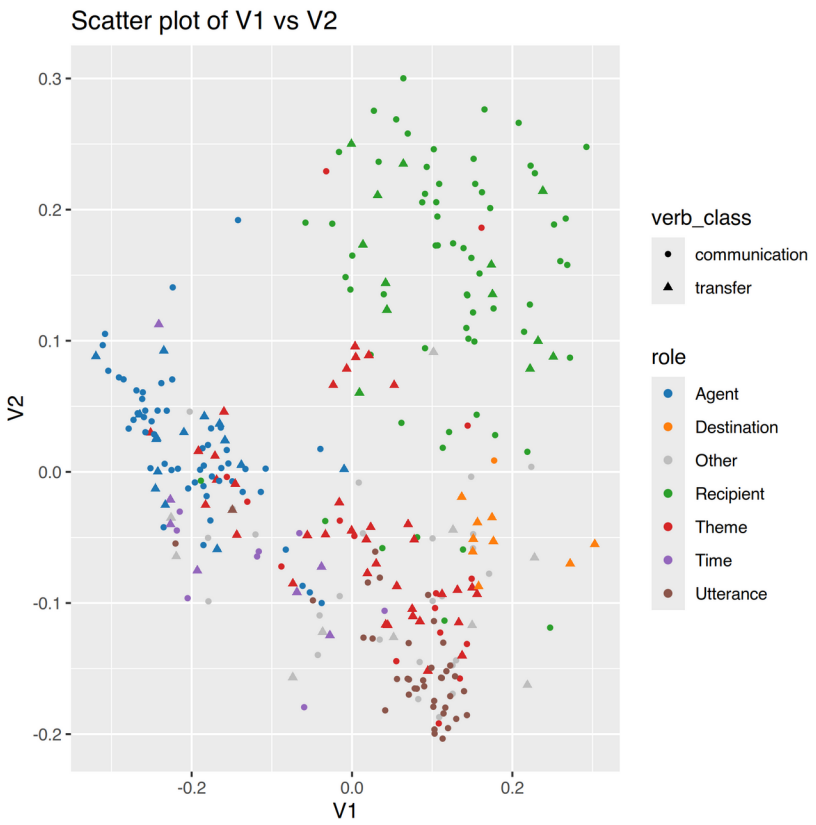


Figure 16. Distribution of the occurrences using V1 and V2 as dimensions.

One particularly notable observation comes from the plot where the dimensions V1 and V2 are used as coordinates (Figure 16). In this plot, three major groups of semantic roles are clearly delineated: recipients are concentrated in the top-right quadrant, themes and utterances are clustered at the bottom-right, and agents are predominantly located on the left side of the space. An interesting detail is the proximity of destinations to the area occupied by recipients, which aligns with the discussion in Section 3.2.1.

Additionally, some recipients appear close to points representing directions, with most of these recipients associated with communication verbs.

The spatial distribution of semantic roles across the dimensions V1 and V2 reveals meaningful insights into the underlying linguistic structure. By examining the position of the three primary groups—agents, recipients, and themes—it becomes apparent that V1, represented on the x-axis, can be interpreted as a dimension related to the control exerted by these roles over the action. Agents, who typically have the highest degree of control in an action, are positioned on the left side of the plot. In contrast, themes and recipients, which generally have less control or are more passive in the context of the action, are located towards the right. This clear separation suggests that V1 effectively captures the variation in control or agency across the semantic roles.

Turning to V2 on the y-axis, its interpretation appears to be closely related to the concept of affectedness. Affectedness refers to the extent to which a participant in an event undergoes a change or is impacted by the action. In the plot, themes and recipients are distinctly separated along this axis, reflecting their differing degrees of affectedness. Themes, which often represent entities that are directly impacted or altered by the action, are positioned lower on the plot, indicating a higher level of affectedness.

Recipients, on the other hand, are located higher up, signifying a lesser degree of affectedness, as they typically receive something but are not as fundamentally changed by the action.

When we shift our focus to the plot utilizing V1 and V3 as the dimensions (Figure 17), the separation between the three main groups—agents, recipients, and themes—is less distinct compared to the previous plot with V1 and V2. However, this plot reveals a more nuanced structure, with smaller, more specialized groups becoming more clearly delineated.

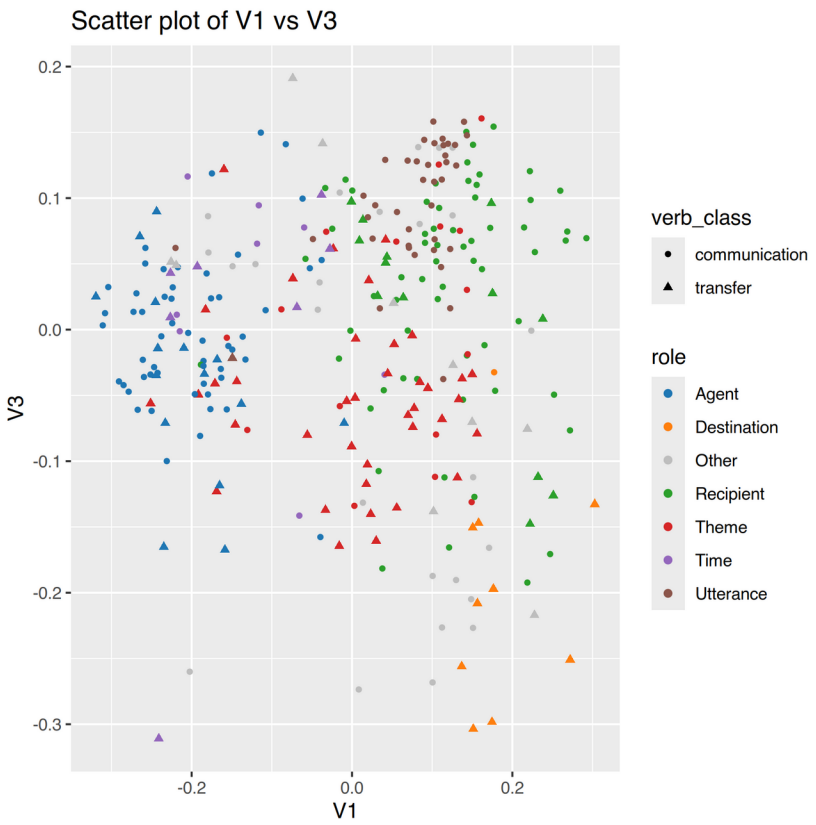


Figure 17. Distribution of the occurrences using V1 and V3 as dimensions.

Specifically, we observe that the points representing occurrences of destinations and time adverbial roles are notably distinct. Destinations cluster towards the bottom-right of the plot, while time adverbials are positioned in the top-left corner. This separation highlights the differing semantic roles these elements play within the overall structure of the data.

The clarity of this distinction is even more pronounced in Figure 18, where only the points representing destinations and time adverbials are

plotted. In this focused view, the significant distance between these two roles becomes apparent, suggesting that V3 may capture a dimension related to the temporal or locative aspects of the events.

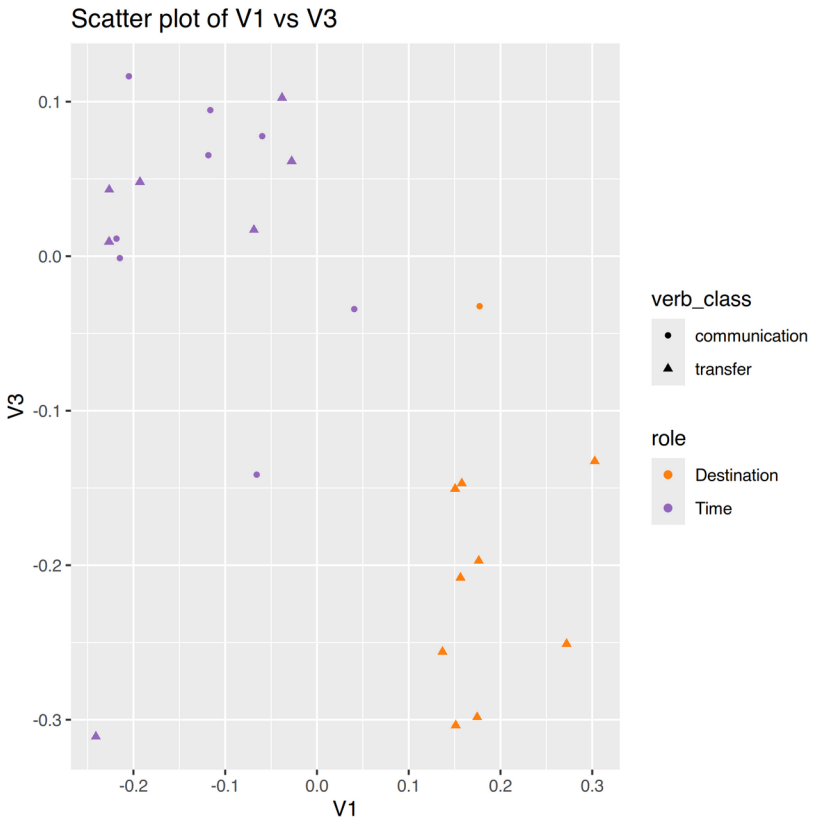


Figure 18. Distribution of the occurrences (Destination and Time) using V1 and V3 as dimensions.

V1, as previously interpreted, continues to represent control or agency, influencing the distribution along the x-axis. Meanwhile, V3 on the y-axis appears to differentiate roles based on their temporal or spatial relevance within the action. The separation of destinations and time

adverbials along this axis suggests that V3 might be capturing a dimension related to the temporal progression or spatial orientation of events. Destinations, which are more spatially oriented, are contrasted with time adverbials, which are temporally focused.

The multidimensional scaling (MDS) approach employed in this analysis has proven to be a highly effective method for visualizing and interpreting the relationships between semantic roles within a parallel corpus. By mapping these roles onto a bidimensional space, MDS allows us to uncover underlying patterns and dimensions that might not be immediately apparent through traditional analysis.

The ability to differentiate roles based on dimensions such as control, affectedness, and spatial-temporal orientation underscores the utility of this method in linguistic research. The clear separation of roles such as agents, recipients, themes, destinations, and time adverbials in the multidimensional plots demonstrates how MDS can reveal the nuanced interplay between different semantic roles, offering insights into their functional relationships within sentences.

Moreover, MDS provides a flexible framework for exploring the semantic landscape, accommodating both well-defined groups and more subtle distinctions among roles. By plotting roles in various dimensional

pairings, we can generate a comprehensive view of the data, highlighting both broad categories and specific relationships.

4 Conclusions

In this chapter, I will revisit the main objectives of the research, reflect on the implications of the findings, and discuss the limitations I encountered during the study. I will also suggest potential directions for future research that could further enhance our understanding of semantic roles and their applications in natural language processing and cross-linguistic analysis.

4.1 Summary of findings

Throughout the course of this research, I employed various analytical techniques to investigate the encoding of semantic roles across different languages, with a particular focus on verbs related to transfer and communication. Two key methods used in this analysis were clustering experiments, specifically hierarchical clustering, K-Means and Partitioning Around Medoids (PAM), as well as Multidimensional Scaling (MDS). These methods proved instrumental in uncovering patterns and relationships within the data, leading to some important findings.

First, the clustering experiments, in particular when using K-Means and PAM, were highly effective in identifying meaningful clusters within the dataset. These clusters corresponded to the semantic roles typically associated with transfer and communication verbs. Specifically, the clusters highlighted distinct groupings that align with the roles of agents, recipients and themes, as well as groupings containing utterances and adverbial roles, reflecting the underlying semantic structure of these verb types. The clarity and consistency of these clusters suggest that the clustering algorithms were successful in capturing the essential features that differentiate these roles, thereby providing a robust framework for understanding how these roles are encoded across languages.

Second, the distribution of parallel occurrences as analyzed through Multidimensional Scaling (MDS) further supported the results obtained from the clustering experiments. The MDS technique allowed for the visualization of the relationships between different linguistic instances, revealing a distribution that was both meaningful and coherent with the clustering results. This consistency between the MDS distribution and the clustering outcomes reinforces the validity of the findings, demonstrating that the semantic roles of transfer and communication verbs are represented

in a manner that is both structured and predictable across different languages.

One particularly intriguing finding emerged from both the clustering and MDS analyses: the identification of a group of recipients associated with communication verbs that appeared to occupy a position at the border between typical recipients and directions. This borderline position suggests that these instances may represent a distinct category that does not fully align with traditional recipient roles. Instead, these cases might be better understood as addressees, as suggested by Daniel (2014). This observation opens up new avenues for research, indicating that the categorization of these role could benefit from further exploration. It suggests that the traditional distinction between recipients and addressees may need to be reconsidered or refined to account for these nuanced cases, especially within the domain of communication verbs.

The findings from the clustering experiments and MDS analysis provide significant insights into the encoding of semantic roles, particularly in the context of transfer and communication verbs. The results not only confirm the existence of clear and meaningful role clusters but also highlight the potential need for a more nuanced definition of certain roles, such as addressees, which sit at the intersection of established categories.

Importantly, these observations underscore the value of a multilingual comparison in linguistic research. By examining data across different languages, I was able to identify patterns and distinctions that might otherwise go unnoticed in monolingual studies. The diversity of linguistic structures and semantic encodings revealed through this comparative approach not only enriche our understanding of how roles are defined across languages but also provides a broader perspective, which is essential for refining theoretical frameworks in semantic role theory. This underscores the critical importance of using cross-linguistic data to uncover the full complexity of how meaning is encoded and to ensure that linguistic models are robust and universally applicable.

4.2 Implications of the study

The findings of this study largely reinforce existing knowledge regarding the semantic roles associated with communication and transfer verbs. These verbs are well-known to exhibit a set of clearly defined semantic roles, such as agent, recipient and theme, which have been thoroughly studied in the field of linguistics. However, this research has also uncovered a potentially significant distinction that warrants further investigation: the difference

between recipients of transfer verbs and recipients of communication verbs. The analysis suggests that the recipients of communication verbs may be more accurately described as addressees rather than traditional recipients. This distinction, if validated by further research, could have important implications for the way we conceptualize and categorize these roles within semantic frameworks.

Should this distinction turn out to hold cross-linguistically (?), it would be necessary to reconsider how semantic roles are annotated in existing resources like FrameNet, PropBank, and other semantic role labeling tools. Specifically, the tagsets used in these resources might need to be redesigned to include a separate category for addressees, distinguishing them from recipients of transfer verbs. This would allow for a more precise and context-sensitive annotation of semantic roles, particularly in languages where this distinction is more pronounced. The refinement of these tagsets could enhance the accuracy of NLP tools that rely on these resources, leading to better performance in tasks such as machine translation, semantic parsing, and information extraction.

In addition to these theoretical implications, the methodology employed in this study offers practical insights that could be applied to future research. The approach of alignment and transfer of annotation,

especially across different languages, has proven to be effective in identifying semantic roles and their cross-linguistic variations. This methodology could be extended to include a larger and more diverse sample of languages, thereby enriching our understanding of how semantic roles are encoded across linguistic families. While the use of alignment tools and the inclusion of additional languages may introduce some degree of noise and errors, I argue that the benefits of expanding the dataset outweigh these challenges. By incorporating languages from underrepresented families, we can gain a more comprehensive and inclusive understanding of semantic roles, which is essential for developing universally applicable linguistic theories and NLP models.

To achieve this, there is a need for the development of new and improved alignment models, both for word and sentence alignment. The accuracy of these models is crucial for ensuring that the data used in cross-linguistic studies is reliable and that the annotations can be effectively transferred across languages. Investigating with better alignment tools will not only reduce the noise and errors associated with multilingual annotation but also facilitate the inclusion of a broader range of languages in future research. This, in turn, will contribute to the preservation and study of

lesser-known languages, which are often underrepresented in cross-linguistic research.

4.3 Limitations of the study

While this study offers valuable insights into the encoding of semantic roles across languages, it is important to acknowledge its limitations. These limitations primarily relate to the language sample, the reliance on computational models, the nature of the texts used, and the scope of verb classes investigated.

One of the main limitations of this study is the unbalanced nature of the language sample, which is heavily skewed towards Indo-European (IE) languages. This imbalance is largely due to the greater availability of texts, linguistic resources, and computational tools for IE languages compared to those for non-IE languages. While this focus on IE languages allowed for more consistent and reliable analysis, it also means that the findings may not fully capture the diversity of semantic role encoding across the world's languages. As a result, the conclusions drawn from this study may be less applicable to languages outside the IE family, which often exhibit different syntactic and semantic structures. A more balanced and diverse language

sample would be necessary in future studies to provide a more comprehensive understanding of semantic roles across a wider range of linguistic contexts.

Another significant limitation is the study's reliance on the outputs of various computational models, including syntactic parsers, alignment models, and semantic role labelers. While these models are essential tools in linguistic research, they are not infallible and can produce errors that affect the accuracy of the analysis. For instance, syntactic parsers may misinterpret complex sentence structures, alignment models may fail to correctly align words or phrases across languages, and semantic role labelers may incorrectly assign roles due to subtle differences in linguistic context. These inaccuracies can introduce noise into the data and potentially skew the findings. Although efforts were made to mitigate these issues, the inherent limitations of the models must be acknowledged, as they may have impacted the results of the study.

The corpus used in this study, comprising the Gospels and the Acts, is relatively small. The choice of these texts was primarily driven by practical considerations, such as their availability in multiple languages and their public domain status, which avoids potential copyright issues that could arise with other texts. However, the limited size of this corpus restricts

the breadth of linguistic data that could be analyzed. A larger corpus would provide more extensive data and could potentially reveal additional patterns and insights that were not apparent in this study. Nevertheless, expanding the corpus to include texts beyond the Bible would have required significantly more time for processing and could have introduced further challenges related to copyright and data accessibility. Despite these constraints, the selected texts provided a consistent basis for cross-linguistic comparison, albeit within a limited scope.

Finally, the set of verbs chosen for this study was deliberately limited. The focus on a specific set of well-known and well-described verbs, particularly those related to transfer and communication, was a strategic decision aimed at testing the feasibility of the methodology. This preliminary research was intended to explore the potential of the alignment and annotation transfer techniques, and thus, it was necessary to restrict the analysis to a manageable set of verbs. While this approach allowed for a focused and in-depth examination, it also meant that the study did not explore a broader range of verb classes, which could have provided a more comprehensive picture of semantic role encoding. Future research could build on this foundation by applying the methodology to a wider variety of

verbs, thereby extending the findings to cover a more diverse set of semantic roles and linguistic contexts.

4.4 Directions for future research

The findings of this study open up several promising avenues for future research. While this work has made significant contributions to our understanding of semantic roles, particularly in the context of communication and transfer verbs, there are numerous opportunities to expand and refine the research further.

One of the most pressing directions for future research is to extend the language sample beyond the predominantly Indo-European languages analyzed in this study. A more diverse language sample, including languages from underrepresented linguistic families, would provide a broader understanding of how semantic roles are encoded across different language structures. Expanding the study to include a wider range of languages will help to validate the findings and ensure that they are applicable across a variety of linguistic contexts. Additionally, it would be valuable to test the feasibility of the methodology used in this study on other texts beyond the Biblical corpus. Exploring different genres and types of texts could reveal

new insights and help determine whether the alignment and annotation transfer techniques are robust across various linguistic materials.

Future research should also consider extending the analysis to other classes of verbs, such as motion verbs. Motion verbs often involve complex semantic roles, such as source, goal, and path, which can vary significantly across languages. Investigating these roles would provide a more comprehensive understanding of how different types of actions and events are represented in language. By applying the methodology developed in this study to motion verbs and other verb classes, researchers could explore whether the patterns observed with communication and transfer verbs hold true for a broader range of semantic roles, or whether new patterns emerge that require additional theoretical refinement.

As computational models continue to evolve, there is great potential to leverage new and improved tools for future research. The current study relied on the outputs of existing models, which, while effective, have limitations that may have affected the results. It is anticipated that in the near future, more accurate syntactic parsers, alignment models, and semantic role labelers will be developed and made available to the research community. These advancements will likely reduce errors and improve the precision of linguistic analysis, making it possible to undertake even more

ambitious and comprehensive studies. Future research should aim to incorporate these cutting-edge tools, which will enhance the reliability and scope of the findings.

Finally, a crucial direction for future work is to revisit and potentially refine the semantic role annotations of the two verb classes analyzed in this study—communication and transfer verbs. In light of the distinctions suggested by this research, particularly the potential differentiation between recipients of transfer verbs and addressees of communication verbs, it may be necessary to re-examine the existing annotations. Correcting and updating these annotations will not only improve the accuracy of the current dataset but also set a more rigorous standard for future studies. This refinement process could involve a detailed review of the data, guided by the insights gained from this work, to ensure that the semantic roles are categorized in a way that reflects their true linguistic functions.

References

Agić, Ž., Hovy, D., & Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 268–272. https://doi.org/10.3115/v1/P15-2044

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics -*, *1*, 86. https://doi.org/10.3115/980845.980860

Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling*. Springer New York. https://doi.org/10.1007/0-387-28981-X

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In L. Màrquez & D. Klein (Eds.), *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)* (pp. 149–164). Association for Computational Linguistics. https://aclanthology.org/W06-2920

Chi, Z., Dong, L., Zheng, B., Huang, S., Mao, X.-L., Huang, H., & Wei, F. (2021). Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3418–3430. https://doi.org/10.18653/v1/2021.acl-long.265

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, *49*(2), 375–395. https://doi.org/10.1007/s10579-014-9287-y

Cysouw, M. (2014). Inducing semantic roles. In S. Luraghi & H. Narrog (Eds.), *Typological Studies in Language* (Vol. 106, pp. 23–68). John Benjamins Publishing Company.

Daniel, M. A. (2014). Against the addressee of speech – Recipient metaphor: Evidence from East Caucasian. In S. Luraghi & H. Narrog (Eds.), *Typological Studies in Language* (Vol. 106, pp. 205–240). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.106.07dan

Das, D., & Petrov, S. (2011). Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 600–609). Association for Computational Linguistics. https://aclanthology.org/P11-1061

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 4585–4592). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Di Fabio, A., Conia, S., & Navigli, R. (2019). VerbAtlas: A Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 627–637. https://doi.org/10.18653/v1/D19-1058

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, *67*(3), 547. https://doi.org/10.2307/415037

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95–104. https://doi.org/10.1080/01969727408546059

Fillmore, C. J. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1–88). Holt, Rinehart, and Winston.

Fillmore, C. J. (1982). Frame semantics. In L. S. of Korea (Ed.), *Cognitive Linguistics Bibliography (CogBib)*. De Gruyter Mouton. https://www.degruyter.com/database/COGBIB/entry/cogbib.3801/html

Fillmore, C. J., & Baker, C. (2012). A Frames Approach to Semantic Analysis. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (1st ed., pp. 313–340). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0013

Fürstenau, H., & Lapata, M. (2012). Semi-Supervised Semantic Role Labeling via Structural Alignment. *Computational Linguistics*, *38*(1), 135–171. https://doi.org/10.1162/COLI_a_00087

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75–102.

Georgakopoulos, T., & Polis, S. (2021). Lexical diachronic semantic maps: Mapping the evolution of time-related lexemes. *Journal of Historical Linguistics*, *11*(3), 367–420. https://doi.org/10.1075/jhl.19018.geo

Goldberg, A. (1995). A Construction Grammar Approach to Argument Structure: A Construction Grammar Approach to Argument Structure. In *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Gomes, L., & Lopes, G. P. (2016). First Steps Towards Coverage-Based Sentence Alignment. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2228–2231). European Language Resources Association (ELRA). https://aclanthology.org/L16-1354

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, *29*(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Hartmann, I., Haspelmath, M., & Taylor, B. (Eds.). (2013). *The Valency Patterns Leipzig online database*. Max Planck Institute for Evolutionary Anthropology. https://valpal.info/

Haspelmath, M. (1997). *From space to time: Temporal adverbials in the World's Languages*. LINCOM Europa.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *New Psychology of Language*, *2*, 211–242.

He, L., Lee, K., Levy, O., & Zettlemoyer, L. (2018). Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 364–369. https://doi.org/10.18653/v1/P18-2058

He, L., Lee, K., Lewis, M., & Zettlemoyer, L. (2017). Deep Semantic Role Labeling: What Works and What's Next. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 473–483. https://doi.org/10.18653/v1/P17-1044

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1607.01759

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (1st ed.). Wiley. https://doi.org/10.1002/9780470316801

Kingsbury, P., & Palmer, M. (2002). From TreeBank to PropBank. In M. González Rodríguez & C. P. Suarez Araujo (Eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/283.pdf

Kipper Schuler, K. (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon* [Ph.D. thesis]. University of Pennsylvani.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In S. Ananiadou (Ed.), *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Association for Computational Linguistics. https://aclanthology.org/P07-2045

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Lang, J., & Lapata, M. (2010). Unsupervised Induction of Semantic Roles. In R. Kaplan, J. Burstein, M. Harper, & G. Penn (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 939–947). Association for Computational Linguistics. https://aclanthology.org/N10-1137

Lang, J., & Lapata, M. (2011). Unsupervised Semantic Role Induction via Split-Merge Clustering. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1117–1126). Association for Computational Linguistics. https://aclanthology.org/P11-1112

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Li, P., Sun, M., & Xue, P. (2010). Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. In C.-R. Huang & D. Jurafsky (Eds.), *Coling 2010: Posters* (pp. 710–718). Coling 2010 Organizing Committee. https://aclanthology.org/C10-2081

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Luraghi, S. (2014). Plotting diachronic semantic maps: The role of metaphors. In S. Luraghi & H. Narrog (Eds.), *Typological Studies in Language* (Vol. 106, pp. 99–150). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.106.04lur

Luraghi, S., & Narrog, H. (2014). Perspectives on semantic roles: An introduction. In S. Luraghi & H. Narrog (Eds.), *Typological Studies in Language* (Vol. 106, pp. 1–22). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.106.01lur

Luraghi, S., Palmero Aprosio, A., Zanchi, C., & Giuliani, M. (2024). Introducing PaVeDa – Pavia Verbs Database: Valency Patterns and Pattern Comparison in Ancient Indo-European Languages. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024* (pp. 79–88). ELRA and ICCL. https://aclanthology.org/2024.lt4hala-1.10

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

Malchukov, A., Haspelmath, M., & Comrie, B. (2010). Ditransitive constructions: A typological overview. In A. Malchukov, M. Haspelmath, & B. Comrie (Eds.), *Studies in Ditransitive Constructions* (pp. 1–64). DE GRUYTER MOUTON. https://doi.org/10.1515/9783110220377.1

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1301.3781

Ni, J., Dinu, G., & Florian, R. (2017). Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1470–1480. https://doi.org/10.18653/v1/P17-1135

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, *36*(1), 307–340.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71–106. https://doi.org/10.1162/0891201053630264

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Schäfer, H., Idrissi-Yaghir, A., Horn, P., & Friedrich, C. (2022). Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language. *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 53–62. https://doi.org/10.18653/v1/2022.clinicalnlp-1.6

Sennrich, R., & Volk, M. (2010, October 31). MT-based Sentence Alignment for OCR-generated Parallel Texts. *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*. https://aclanthology.org/2010.amta-papers.14

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shi, P., & Lin, J. (2019). *Simple BERT Models for Relation Extraction and Semantic Role Labeling* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1904.05255

Strubell, E., Verga, P., Andor, D., Weiss, D., & McCallum, A. (2018). Linguistically-Informed Self-Attention for Semantic Role Labeling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5027–5038. https://doi.org/10.18653/v1/D18-1548

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Kincksieck.

Van Valin, R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press.

Wälchli, B. (2010). Similarity Semantics and Building Probabilistic Semantic Maps from Parallel Texts. *Linguistic Discovery*, *8*(1). https://doi.org/10.1349/PS1.1537-0852.A.356

Wälchli, B., & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, *50*(3). https://doi.org/10.1515/ling-2012-0021

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *arXiv Preprint arXiv:2402.05672*.

Zhou, J., & Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. *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)*, 1127–1137. https://doi.org/10.3115/v1/P15-1109

Appendix A Additional figures

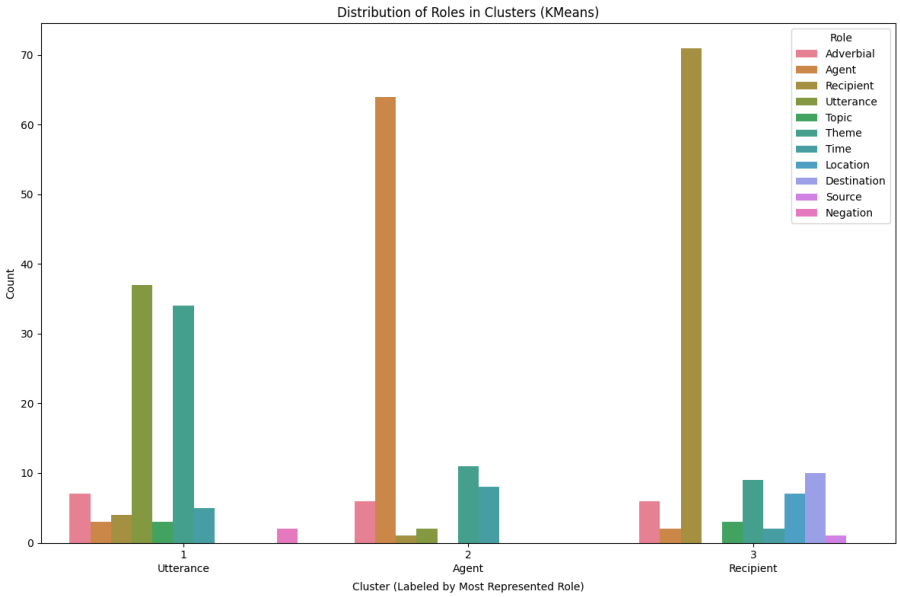


Figure 19. *Distribution of the roles in three clusters (K-Means).*

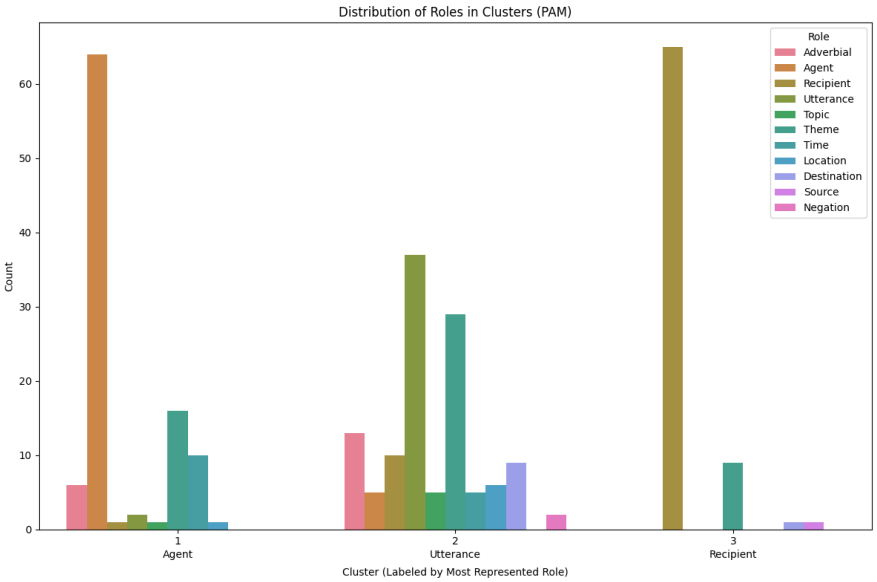


Figure 20. *Distribution of the roles in three clusters (PAM).*

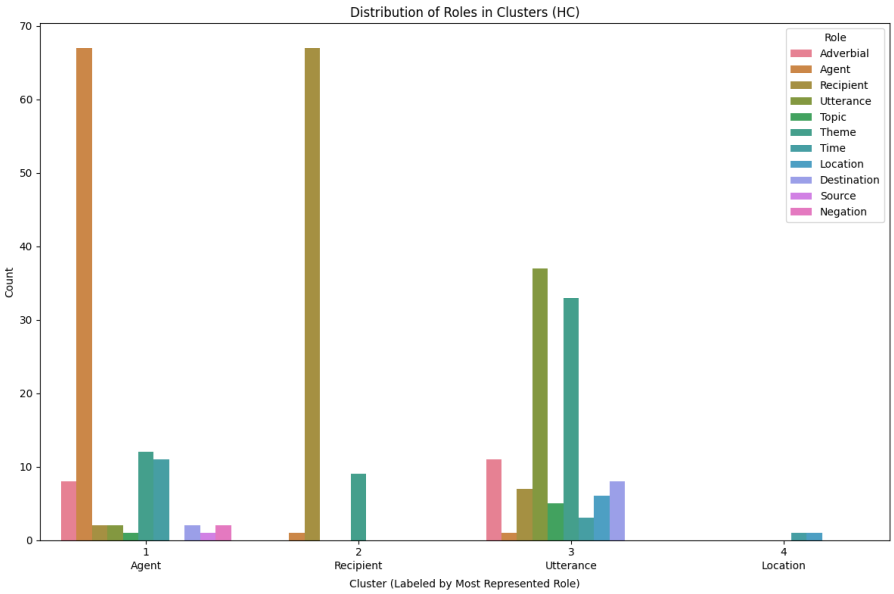


Figure 21. *Distribution of the roles in four clusters (hierarchical clustering).*