INAUGURAL-DISSERTATION
zur Erlangung der Doktorwürde der Philosophie (Dr. phil.)

des Fachbereichs Germanistik und Kunstwissenschaften der
Philipps-Universität Marburg

A Cross-linguistic Comparison of Content Interrogatives

Vorgelegt von

Siyu Liu, M.A.

aus

Guangdong, China

Marburg/Lahn im August 2023

Vom Fachbereich Germanistik und Kunstwissenschaften der Philipps-Universität Marburg als Dissertation angenommen am 17.08.2023

Tag der mündlichen Prüfung/Disputation am 14.12.2023

Erstgutachter: Prof. Dr. Michael Cysouw

Zweitgutachter: Prof. Dr. Martin Haspelmath
(Abteilung für Sprach- und Kulturevolution, Max-Planck-Institut für evolutionäre Anthropologie)

A Cross-linguistic Comparison of Content Interrogatives

INAUGURAL-DISSERTATION

zur Erlangung der Doktorwürde der Philosophie (Dr. phil.)

des Fachbereichs Germanistik und Kunstwissenschaften der Philipps-Universität Marburg

Vorgelegt von

Siyu Liu, M.A.

Geboren am 16. Juni 1993 in Guangdong, China

Marburg/Lahn im August 2023

Erstgutachter: Prof. Dr. Michael Cysouw

Zweitgutachter: Prof. Dr. Martin Haspelmath
(Abteilung für Sprach- und Kulturevolution, Max-Planck-Institut für evolutionäre Anthropologie)

Liu, Siyu: A Cross-linguistic Comparison of Content Interrogatives

Marburg, Philipps-Universität Marburg (1180)

Vom Fachbereich Germanistik und Kunstwissenschaften der Philipps-Universität Marburg als Dissertation angenommen am 17.08.2023

Tag der mündlichen Prüfung/Disputation am 14.12.2023

Jahr der Veröffentlichung der Dissertation an der Universitätsbibliothek Marburg: 2024

Table of Contents

Acknowledgement

First of all, I am truly thankful to my supervisor Michael Cysouw, who introduced me into linguistic typology and supported me throughout the PhD process. I will never forget the day he explained how to document a lesser-described language in a workshop in the reading week, where I found my enthusiasm for language diversity. He also showed me the possibilities of using various quantitative methods for linguistic research, which helped me to overcome my long-standing fear of maths and statistics. During the writing of this thesis, he was always generous with his time in weekly supervision meetings and provided insightful suggestions for my research. I am very fortunate to have such an inspiring teacher like him. I would also like to thank Martin Haspelmath for his stylistic suggestions for this dissertation.

I am so grateful to have met good teachers along the way. Wu Meiling Laoshi at Caitian School initiated me to the fun to learn a foreign language. The teaching of all the professors and lecturers at the German Department of Sun Yat-sen University made it possible for me to study in Germany. Johanna Mattissen introduced me into linguistic with a comprehensive seminar at the University of Cologne. Gea de Jong-Lendle encouraged me to do my PhD in Marburg and is always willing to give me advice on both academic and practical matters.

I am also indebted to all my friends. It is warming to receive greetings, funny jokes, stories and life updates from you. I would also like to thank Huang Shanshan, Ayse Gül Ünlü and Wenxun for answering my questions on Korean, Turkish and Japanese language.

I am incredibly fortunate to have my family cheering me on during all these years away from home. Thank you, my mum, for always being there for me. You are the one I can always rely on to listen to my ideas, give me advice and share both my worries and joys. This thesis is dedicated to you. Even though the meeting was only virtually possible, each family member was the strength that got me through the lonely days during the Corona pandemic. It is really exciting to reunite with you soon. Finally, to my beloved grandpa and grandma, even though you are no longer by my side, the love and trust you gave me up until the last day walked me through the hard times. How I wish you were still here to share everything in life!

Abstract

This dissertation explores the diversity of content interrogatives and the intricate semantic distinctions both between and within interrogative categories. The research employs the Massively Parallel Text method with which 413 interrogative contexts are collected from Bible translations in 88 languages. By observing the interrogative codings used in these contexts, this study inductively identifies six major categories and 38 sub-categories with the aid of the quantitative technique Cluster Analysis. Furthermore, the statistical results suggest the representative context for each interrogative category and sub-category. These exemplars can be used as a template for the characterisation of content interrogatives in language description. This research also illustrates cross-linguistically typical derivations within and across the identified interrogative categories based on the codings applied in these representative contexts. These derivations can be interpreted as typical diachronic pathways for content interrogatives.

Chapter 1 begins with an elaborate introduction of the research object, i.e., content interrogatives (Section 1.2). Subsequently, Section 1.3 presents a discussion of methodological issues relevant to linguistic comparison across languages. Section 1.4 lists the research questions. Section 1.5 provides an overview of the previous studies carried out on content interrogatives. The outline of this book is found in Section 1.6.

Chapter 2 focuses on the research methods used in this study. Section 2.1 presents the Massively Parallel Text method. Section 2.2 introduces quantitative approaches used for data analysis. Chapter 3 embarks on providing information on the Parallel Bible Corpus, which is the source of data for this investigation (Section 3.1). Section 3.2 presents the sampling strategy and the sampled languages. Section 3.3 and 3.4 describe the procedure of data collection and processing in detail. Section 3.5 introduces the online repository in which all data of this study is stored.

Chapter 4 discusses the results of this study. Section 4.1 provides general information on the clustering of interrogative contexts. A total of 413 interrogative contexts are classified into six primary categories: TIME (Section 4.2), PLACE (Section 4.3), PERSON (Section 4.4), THING (Section 4.5), INTENTION (Section 4.6) and MANNER/EXTENT (Section 4.7). Furthermore, the second level of clustering identifies 38 interrogative sub-classes within these six categories. These sub-classes are illustrated in their respective sections.

Chapter 5 presents the significant derivational links between interrogative sub-categories based on the interrogative constructions used in the representative contexts. Section 5.1 first provides general notes on identifying derivations. Section 5.2 to 5.6 present the derivational connections within TIME, PLACE, PERSON, THING, MANNER/EXTENT, respectively. Finally, Section 5.7 illustrates the derivations across six primary interrogative categories.

Zusammenfassung

Diese Dissertation untersucht die Vielfalt der Content Interrogativen und die komplizierten semantischen Unterschiede sowohl zwischen als auch innerhalb der Interrogativkategorien. Dazu verwendetet die Untersuchung die Methode Massively Parallel Text, womit 413 Interrogativkontexte aus Bibelübersetzungen in 88 Sprachen gesammelt wurden. Sechs Hauptkategorien und 38 Subkategorien sind mit Hilfe des quantitativen Ansatzes Clusteranalyse identifiziert, indem man die in Interrogativkontexten benutzten Fragekodierungen beobachtet. Darüber hinaus geben die statistischen Ergebnisse Aufschluss über den repräsentativen Interrogativkontexte für jede Kategorie und Unterkategorie. Diese repräsentativen Kontexte können als Vorlage für die Charakterisierung von Content Interrogativen in der Sprachbeschreibung verwendet werden. Weiterhin werden die Ableitungen innerhalb und zwischen den identifizierten Interrogativkategorien dargestellt. Diese Ableitungen lassen sich als typische diachronische Pfade für Content Interrogative interpretieren.

Kapitel 1 bietet zuerst eine ausführliche Einführung in den Forschungsgegenstand, nämlich Content Interrogative in Abschnitt 1.2. Anschließend werden die methodischen Fragen, die für den sprachübergreifenden Vergleich relevant sind, in Abschnitt 1.3 erörtert. Abschnitt 1.4 listet die Forschungsfragen dieser Untersuchung auf. Dann wird in Abschnitt 1.5 ein Überblick über die bisherige Forschungen zu Content Interrogativen gegeben. Die Gliederung des vorliegenden Buches befindet sich in Abschnitt 1.6.

Kapitel 2 befasst sich mit den in dieser Studie verwendeten Forschungsmethoden. In Abschnitt 2.1 wird die Methode Massively Parallel Text vorgestellt. Anschließend werden die quantitativen Ansätze für die Datenanalyse in Abschnitt 2.2 gezeigt. In Abschnitt 3.1 des Kapitels 3 werden zunächst Informationen über das Parallel Bible Corpus, das als die Datenquelle für diese Forschung dient, gegeben. Danach werden die Stichprobenstrategie und die untersuchten Sprachen in Abschnitt 3.2 dargestellt. In den Abschnitten 3.3 und 3.4 wird das Verfahren der Datenerhebung und -verarbeitung im Detail beschrieben. Das Online-Repository, in dem alle Daten dieser Studie gespeichert sind, wird in Abschnitt 3.5 vorgestellt.

In Kapitel 4 werden die Ergebnisse dieser Untersuchung behandelt. Zunächst beinhaltet Abschnitt 4.1 allgemeine Informationen über die Cluster der Interrogativkontexten. Insgesamt werden 413 Interrogativkontexte in sechs Hauptkategorien eingeteilt: TIME (Abschnitt 4.2),

PLACE (Abschnitt 4.3), PERSON (Abschnitt 4.4), THING (Abschnitt 4.5), INTENTION (Abschnitt 4.6) und MANNER/EXTENT (Abschnitt 4.7). Auf der zweiten Ebene der Clusterbildung werden 38 Unterklassen innerhalb dieser sechs Interrogativkategorien identifiziert. Diese Unterkategorien werden in den entsprechenden Abschnitten beschrieben.

Anhand der Interrogativkonstruktionen, die in den repräsentativen Kontexten benutzt sind, werden in Kapitel 5 die signifikanten Ableitungen zwischen den interrogativen Unterkategorien präsentiert. Abschnitt 5.1 bietet zuerst allgemeine Hinweise zur Identifizierung der Ableitungen. Danach werden die Ableitungsbeziehungen jeweils innerhalb der sechs Hauptkategorien in den Abschnitten 5.2 bis 5.6 dargestellt. Abschließend werden die Ableitungen zwischen den Hauptkategorien in Abschnitt 5.7 veranschaulicht.

List of Abbreviations and Symbols

2	second person
3	third person
↘	falling intonation
↗	rising intonation
=	clitic
-	suffix
/	or
…	position of other possible grammatical elements
CONT	continuous aspect
CL	classifier
DES	desiderative (Central Alaskan Yupik; Miyaoka 2012: 446)
INAL	inalienable possession (Abui; Kratochvíl 2007: 129)
INT.SUFF	interrogative suffix
LOC	localizer (Wolof; Robert 2016: 7)
PFV	perfective aspectual suffix
PL	plural
PROG	progressive aspect
PROX	proximal spatial suffix (Wolof; Robert 2016: 7)
Q	question word
SG	singular
SUBJ.FOC	subject focus conjugation (Wolof; Robert 2016: 7)

List of Tables

List of Figures

1 Introduction

1.1 A glimpse of interrogative diversity

Human languages have special structures to ask questions. In order to inquire for missing information, there exists an apparently universal class of function words — content interrogatives. In questions, content interrogatives are the linguistic expression of the information which is unknown to the speaker. Syntactically, they can be considered as substitutions for questioned noun phrases, determiners, adverbs, verbs or phrases expressing time, location, manner, etc. Despite the universal existence of content interrogatives, there is massive variation across languages as to the queried content of individual interrogatives and the semantic distinctions between them. To illustrate this variation, consider the following temporal questions shown in (1.1) and (1.2).

(1.1) Time of an action

a. English ***When*** *do you go?*

b. German ***Wann*** *gehst du?*
when go.2SG 2SG
lit. '**When** do you go?'

c. Korean 너 언제 가?
[neo ***eonje*** *ga]*
2SG when go
lit. '**When** do you go?'

d. Mandarin 你 什么 时候 走?
[nǐ ***shénme*** ***shíhòu*** *zǒu]*
2SG what time go
lit. '**What time** do you go?'

e. Cantonese 你 幾 時 走?

[nei⁵ ***gei²*** ***si⁴*** *zau²]*

2SG how many time go

lit. '**How many time** do you go?'

(1.2) Time of the day

a. English ***What*** ***time*** *is* *it* *now?*

b. German ***Wie*** ***spät*** *ist* *es* *jetzt?*

how late be.3SG 3SG now

lit. '**How late** is it now?'

c. Korean 지금 몇 시-야?

[jigeum ***myeoch*** ***si****-ya]*

now how many o'clock-INT.SUFF

lit. '**How many o'clock** is it now?'

d. Mandarin 现在 几 点?

[xiànzài ***jǐ*** ***diǎn]***

now how many o'clock

lit. '**How many o'clock** is it now?'

e. Cantonese 现在 幾 點?

[jin⁶zoi⁶ ***gei²*** ***dim²]***

now how many o'clock

lit. '**How many o'clock** is it now?'

Both situations of (1.1) and (1.2) involve time. According to the semantic distinction of the questioned content, the languages in the examples apply different interrogative constructions.

In (1.1a), English uses an unanalyzable word *when* to ask for the time of an action, while a different construction *what time* is applied in (1.2a) for the question of time of the day. An opposite case is presented by Cantonese (1.1e & 1.2e). This language utilizes two highly similar constructions including the same interrogative, roughly translated as 'how many' in English, for these questions.

Referring to the same content, languages can adopt different interrogative constructions. When asking for time of the day in (1.2), English and German respectively employ 'what time' and 'how late', while Korean, Mandarin, and Cantonese apply the same construction 'how many o'clock'. Considering the areal and genealogical relations between these languages, there still exists differentiation. In (1.1), Korean shows a similar interrogative construction to the European languages English and German, even if Korean is areally more closely connected to Mandarin. Meanwhile, Mandarin and Cantonese are genealogically related. But when questioning the time of an action, these two languages tend to use different interrogative expressions (1.1d vs. 1.1e). The same situation is also displayed in English and German when asking for time of the day (1.2a vs. 1.2b).

These two examples provide a brief glimpse into the diversity of content interrogatives. Even though there are only five well-described languages and just two interrogative situations included in (1.1) and (1.2), it can already be seen that no identical interrogative expression is applied by all languages for particular content. When the horizon is expanded to more languages in the world and a wider variety of interrogative contexts are taken into consideration, more various and fascinating findings can be expected.

1.2 The object of the study

This section will elaborate on aspects of interrogatives and questions. The content of the present section is organized as follows. In §1.2.1 I will first distinguish two notions — INTERROGATIVE, a kind of sentence type, and QUESTIONS, a speech act type. Then in §1.2.2, I will differentiate subtypes of interrogative clauses. The focus will be given to interrogative clauses with the function to inquire, i.e., polar questions, content questions, and alternative questions. §1.2.3 will be dedicated to the main topic of this research, i.e., content questions. Grammatical strategies attested cross-linguistically to construct content questions are presented subsequently. Some central characteristics of interrogative words and the typical

devices to mark content questions will be introduced. Afterward, I will discuss semantic features expressed with content questions in §1.2.4. The goal is to present a general picture of semantic categories distinguished in interrogatives. Following this, §1.2.5 will focus on the association between semantic categories and interrogative words. The way in which languages code information in interrogatives will also be presented in this section. §1.2.6 will illustrate the structure of an interrogative paradigm. The final §1.2.7 will discuss the terminology used in the remainder of this book.

1.2.1 Interrogative vs. question

The term INTERROGATIVE refers to a clause type. In terms of its syntactic properties, interrogative contrasts with other categories of clauses, i.e., declarative, imperative, and exclamative (cf. Sadock & Zwicky 1985: 160; Crystal 2008: 252, 433; König & Siemund 2007: 277; Dixon 2012: 376). The grammatical devices to form interrogatives vary across languages. The most common ones are inverted constituent order, final sentential rising pitch, special verb form and the use of interrogative particles and interrogative words.

Primarily, interrogative clauses are associated with the function of asking questions. The term QUESTION indicates a kind of speech act with which the inquirer seeks information or requests a response from the listener. In many linguistic works, question is reckoned to be equal to interrogative form, as Crystal (2008: 400) notes. Other typical clause functions opposed to question are statement, command, and exclamation. They are the main function of declarative clauses, imperative clauses, and exclamative clauses mentioned above, respectively (Crystal 2008: 433). A correspondence between clause forms and functions is given in (1.3):

(1.3)

FORM	**FUNCTION**
interrogative	question
declarative	statement
imperative	command
exclamative	exclamation

However, clause types and functions do not always have a one-to-one correspondence to each other, just as some interrogatives are pragmatically interpreted by the addressee other than requesting information or response. Besides asking questions, there are more functions of interrogatives. Some instances are listed with examples from Levinson (2012: 12) in (1.4).

(1.4) a. *How do you do?* (introduction)
b. *He said what?* (repair)
c. *Why don't we get a coffee?* (suggestion)
d. *Would you mind taking this?* (request)
e. *Well, what damn fool would trust a bank with their money?* (statement)
f. *Who do you think you are?* (reprimand)

In this regard, there exist mismatches between interrogative forms and question functions. That is to say, a sentence or a clause can be in an interrogative form, but is not used to formulate a question, given that the speaker does not really expect any answer from the opposite party. Pragmatically, the function of this kind of interrogative clause is possibly to convey a command or make a statement. Body language, hand gestures, or different intonations will normally feature in the utterance to give a hint of appropriate interpretation, as the two examples provided by Dixon (2012: 376) in (1.5):

(1.5) a. *Could you please close the window?*
(with a friendly intonation)
b. *Who knows?*
(with spreading hands in a gesture of despair)

Formally, (1.5a) is an interrogative clause. But functionally, it serves as a polite imperative to require the listener to close the window. The interpretation of an expression like (1.5b) depends largely on the tonal or gestural indications made by the speaker as well as the context. In the case of (1.5b), the concomitant body language discloses that the addresser is actually making a statement of *no one knows*. This kind of expression occurs quite often in

daily conversation and is termed a RHETORICAL QUESTION. The usage of the word 'question' here is not completely in accordance with the definition posed previously. Nevertheless, since this term is already widely accepted and used, I will use it in the following discussion.

Another relevant situation involves the formal overlap between interrogatives and indefinites in many languages. It refers to the phenomenon that indefinites and interrogatives are identical in form. For instance, in Chamorro, the same set of pronouns is employed for interrogatives and indefinites, as displayed in (1.6). A counter-example is English in which the formal affinity between interrogatives and indefinites is much less attested.

(1.6) Chamorro (Chung 2020: 192, 194)

håyi	'who, someone'
håfa	'what, something'
(a)månu	'where, somewhere'
ngai'an	'when, sometime'

The functional differentiation between interrogatives and indefinites is usually realized by means of syntactic devices or suprasegmental signals (Haspelmath 1997: 170). A comparison from Mandarin is offered in (1.7). In Mandarin, interrogative pronouns are identical to indefinite pronouns, as *shěnme* means both 'what' and 'something'. The cue of disambiguation is the intonation — (1.7a) is a statement containing indefinite meaning which must be finished with the falling intonation, whereas the final rising intonation in (1.7b) is indicative of the interrogative reading.

(1.7) Mandarin

a.	*nǐ*	*chī*	*le*	*shěnme.*↘
	you	eat	PFV	something
	'You ate something.'			
b.	*nǐ*	*chī*	*le*	*shěnme?*↗
	you	eat	PFV	what
	'What did you eat?'			

Conversely, it is also not uncommon that other clause types are applied to elicit information or seek confirmation, such as the example in (1.8):

(1.8) *You've finished your homework?↗*

Syntactically, the utterance in (1.8) is formulated as a declarative sentence without any formal marker signifying an interrogative in English. However, with the help of the final rising intonation, this utterance is pragmatically perceived as the speaker expressing a sense of uncertainty towards the statement. The real intention of making such an utterance is to enquire about whether the action is actually done. Therefore, it should be regarded as a declarative clause with the purpose to ask a question. This case is also called a *declarative question* (Haan 2002: 16), as will be seen in (1.9e) below.

1.2.2 Categorization of interrogative clauses

Interrogative clauses can be divided into several types in different ways, according to various aspects, e.g., speech functions, syntactic features, and concrete usage. For example, Haan (2002: 12-18) distinguishes nine subcategories of interrogative clauses in (1.9):

(1.9) Subtypes of interrogative clauses

a. Polar questions
b. Content questions
c. Alternative questions
d. Tag-questions
e. Declarative questions
f. Echo questions
g. Elliptic questions
h. Rhetorical questions
i. Embedded questions

Here, Haan (2002) seems to equate the term 'question' with interrogative clauses, since it is applied to all interrogative structures in (1.9). As aforementioned, it should be noticed that

not every type of (1.9) is utilized to address a 'real' question, i.e., to seek information. The most conspicuous ones are echo questions (1.9f) and rhetorical questions (1.9h). The actual function of echo and rhetorical questions is not to solicit information or confirmation. Instead, the former type is used to reveal the surprise or disbelief of the speaker, while the latter one is a statement with which the addresser assumes affirmation from the recipient. Given that it might lead to terminological confusion, I prefer keeping the difference between two terms, question and interrogative, apart in my succeeding discussion.

Despite the different functions, some interrogative types are often related in form and share similar marking strategies, as Haan (2002: 12-18) argues. For example, an alternative question consists of multiple polar questions that are combined through the conjunction *or*. In this way, the addressee is provided with a set of options. An elliptic question can be seen as a content question that dispenses with repetitive elements of the preceding discourse. The interrogative form of embedded questions can overlap with content questions and polar questions, depending on the grammatical devices, such as interrogative words or inversion. Rhetorical questions, in spite of the pragmatical function, cannot be differentiated from content questions or polar questions only by their form.

If we only consider the interrogative clauses that address questions in the functional sense, the categorization can be conducted on a different ground. According to the traits of the expected answer, questions fall into three main types: polar questions, content questions, and alternative questions, as demonstrated in (1.10) (cf. Sadock & Zwicky 1985: 179; Huddleston 1994: 416; Siemund 2001: 1010).

(1.10) a. *Will it rain today?* (polar question)
b. *How is the weather now?* (content question)
c. *Is it raining or snowing outside?* (alternative question)

With polar questions, as shown in (1.10a), the speaker is expecting the listener to judge whether the statement is true or false, so the answer is a close class of 'yes' and 'no'. Thus, this kind of question is also called yes-no question. However, any value of the scale between 'yes' and 'no' is also the possible response to polar questions, such as 'perhaps' or 'maybe' (König & Siemund 2007: 291; Dixon 2012: 377; Aikhenvald 2015: 236).

By contrast, content questions like (1.10b) come into play when there is a missing piece of knowledge by the speaker. Thus, the answer to this kind of question should encompass specific information other than a truth value. The answer of content questions are theoretically open and unlimited, as long as they are discourse-relevant.

Finally, questions like (1.10c) are named alternative questions. With alternative questions, the inquirer gives a set of options and expects the addressee to make a decision among them. In this respect, although they have a formal affinity with polar questions, alternative questions are not satisfied just with a yes/no answer but are waiting for information to complete the discourse, which is functionally akin to content questions.

Not all of the above exhibited types of questions are attested cross-linguistically. Dixon (2012: 377, 426) considers two core practical purposes of asking questions as universally existent, i.e., requesting confirmation of an old or known status and seeking information about a new or unknown situation. Based on this criterion, questions can be classified into two basic sorts — polar questions serving the former aim and content questions used to operate the latter purpose. Alternative questions, on the contrary, are not found universally (Dixon 2012: 398). They are not irreplaceable in the form, since an alternative question can be separated into multiple polar questions. Moreover, their function can be performed with polar questions and content questions. Thus, alternative questions are considered secondary on this ground.

1.2.3 Content questions

The interest of this study lies in content questions. Summarized from the illustration above, I will define CONTENT QUESTION as an interrogative clause with the purpose to obtain specific information. Other alternative terms of content question are, for example, constituent question, information question, and wh-word question. The use of content questions is "very nearly" universal (Sadock & Zwicky 1985: 179). Content questions arise when there is an information gap between the speaker and the listener. In this case, the speaker lacks knowledge of a certain proposition in a discourse. The addressee is requested to provide relevant information. The expected information may pertain to various domains. Some major ones, as Siemund (2001: 1018) poses, are "participants and objects", "more circumstantial information relating to the relevant locational or temporal setting", and "issues like the manner of execution and the purpose".

Usually, languages distinguish content questions from polar questions not only in their function but also in their formal marking strategies. In terms of polar questions, some common grammatical features are listed in (1.11) (cf. Siemund 2001: 1012; König & Siemund 2007: 292; Dixon 2012: 391-394):

(1.11) a. Intonation patterns or pitch
b. Interrogative particles
c. Interrogative tags
d. Disjunctive-negative structures
e. Interrogative word order
f. Interrogative mood / verbal inflection
g. Addition or omission of phonological or morphological features

A language can combine different grammatical means to establish its own polar question mechanism. Among all, it is reported that intonation is the most distinctive and ubiquitous strategy to mark polar questions (Ultan 1978: 7). More specifically, rising intonation is most often employed in polar questions across the world (König & Siemund 2007: 292). Dryer (2013) also surveys the cross-linguistic distribution of the grammatical methods to build polar questions. In a sample of 955 languages, 585 of them utilize interrogative particles to distinct polar questions from declarative sentences.

All the in (1.11) presented strategies of polar questions are also applicable to form content questions. Only the frequency of occurrence is different. For instance, intonation is no longer predominant in identifying content questions. According to the data of Ultan (1978), only one-third of the sample languages optionally use intonation to mark content questions, whereas this strategy is not adopted in another one-third of languages at all.

Even though there are many possible grammatical devices and the strategies vary across languages, INTERROGATIVE WORDS are always the defining hallmark to construct content questions.[1] An interrogative word substitutes a certain component of a statement in the

1 Interrogative words are the most vital tool to form content questions. Yet, not every language has a set of grammatical structures that function dedicatedly as interrogative words. For example, Velupillai (2012: 358) points out that Wari', a language of the Chapacuran family, forms content questions through fronting the demonstrative *ma'* in clause-initial position.

corresponding question. The information that the inquirer is seeking is encoded in interrogative words. As Haan (2002: 13) states, interrogative words have the function to limit the domain to which the expected answer belongs. For instance, *who* directs that the answer is related to a human concept, while *when* specifies that the answer is involved with a temporal relation.

In fact, there is no agreement reached on the terminology for the grammatical strategy transmitting semantic properties of the expected answer in content questions. The reason for the use of the term 'interrogative word' in the last paragraph is that it occurs most frequently in references and is also widely comprehensible. However, since I have defined interrogative in §1.2.1 as a formal representation whose major function is to ask questions, I prefer maintaining the terminological consistency. Thus, the term CONTENT INTERROGATIVE is applied to refer to interrogative word in the following discussion. This label is commonly accepted in many linguistic works as well. Also, it can be abbreviated as interrogative. A more detailed discussion related to this point will be given in §1.2.7.

Several facets of content interrogatives are worthy of discussion. The most classic one must be their position in the clause. It refers to the fact that content interrogatives are arranged into different syntactic positions across languages. Since it is not the main concern of this research, this point will only be presented briefly in the following. There are three attested types into which languages are classified according to the position of content interrogatives (Siemund 2001: 1019), see (1.12):

(1.12)

	Position of content interrogatives	**Language type**
a.	clause-initial	fronting
b.	the same position of the questioned constitute	in-situ
c.	no obligatory position	optional fronting

Similarly, Dryer (2013) identifies two common classes cross-linguistically: a) content interrogatives are situated obligatorily at the beginning of the clause, as (1.12a), and b) they occur optionally in the clause. In Dryer's description, the latter type includes the in-situ languages, as (1.12b), in which content interrogatives naturally take the position of the inquired constituents. He then finds a split positioning in some languages. In this case,

grammatical properties of content interrogatives, such as word classes, have an influence on their position in the clause. An example is encountered in Malakmalak in which interrogative noun phrases must be placed clause-initially, whereas the position of interrogative adverbs is not strict (Birk 1976: 26; Dryer 2013). Within 902 sampled languages, Dryer (2013) notes that 68% of them do not obligatorily locate content interrogatives in the clause-initial position, whereas 29% require them to be fronted at the beginning of the clause.

Another important topic of content interrogatives is their semantic properties. Given that this subject is highly relevant to the current study, the next two sections will elaborate on it.

1.2.4 Semantic features of content interrogatives

— What kinds of information can be asked?

As mentioned previously, almost every language in the world has a set of interrogatives to generate content questions. The content interrogatives carry the semantic properties of the expected information. Many studies come across this topic and have identified sets of categories that are expressed in content interrogatives either across languages or in particular languages. However, given the fact that these works are conducted with different collections of sampled languages or with various research emphases, there exist terminological inconstancy and divergent results. In the following, I summarize several attested semantic categories from different authors and make a comparison of them. Although the summary is non-exhaustive, it includes the most prominent ones.

Firstly, Mackenzie (2009: 1132) distinguishes six basic semantic categories in content interrogatives through observing a sample of 50 languages, as given in (1.13).

(1.13) a. INDIVIDUAL
b. LOCATION
c. TIME
d. MANNER
e. QUANTITY
f. REASON

Heine et al. (1991: 55-59) investigate the so-called metaphorical relations between conceptual domains and interrogatives with a sample of 14 languages. In this study, seven semantic categories are involved, as listed in (1.14).

(1.14) a. PERSON
b. THING
c. ACTIVITY
d. PLACE
f. TIME
h. MANNER
i. PURPOSE/CAUSE

Compared to (1.13), some semantic categories are classified into finer subtypes in Heine et al. (1991). Based on the parameter ANIMACY, INDIVIDUAL in (1.13a) can be separated into two classes, i.e., PERSON ('who') and THING ('what'). The distinction between these two categories is universally existent in interrogative encodings. Interestingly, the category ACTIVITY ('do what') is individually sorted out in (1.14). This aspect comprises concepts like EVENT, PROCESS, and ACTION (Heine et al. 1991: 57). It is in many well-known languages indistinguishable from THING because of the formal unmarkedness. However, there are also languages in the world that possess a content interrogative particularly to ask for ACTIVITY. The category REASON can also be separated into two finer-grained classes, i.e., PURPOSE and CAUSE, as they are distinctively marked in interrogatives in some languages. Yet, the distinction between PURPOSE and CAUSE is much less frequently attested than that between PERSON and THING in interrogatives across languages.

Diessel (2003) compares demonstratives and interrogatives. In this work, he also identifies seven semantic categories which are considerably overlapped with the last two sets in (1.13) and (1.14): PERSON, THING, PLACE, DIRECTION, TIME, MANNER, and AMOUNT. Here, AMOUNT is a terminological alternative to QUANTITY. What is especially noticeable with this classification is that there are two categories related to spatial relations, i.e., PLACE and DIRECTION. Under the domain of DIRECTION, Diessel (2003) further differentiates GOAL ('whither') and SOURCE ('whence'). It is quite common that languages have a tripartite

interrogative paradigm to mark spatial semantic categories, such as in German *wo* 'where', *wohin* 'whither', and *woher* 'whence'.

Cysouw (2004) adds more possible categories. A key one left out in (1.13) and (1.14) is SELECTION ('which') which is recurrently marked in interrogatives. This category, as Hölzl (2018: 81) insists, is necessary to be distinguished from KIND ('what kind of'), as the latter denotes the referentiality and delimits a specific object. Noteworthily, Cysouw (2004: 4, 8) discerns two kinds of the category EXTENT. One is established in German with *inwiefern/ inwieweit* with which the inquirer appears to ask for an explanation. It may be translated as 'in what way' in English. The other type is easily confused with MANNER given the homogenous form in many languages. Yet, it denotes the grade of a state, as 'how far' in English. Besides, RANK is also marked in interrogatives in some languages.

On the basis of the above attested basic semantic features, Cysouw (2004) observes further specification of secondary categories in content interrogatives. Apart from the distinction between PLACE and DIRECTION before-mentioned, POSITION ('be where') is sometimes differentiated within the general spatial domain. In terms of temporal concepts, some languages tell apart questions on a specific point of time in a day ('what time'), a general time ('when'), and the start point of time ('since when'). Moreover, whether the time refers to the past or the future might also play a role in conceptual perception (cf. Siemund 2001: 1023, Dixon 2012: 416). Correspondingly, it leads to separated interrogatives in certain languages. In comparison to spatial relations, the specification of TIME occurs relatively less around the globe. The category QUANTITY is in some languages divided into subtypes COUNT ('how many') and MASS ('how much'), which reflects the contrast between countables and non-countables. In addition to the category ACTIVITY in (1.14), which is also called ACTION in Cysouw (2004: 9), UTTERANCE ('say what') can be subtly separated from THING.

1.2.5 Semantic connections between content interrogatives

— How do interrogatives encode information?

In the last section, an overview of semantic categories that content questions express is given. However, there exists a significant cross-linguistic diversity in the way in which languages encode semantic information in interrogatives. That is to say, not every language designates a specialized content interrogative for each category. Some semantic features fall into the same

form in some languages, whereas in other languages they are assigned to heterogeneous interrogative constructions. Moreover, interrogatives of a paradigm are not completely structurally different. Rather, there exist certain formal similarities in their construction. The interrogatives of some domains can be derived from the same root or have overlapping linguistic material. Such connections vary across languages as well. In this section, I will provide a glimpse of how the semantic scope of content interrogatives ranges among categories.

To start the illustration with a well-known language, English. Dixon (2012: 407) suggests eight basic interrogative words in English, which are listed along with the corresponding semantic categories in (1.15):

(1.15)

	Interrogative	**Semantic category**
a.	*who*	PERSON
b.	*what*	THING
c.	*why*	REASON
d.	*where*	PLACE
e.	*when*	TIME
f.	*which*	SELECTION
g.	*how*	MANNER
h.	*how many/how much*	QUANTITY

In English, every basic semantic category in (1.15) is allocated a unique form. That is, these semantic features are individually marked. However, these forms are not completely unique. Question words from (1.15a) to (1.15f) are all started with *wh-*, while interrogatives for MANNER and QUANTITY share the same question word *how*.

Yet, this kind of system is not universally applicable. Some languages may have a small-size inventory of interrogatives. An extreme case is Asheninca Campa, an Arawakan language spoken in Peru. It seems that there is only one question word *tsika* serving in all interrogative contexts. As Cysouw (2007) claims, the original meaning of *tsika* is 'where'. The disambiguation of interrogative meanings eminently relies on the following verbs.

At the other end of the scale, the interrogative catalog of a language can also be pretty abundant in formally fresh words. According to Frajzyngier & Shay (2002: 357-378), in Hdi, a Chadic language spoken in Cameron and Nigeria, there is almost no formal connection among question words. Compare the following (1.16) with (1.15):

(1.16) Hdi (Frajzyngier & Shay 2002: 357-378)

	Interrogative	**Semantic category**
a.	*wá*	PERSON
b.	*nə́*	THING
c.	*ní-yà*	REASON
d.	*gá*	PLACE
e.	*yà-wú*	TIME
f.	*nú*	SELECTION
g.	*kí*	MANNER
h.	*kí dàrì*	QUANTITY

As noted in §1.2.5, the contrast between PERSON and THING might be the most prevalent feature in the world. The majority of languages have interrogatives respectively for these two categories based on the animateness of the referent. Nevertheless, there also exist languages that consent to the ambiguity of PERSON and THING. According to Idiatov (2007: 563), 7-9% of a 1850 sampled languages belong to this type. One example is Krenák, a Macro-Jê language spoken in Brazil, which uses a single form *hokonim* to ask for 'who' and 'what', as illustrated in (1.17). However, the way of disambiguation is not given.

(1.17) Krenák (Idiatov 2007: 553, citing Ehrenreich 1896: 617, 626)

a. ***hokonim*** *huk* *ńinum* *a-taň?*
Q his arm broke
'Who broke his hand?'

b. ***hokonim*** *akkorune?*
Q 2SG.want
'What do you want?'

Even though most languages differ PERSON and THING, the boundary between two categories is not so straightforward, since animateness is essentially a matter of degree and every language measures the animacy of an object with different yardsticks. A quintessential instance is the question about proper names. Some languages even allow both interrogatives for querying someone's name. Consider the example from Abui in (1.18) provided by Hölzl (2018: 81-82). In this language, the usage of *nale* 'what' (1.18a) or *maa* 'who' (1.18b) has no impact on the understanding.

(1.18) Abui (Hölzl 2018: 81-82, citing Kratochvíl 2007: 129)

a. *a-ne* ***nala****?*
   2SG.INAL-name what
   'What is your name?'

b. *a-ne* ***maa****?*
   2SG.INAL-name who
   'What is your name?'

Next to the widely attested opposition between PERSON and THING, in the majority of languages interrogatives of different categories associate with each other using various morphological connections. Even in Hdi, the interrogatives of MANNER and QUANTITY share the same lexeme *kí*, as can be seen in (1.16). In English, the construction *how many/much* is apparently derived from *how*. The same phenomenon can also be found in German, e.g., *wie* 'how', *wie viele* 'how many' and *wie viel* 'how much'. It appears that the formal affinity between these two categories is more than coincidental. It is highly possible that the interrogative for QUANTITY is derived from MANNER in a number of languages. Thus, new questions arise: which categories are regularly formally associated with each other? Is there a lexeme of a certain category that serves as the origin of every derivation?

In this regard, Cysouw (2004: 12-19) suggests a typology. Based on the analyzability of interrogative words, he classifies the attested interrogative categories into three types, as given in (1.19) to (1.21).

(1.19) Major categories

a. PERSON

b. THING

c. SELECTION

d. PLACE

(1.20) Minor categories

a. QUANTITY

b. MANNER

c. TIME

(1.21) Incidental categories

a. REASON

b. QUALITY

c. EXTENT

d. POSITION

e. ACTION

f. RANK

etc.

Semantic categories of the major type in (1.19) are mostly coded in basic lexemes. In other words, these interrogatives are rarely decomposable into smaller meaning units. Within this class, the analyzability also diversifies. Compared to the high non-analyzability of PERSON and THING, languages are more prone to derive SELECTION and PLACE from other domains. Interestingly, these four categories can mutually be the derivational source. For example, the interrogatives for SELECTION are in some languages derived from PLACE, PERSON or THING, while in other languages SELECTION, PERSON and THING can be used to generate interrogatives of PLACE.

In the next minor group in (1.20), it is less attested that the forms of these categories serve as the base of other interrogatives. Only about 40% of languages in the world build a single form without derivation or composition for MANNER and TIME. Such a situation for QUANTITY

is found in 60% of languages. The main source category for MANNER is THING, while the derivation for QUANTITY usually originates from THING, SELECTION, and PLACE. The interrogatives for TIME are recurrently composed of THING, SELECTION, MANNER, and QUANTITY, depending on the precise temporal relations (see the subtypes of TIME in §1.2.4). However, unlike QUANTITY, TIME is rarely derived from PLACE. The last class in (1.21) comprises semantic categories that are seldom attested as unanalyzable and lexicalized around the globe.

Besides derivational relationships, Cysouw (2005b) further documents some possible ambiguities between interrogatives, as the case PERSON = THING in Abui in (1.18). A summary of Cysouw (2005b: 4-5) is given in (1.22).

(1.22) a. PERSON = THING
b. MANNER = QUANTITY
c. THING = REASON
d. THING = QUALITY
e. THING = MANNER
f. PLACE = TIME
g. QUANTITY.MASS = TIME

1.2.6 The structure of content interrogatives

— How is an interrogative system constructed?

In the last section, we saw that content interrogatives appear not only in the form of single and unanalyzable lexemes, e.g., *what*, but they can also be a combination of multiple words, as *wie lange* ‘how long’ in German, or lexemes derived from another, e.g., *imanir* ‘why’ < *ima* ‘what’ in Huallaga Quechua (Weber 1989: 327). Another possible structural connection is the joint element *wh-* situated at the beginning of question words in English. Thus, the structure of the interrogative system differs across languages too.

According to the structural complexity of the interrogative paradigm, Muysken & Smith (1990) draw out a classification composed of five types, as given in (1.23) below.

As the second column of (1.23) indicates, the complexity increases from (1.23a) to (1.23e) (cf. Table 4.12 in Hölzl 2018: 87). When an interrogative serves as the source of all other

members within the paradigm and they are thus synchronically analyzable, such a system is considered transparent. Languages of the atrophied type used to have a transparent system, but it is later deprived of the interrogative marker. Mix-transparent class includes interrogative paradigms encompassing both analyzable and unanalyzable forms. A fused system refers to interrogatives in a paradigm that are historically related but synchronically inseparable. Finally, interrogatives in the opaque group vary from all members and are unable to be morphologically analyzed.

(1.23)

	Type	**Complexity**
a.	transparent	simple
b.	atrophied	↓
c.	mixed transparent	↓
d.	fused	↓
e.	opaque	complex

1.2.7 Terminology

As previously mentioned, no consensus has been achieved for the term content interrogative defined in §1.2.3. Some scholars prefer other names in the grammatical description of certain languages. For instance, Miyaoka (2012) uses the term ignorative for Central Alaska Yupik, while Mushin (1995) employs the term epistememe for Australian languages. The other frequent alternatives are, e.g., interrogative word, question word, interrogative pronoun and wh-word. However, they are all tricky for that and other reasons. None of them seems to be comprehensive enough to cover all kinds of feasible content interrogatives.

Firstly, the definition of word is already troublesome. As discussed above, an interrogative can have an appearance of a single lexeme, a complex derived construction, or a combination of lexemes. In some languages, interrogative components are even not allowed to be used alone. An example is Wolof in which the interrogative codings for certain categories is obligatorily composed of a noun class marker and an interrogative suffix (Robert 2016: 4), as an example displayed in (1.24). Thus, the term of word is not optimal to describe the basic unit of interrogatives.

(1.24) Wolof (Robert 2016: 7)

B-an	*moo*	*nekk*	*ci*	*ëtt*	*bi?*
CL-INT.SUFF	SUBJ.FOC.3SG	be.at	LOC	yard	CL:PROX

'Which one is in the yard?' [talking about a dog]

Second, the word class of an interrogative is indefinite. The typical word class of each semantic domain differ. PERSON and THING are normally asked with pronouns or nouns, so the corresponding chapter in grammars is often named interrogative pronoun. The interrogatives indicating SELECTION, KIND and QUANTITY usually locate in front of nouns. Thus they are labeled interrogative adjective. REASON, MANNER, PLACE and TIME are commonly asked with interrogative adverbs. Besides, interrogative verbs exist in some languages to inquire about ACTIVITY and UTTERANCE. Finally, the name of wh-word only reflects a structural trait in English. Thus, in no way it is appropriate for a cross-linguistic depiction.

In order to respect language diversity and avoid terminological biases, I suggest the term CONTENT INTERROGATIVE UNIT as a general notion referring to elements that encode semantic features of information in content questions. For the sake of succinctness, it can be simplified to content interrogative, interrogative coding, interrogative form and interrogative. When the thesis proceeds to the description of language-particular phenomena, other practical alternatives, e.g., interrogative pronoun or interrogative suffix, are also allowed, as long as they felicitously mirror the traits of interrogatives in given languages.

1.3 Methodology

The present chapter focuses on the methodological issues that one might come across in cross-linguistic comparison. In §1.3.1, I will first discuss the difficulty to define a comparable object across languages and illustrate some frequent cases of asymmetric linguistic structures during comparison. Next, §1.3.2 will give some ideas to establish an applicable basis for comparison across languages. Finally, I will depict the workflow of cross-linguistic comparison of this investigation in §1.3.3.

1.3.1 Issue of comparability

The workflow of a cross-linguistic study normally incorporates the following steps: setting the research entity, collecting data across languages, and conducting comparison and analysis. The goal of such an investigation is to identify patterns of language structures shared across languages or discover divergence among languages. The fundamental bottleneck of this course of action is to properly define the object to be studied and compared. In this respect, it leads to a concern for which the spirited debate is still ongoing, that is, how and to what extent linguistic structures are comparable across languages. In other words, a research object, i.e., a specific linguistic phenomenon, must be identifiable across languages, and secondly, their occurrences in the sampled languages are all used to practice the same certain function or express the same meaning. This then creates a common ground based on which a comparison can take place. Without a credibly comparable basis, cross-linguistic research will lack a coherent foundation. Any inference drawn from such a comparison is in fact meaningless. In this light, it is often suggested that a cross-linguistic comparison should be conducted on the basis of linguistic structures defined on a functional or semantic ground (cf. e.g., Dryer 1997; Haspelmath 2007; Haspelmath 2010; Stassen 2011; Croft 2016).

The radical problem of comparability of a cross-linguistic investigation lies in, as Croft (1995: 88), Stassen (2011: 90) and Evans (2020: 417) note, how can we guarantee that all selected linguistic structures from different languages exactly represent the object of interest, neither more nor less. Or in other words, how can we make sure that those instantiations are identical in a formal or functional sense? This consideration is usually not the final aim of a cross-linguistic study, but it is an inevitable precondition of a feasible comparison. This question seems easy to answer, but actually it is not. For example, it is often taken for granted that the function of a given grammatical category or the meaning of a lexical item are cross-linguistically the same across languages. However, structural codings for them do not necessarily have a one-to-one correspondence between languages. It is very possible to inaccurately equate structures, since mismatches between languages are normally so fine-grained that they are easily overlooked. In the following, I will present three examples to discuss the tricky issues of cross-linguistic comparability.

The first case illustrates the different distributions of meanings and functions encoded in morphosyntactic structures. Consider the following two examples in Cantonese and the corresponding translations in English given in (1.25).

(1.25) Cantonese (Matthews & Yip 2013: 231-232)

a. *Kéuihdeih* *léuhng* *go* *paak-gán-tō*
they two CL date-**PROG**
'The two of them are dating.'

b. *Kéuih* *sèhngyaht* *jeuk-jyuh* *ngàuhjái-fu*
s/he always wear-**CONT** cowboy-pant
'She/He is always wearing jeans.'

In English, the suffix *-ing* is indicate of the progressive aspect of action. However, it has two translational counterparts in Cantonese, as shown in (1.25). The example in (1.25a) describes an ongoing activity that might alter after some time. In this case, the progressive marker *-gán* is applied between the compound word *paaktō* 'date'. Conversely, (1.25b) uses a different marker *-jyuh* to explicitly tell that the activity is not dynamic, but continuously lasts or remains unchanged. For Cantonese speakers, such a differentiation between progressive aspect and continuous aspect of a state is always clear and must be signalled by using different grammatical elements. A misuse of *-gán* and *-jyuh* will cause an ungrammatical expression. But when it comes to translating these two sentences into English, this semantic differentiation is no longer formally identifiable, since the motion verbs in both situations are marked with the suffix *-ing*. That is to say, there exists an asymmetry between the meanings of the progressive forms *-ing* in English and *-gán* in Cantonese. One should not simply equate them with each other. Strictly speaking, only *-gán* could be seen as a pure progressive marker. Although the suffix *-ing* is dubbed with the same term, it is not specifically adopted for the progressive aspect but can also carry the meaning of continuity.

The second example that reflects the underlying semantic differentiation of a certain concept across languages is how languages lexically encode the human upper limb, as reported by Brown (2013). In the sample of 617 languages, two types are attested. 389

languages have separate words for hand, the body part from the fingertips to the wrist, and for arm, the segment from the fingertips or from the wrist way up to the shoulder. English is a typical example of this class. The second type comprises languages in which two body segments are denoted with an identical word, exemplified by *ruka* in Czech. In summary, one should always pay attention to the exact denotation of lexemes as well as the delicate discrepancy between languages during translation or language comparison.

Apart from the differences in linguistic structures or lexical concepts across languages, the confusion of function and form of research objects can also bring trouble to the comparison. Take English and German to illustrate. The adjective *beautiful* in English is correspondingly translated as *schön* in German. However, the word *schön* can also function as an adverb, given that in German adverbs of manner are not derived from adjectives by adding a suffix, as *-ly* in English, but instead these two grammatical categories have an identical form. The adverbial equivalent in English of *schön* should be *beautifully*. If the investigated topic is about the domain of adverb, which is defined as a commonly familiar entry in grammar and refers to a word class (cf. Crystal 2008: 14), the homology of adjectives and adverbs in German might misguide to a conclusion that there does not exist the class of adverb in this language and adjectives can also be used as adverbs. Obviously, such a statement is incorrect on the ground that the functional aspect and the formal realization are confounded with each other. Moreover, it is biased in favor of morphological fashion in English. German speakers do have the need, as users of other languages, to specify the manner of action, which refers to the adverbial function. The ways to embody this function in expression differ between languages. If we do not clarify the exact item to be compared in the first place but only take the formal pattern into consideration, a one-sided or even wrong generalization could consequently emerge.

Summarizing the instantiations above, we can see that selecting an object for a linguistic comparison deserves more caution than assumed. Linguistic structures defined in a pure formal domain are clearly not a solid basis. Regarding the functional and semantic structure, there are also pitfalls. In this term, researchers have already attested some possibilities to legitimately compare linguistic phenomena across languages, which will be discussed in the next section.

1.3.2 Language-independent basis of cross-linguistic comparison

Morphosyntactic constructions, as the examples in §1.3.1 show, are not reliable enough to be the basis of a cross-linguistic comparison given the differentiation in all levels of linguistic structures across languages. Besides, definitions purely made on formal grounds are exclusively applicable within a language. That is to say, they are not language-independent but rather only relevant to that specific language. This is also the reason why a structure defined in this way is not suitable to be compared cross-linguistically. Yet, it is common and sometimes necessary to directly compare formal structures between different languages. For instance, many descriptive grammars define grammatical constructions with terminology from languages with broader influence and longer linguistic tradition, e.g., Latin or English. On the one hand, high similarities between languages rationalize sharing grammatical terms, which spares the labor to create new ones. On the other hand, it will inevitably mislead readers to miss out peculiarities of the described language if we simply equate linguistic structures between different languages.

Despite their occasional role in linguistic comparison, it is increasingly clear that formal structures are unsuitable to be a comparative object. Dryer (1997: 117-118), for example, points out that the grammatical properties defining word classes in a language are unique to that language only. Even though they might still be similar to those in other languages in terms of their function, the formal realization and the distributional pattern of these grammatical properties are specific to that language. Haspelmath (2007: 121, 127) further argues that grammatical categories of language structure can only resemble each other, but not be completely identical, across languages, which indicates that they should be viewed as language-particular. The misidentification of a language-particular category as universal may lead to the consequence, as Dryer (1997: 140) addresses the concern, that the investigation of cross-linguistic patterns and the resulting distribution will be hindered.

Then what is a practicable language-independent basis for comparison across languages? As an example, Haspelmath (2007: 126) proposes that cross-linguistic comparison should be based on “substance” instead of “category”. However, he does not further clarify the definition of “substance”. More specifically, there is a suggestion to adopt a non-structural definition for a comparable object, such as a definition in a semantic or functional domain. Accordingly, it will be much more reliable to employ linguistic structures that are definable

by such criteria across languages, e.g., a particular function or lexical meaning, as the basis for cross-linguistic comparison. Stassen (2011: 94) advises the term "external criteria" to represent the semantic or functional foundation, since it is clearly opposed to formal or structural criteria that are often associated with the language-internal system. Besides, concepts based on pragmatic or discourse-functional criteria are also useful to define a comparable basis.

One of the most influential and comprehensive ideas proposed for comparison is the COMPARATIVE CONCEPTS put forward by Haspelmath (2010: 664-666, 673-677). Comparative concepts are a set of notions that are specifically created for comparison across languages. It must be clear in the first place that a comparative concept is not a category belonging to any particular language but rather a specifically designed artefact. Therefore, in accordance with concrete research interest, comparative concepts can be adjusted anytime to meet the demand. Comparative concepts are established based on general conceptual-semantic notions and some formal concepts that are assumed to be universal. Given the flexible nature of comparative concepts, they are not restricted to a single linguistic domain. Instead, they can be defined across several categories, as long as they fit the research goal. In this light, a comparative concept incorporates various domains and it can be internally fine-grained. Considering this feature, comparative concepts are ETIC (cf. Levinson et al. 2003: 487; Evans 2011: 509, Haspelmath 2018: 88-89). In sum, the underlying idea of comparative concepts is to set a frame for cross-linguistic comparison based on which language-particular structures are selected from different languages and compared with each other. With this process, linguists are able to generalize similarities and differentiations between linguistic structures across languages.

In contrast to the language-independent notion of comparative concepts, Haspelmath (2010: 664, 666-668) uses the term DESCRIPTIVE CATEGORIES to specifically refer to language-particular structures. Unlike comparative concepts, descriptive categories belong to the internal system of a language and they are defined by properties of that language. In this sense, descriptive categories are EMIC (cf. Evans 2011: 509; Haspelmath 2018: 109), compared to the etic characteristic of comparative concepts mentioned above. Descriptive categories in different languages may share many similarities. Nevertheless, since similarity cannot be simply equated to identity and each language has its own descriptive categories, it

is inappropriate to employ them as cross-linguistically comparative objects. Another purpose of posing the idea of descriptive categories is to better describe the unique traits and the grammatical mechanism of a specific language. After all, it is also valuable to gain a more discerning understanding of a single language via comparison across languages.

Although there exists a fundamental distinction between comparative concepts and descriptive categories, they are often assigned with the same terms from the grammatical tradition. To differentiate the two usages, comparative concepts are written in ordinary lower-case, e.g., 'adjective', whereas descriptive categories are marked with an initial capital, e.g., 'Adjective', as Comrie (1976: 10) suggests. This way of labeling properly indicates the relation between two notions, as LaPolla (2016: 367) states, that "[comparative concepts] are idealizations or prototypes formed on the basis of the family resemblances found in the descriptions". In contrast, descriptive categories are like proper names which are capitalized in English orthography.

However, linguists further advise refinements on comparative concepts. Stassen (2011: 94-96) poses the concern that a comparative basis purely defined with external criteria cannot entirely ensure the feasibility of cross-linguistic comparison, since the domain established on this ground could be too broad to delimit qualified structures. The same opinion is given by Haspelmath (1997: 9): "[…] purely functional definitions have the disadvantage that they tend to pick out quite heterogeneous expressions". In this regard, Stassen (2011) recommends employing formal criteria as a supplementary definition to filter out trivial phenomena and control the demarcation of the domain. He labels such an approach "a mixed functional-formal domain".

Beck (2016) expresses worry about the difficulty to connect comparative concepts with descriptive categories during comparison when comparative concepts are defined too abstractly in a semantic or functional domain, or when terms of descriptive categories are so unique that they can be only understood in a specific context. In the spirit to fill the gap between these two notions, he suggests more "portable terms" should be used. They refer to descriptive labels that are applicable for both comparative concepts and descriptive categories. Just as language-particular terms often describe the correspondences between certain forms and their function or meaning, it will suit the portability that comparative concepts are also defined in both formal and functional/semantic aspects (Beck 2016: 401).

This coincides with the beforehand-mentioned mixed functional-formal domain of comparative concepts.

1.3.3 Comparative workflow and terminology

Summing up the theory above, it is presupposed that pure formal categories are not cross-linguistically identical and hence they are not suitable to define a comparative domain across languages. Such a claim is developed based on the viewpoint of categorial particularism which argues that each language has its own grammatical categories and cross-linguistic comparison should not depart from these a priori categories. Instead, it is requisite to find a reliable crosspoint that is apt for all human languages and allows languages to be compared. In this sense, two notions, i.e., descriptive category and comparative concept, are created and introduced above. The former stands for language-particular categories or language-internal instantiations, while the latter refers to universal and language-external structures.

After an overview of the underlying theory, we shall now approach its realization and relevant terminology in this investigation. A language expression is always realized in a concrete verbal situation. A context serves as the substantial actualization of a language expression in a discourse. As discussed beforehand, a non-structural basis, or a comparative concept defined on an external domain, should be adopted to enable a cross-linguistic comparison. In this regard, contexts provide contextually situated and consistent utterances for the establishment of comparative concepts. Considering the current research object, i.e., content interrogatives, comparative concepts in this investigation are built on the basis of contexts in which information is inquired and correspondingly content interrogatives occur.

The functions of content interrogatives can incorporate multiple domains that are termed FUNCTIONAL DOMAINS by Miestamo (2007: 293). Functional domains are parts of a semantic or pragmatic function that is formally expressed across languages. Some of these functional domains are conspicuously distinct and widely familiar, whereas others can be rather specialized so that they are easily missed out. In this research, a function domain of content interrogatives is identified as a cluster of similar contexts. The goal of this investigation is to explore clustering of content interrogative contexts and propose some interrogative contexts that represent certain functional domains.

Subsequent to the establishment of comparative concepts, the next step is to instantiate linguistic structures used for the selected contexts in each language and make a comparison between them. In this stage, the substantial way to encode content questions, i.e., content interrogative units defined in §1.2.7, will be collected. They are language-specific constructions utilized in interrogative contexts and demonstrate various means to process content questions in different languages. The lexical and morphological differences of content interrogative units indicate the possible semantic distinction between underlying contexts. Through observing the formal realization of content interrogative units, it is possible to analyze similarity and differentiation between interrogative contexts across languages.

During the comparison, when the same interrogative unit is recurrently employed for a set of contexts in a language, these contexts might be associated in a certain sense and hence they compose an interrogative class. This notion can be equated to a descriptive category, given that these two concepts are both language-particular and contain all exemplifications of a comparative concept within a language. As it is impossible to equate descriptive categories cross-linguistically, interrogative classes are not identical to each other across languages either. Instead, they are only similar to different degrees.

If a group of interrogative contexts recurrently belong to the same class in all languages, they are regarded as constituting a universal interrogative category. All contexts pertaining to a category must share some overlapping properties and display a high degree of similarity in a functional domain. Thus, a universal category might reflect a certain kind of uniformity in human cognition.

1.4 Research questions

The main interest of this research is to investigate the diversity of content interrogatives across languages and uncover the underlying semantic distinctions between different kinds of question content. The goal is to gain new cross-linguistic as well as comprehensive languages-particular insight into the structure of content interrogatives. The central questions are as follows:

(1.26) a. Are there universal interrogative categories?

b. Specifically, are there domains (i.e., groups of interrogative contexts) for which the same coding is used within each language? If so, such a domain would be considered as an interrogative category.

c. What sub-domains are differentiated within each interrogative category?

d. Is it possible to identify a prototypical interrogative context ('specimen') that semantically represents a domain or sub-domain across languages?

During the data interpretation and language comparison, the more specific questions in (1.27) related to a certain grammatical aspect will also be taken into consideration:

(1.27) a. **Lexicon**:

- Within a language, which contexts are marked with identical interrogative forms?
- Across languages, which contexts are recurrently encoded with the same interrogative construction? Differently formulated, what kind of colexification appears frequently across languages?

b. **Markedness**:

- What kind of interrogative construction, e.g., a single morpheme and a compound structure with multiple elements, is preferred for certain contexts?
- Is the interrogative construction for a specific context consistently longer that others across languages?

c. **Morphology**:

- Given two different interrogative forms, what kind of language-internal similarity exist within a language (e.g., *wh-* in English)?
- Do such similarities recur across languages?
- What derivational pattern is employed for a certain interrogative construction?

d. **Syntax**:

- For a certain interrogative context, does the applied content interrogative cross-linguistically belong to the same word class?
- Can we observe some lesser-attested word classes, e.g., interrogative verbs?
- And in which contexts do these special word classes usually appear?

Apart from these questions, the other goal of the investigation is to test the feasibility to compare interrogative codings and to answer the questions above in the environment of a parallel text with the help of computational approaches.

1.5 Previous Investigations

1.5.1 Some issues about description of content interrogatives

A description of interrogative clauses is never absent in grammars. The length and concrete content of this part vary between languages, which depends on the complexity of the interrogative paradigm and the sources that the authors possess. However, as for the usage of a certain content interrogative, the most common information available in grammars is composed of just a few of examples, normally no more than three with a few lines of explanation. With such limited information provided in the grammar, readers can only obtain an approximation of the way in which the language constructs content questions.

Besides, some grammarians present the interrogative system on the basis of question words in English. Take the grammar of Nias Selatan as an example, Brown (2001: 128) simply lists the attested interrogatives while succinctly mentioning their shared initial element *ha*. The sequence of the list fails to manifest the morphological relationship between content interrogative of this language. For instance, although *haega* 'where' is apparently related to *haega iβaisa* 'how', these two expressions are placed far apart from each other in the list without explicitly pointing out the structural similarity. On the one hand, this type of arrangement in the grammar already satisfies the need to initially understand the content interrogatives paradigm in unfamiliar languages, especially those lesser-known. Yet, on the other hand, it undermines the exhibition of language-particular characteristics of content interrogatives.

A rash equation of content interrogatives in a lesser-known language with English could also lead to the pitfall of mismatches between form and function. For example, Dixon (2012: 415) notes that in Rukai there are "three 'how' interrogatives", i.e., *amokoa*, *apokoa* and *pikoa* (Zeitoun 2007: 375-376). Through his subsequent description, we know that these three question words carry different interrogative meanings. The first *amokoa* denotes a degree or quality, while the latter two indicate the means in the realis and irrealis mood, respectively. Despite the etymological relation implied by the common element *koa*, these three question words are deemed to be synchronically independent.[2] Yet, they are all translated as 'how' in English. This reflects the versatility of the form *how* in English. Also, this tells that the functional correspondence between these two languages is not one-to-one in terms of question words. If one just simply translates *how* as *amokoa* in Rukai without further examining the exact inquired content, it is very possible to bring the wrong interpretation.

1.5.2 Prominent research about content interrogatives

From a typological perspective, facets of content questions are provided with more detail in research specifically targeting interrogative clauses. Based on 79 sample languages, Ultan (1978) draws a general picture of the interrogative structure by means of classifying and portraying features of different question types. This work has been recurrently cited by the succeeding investigations of interrogatives. In this respect, it can be reckoned as a cornerstone for the modern typological study of this topic. Following this, Wąsik (1982) lists thirteen attested combinations of interrogative strategies with a small sample of seven languages. Sadock & Zwicky (1985) distinguish different sentence types and in this context introduce subtypes of interrogative clauses. The volume edited by Chisholm et al. (1984) also displays different kinds of interrogative clauses in seven languages. In contrast to the last three investigations, it is regrettable that no cross-linguistic comparison of interrogative structure is performed in this collection. The next influential discussion that provides an overview of interrogative construction and its grammatical properties is found in Siemund (2001), which is followed by König & Siemund (2007) with similar content. The latest general description that probably covers most detailed scopes of questions is given in the third volume of Dixon

[2] The meanings of the former part of these three question words, i.e., *amo*, *apo*, and *pi*, are not provided in the reference. It is not given whether these three question words are still morphologically analyzable.

(2012: 376-433). Compared to all the before-mentioned references, Dixon's introduction supplies a huge amount of examples from languages around the globe and thereby yields a much broader typological perspective.

Besides the overall illustration of interrogative clauses across languages, there are also studies revolving around the general interrogative structure of a certain area. There are a plethora of brilliant works with this theme. Some recent and comprehensive ones are, for instance, Luo (2016) demonstrates interrogative strategies in 138 languages of China, Köhler (2016) summarizes those in African languages, and Hölzl (2018) explores the diversity of interrogative clauses in over 450 languages and dialects in Northeast Asia with an ecological perspective. For the interrogative construction in signed languages, Zeshan (2004) provides a cross-linguistic elucidation.

In the typological literature above, content questions usually have to share the stage with other kinds of questions. Not to mention that the discussion dedicated to content interrogative units usually takes up just a small portion. In comparison to other interrogative strategies, it is insufficient to investigate the diversity of content interrogatives. In this respect, the attempt of some scholars to describe content interrogatives in particular languages is valuable, since it creates an important foundation for cross-linguistic studies. A non-exhaustive exemplification is as follows:

Authors	**Languages (Area)**	**Number of languages**
Muysken & Smith (1990)	pidigins and creoles	ca. 25
Mushin (1995)	Australian languages	26
Nau (1999)	European and Australian languages	ca. 19
Cysouw (2007)	Pichis Ashéninca and related languages	ca. 5
Lichtenberk (2007)	Oceania languages	ca. 24
Hengeveld et al. (2012)	Brazilian languages	24
Mus (2015)	Tundra Nenets	1

Table 1.1: Investigations of content interrogatives in certain languages

Among aspects of content interrogatives, their semantic categories have much less often stood in the center, compared to that most of the studies place the emphasis on the syntax of content interrogatives, e.g., word order in interrogative clauses (e.g., Dryer 2013), *wh*-movement (e.g., Cheng 1991), etc. Cysouw (2004, 2005b) makes the first endeavor to distinguish and generalize semantic features encoded in content interrogatives across languages. His survey also reveals the existence of further semantic specifications. Besides, through observing the recurrent lexical relations between content interrogatives, he tries to figure out the derivational connections between those categories and visualizes them with a map.

There are also various studies focusing on individual semantic categories. The opposition between HUMAN and NON-HUMAN is widely so conspicuous that it always comes to the attention of research relating to content interrogatives. Ultan (1978) already notices that this contrast is commonly attested in his sample languages. Lindström (1995) compares the interrogative pronouns equated to *who* and *what* in English with 3rd person pronouns grounded on the animacy hierarchy. Although the corpus is limited to 24 languages, Lindström (1995) starts the exclusive focal point on semantic distinctions of content interrogatives. Nau (1999) examines the morphological cases of *who* and *what* that indicate the semantic and pragmatic features of the referents. The differentiation between HUMAN and NON-HUMAN is then claimed as "almost universal" in Siemund (2001). By contrast, Idiatov (2007) conducts an extensive and insightful typological investigation in which he surveys the lack of differentiation between non-selective interrogative pronominals like *who* and *what* in a sample of 1850 languages. In this work, Idiatov innovatively discriminates functional aspects of categories PERSON and THING as well as classifies their prototypical and non-prototypical combinations. Idiatov (2007) sheds a light on the concrete studies of fine-grained functions of interrogative pronominals and confirms the feasibility of exploring content interrogatives from the perspective of lexical typology.

Compared to the domains of HUMAN and NON-human, the number of cross-linguistic studies is much smaller for other semantic categories of content interrogatives. Stolz et al. (2017) carry out the first systematic typology of spatial interrogatives with a sample of 450 languages. On the basis of a tripartite paradigm of spatial relations, as *where*, *whither*, *whence* in English, this study investigates the morphological patterns and structural complexity of

content interrogatives. It is also noteworthy that this work applies the translational equivalents of *Le Petit Prince* along with descriptive grammars as the data sources, which corroborates the practicability of the research method for the current thesis. Idiatov & van der Auwera (2004), Hagège (2008) and Lin (2012) get involved with interrogative verbs, a comparatively rare type of content interrogatives. The before-mentioned sophisticated semantic categories like UTTERANCE, ACTION, and meanings like 'what happen', 'do how' are in some languages denoted in those interrogative verbs.

Some effort has been made to find out the relations between semantic categories and the formal complexity of content interrogatives. Heine et al. (1991) propose that the phonological and morphological structure of content interrogatives is prone to mirror semantic features (see (1.14) in §1.2.4). Based on this, a hierarchy is established with 14 languages. According to this ranking, PERSON, THING, ACTIVITY and PLACE are coded in interrogatives with the minimal formal complexity, whereas content interrogatives of TIME and QUALITY have a more complex construction. The highest level of complexity is attested in interrogative forms of CAUSE and PURPOSE. Another similar hierarchy is produced by Mackenzie (2009) with a small shift of semantic domains (see (1.13) in §1.2.4) and subtler consideration of formal properties. On this occasion, the rank and position of PERSON, THING, PLACE, TIME and CAUSE remain the same in the hierarchy as in Heine et al. (1991). The differences lie in that ACTIVITY and QUALITY do not participate in the ranking, whereas MANNER and QUANTITY come up between TIME and CAUSE in the hierarchy.

Finally, content interrogatives can be associated with other grammatical categories. For example, Haspelmath (1997) and Bhat (2000, 2004) discuss the widely attested connection between indefinites and content interrogatives. Diessel (2003) is dedicated to the relationship between content interrogatives and demonstratives. Since this topic is beyond the scope of the present investigation, it will not be further discussed.

1.6 Outline of the book

This book is composed of six chapters, including the current Introduction. Chapter 2 presents the research method and semi-automatic approaches used in this investigation. Chapter 3 provides detailed information about the data of this study, including the corpus, the sampled languages and the procedures of data collection as well as processing. Chapter 4 extensively

discusses the clustering results of interrogative contexts. In Chapter 5, the derivational relations within and across the identified interrogative categories are demonstrated. Finally, Chapter 6 summarizes the main conclusions and gives some prospects for future research.

Appendix A lists the abbreviations, family and translational version of each sampled language. In the following Appendix B, interrogative contexts are presented in detail with their semantic label, cluster number and content.

2 Methods

This chapter will explain the methods used in the present study. In §2.1, I will introduce the idea of massively parallel texts which serves as the foundation of data collection and comparison of this thesis. Following this, relevant aspects of this method, i.e., corpora for language comparison, sources of parallel corpora, advantages and limitations of using parallel text, will be discussed from §2.1.2 to §2.1.5 in detail. Then in §2.2, I will describe the various semi-automatic approaches for data analysis.

2.1 Massively Parallel Text

2.1.1 Introduction

The research idea and data collection for this cross-linguistic investigation are based on the PARALLEL TEXT method. A parallel text is a text with its translational equivalents in different languages. On the basis of a parallel text, linguistic structures can be surveyed in a consistent context across languages using different contextually-embedded instantiations throughout the text.

It is not rare to use parallel texts in linguistic research and language comparison. Harris (1988) puts forward the term BITEXT for translation theory, which is frequently used as the synonym of parallel text. Since then, this concept has received wide attention from different linguistic domains, especially studies about language engineering (e.g., Somers 2001) and automatic translation (e.g., Tiedemann 2011). *Language Typology and Universals* (STUF) publishes a special issue in 2007, which extensively discusses different facets of this approach. It includes the general information about parallel text (Cysouw & Wälchli 2007), its pros and cons (Wälchli 2007), some cross-linguistic research exercising this method (da Milano 2007; Dahl 2007), practical parallel corpora (de Vries 2007; Stolz 2007) and statistical development (Cysouw et al. 2007).

In recent years, more and more studies employ the parallel text approach to compare linguistic structures across languages. For example, Dahl & Wälchli (2016) investigate the relationship between perfects and iamitives based on the translations of the New Testament in 1107 languages. Stolz et al. (2017) extract sentences containing content questions from translational equivalents of the novel *Le Petit Prince* in order to survey spatial interrogatives.

In Wälchli (2018), the New Testament is again recruited to study the temporal connectors in 72 languages varieties. And de Swart et al. (2022) draw data from the parallel text corpus *Europarl* in order to investigate the temporal construction 'not…until' in 21 European languages.

Typological or cross-linguistic research normally demands resources from a broad range of languages. In response to this reality, Cysouw & Wälchli (2007) propose the concept of a MASSIVELY PARALLEL TEXT (MPT) referring to texts that are available in translations in a large amount of languages and are consequently well-qualified for cross-linguistic investigations. This term will be used extensively in this research. Compared to other familiar approaches, such as questionnaires and interviews, MPTs provide a much more economical way to collect data in respect of money, time and labor. The most prominent theoretical advantage of MPTs is the contextual parallelism. Still, limitations also exist, which will be elaborated in the following subsections §2.1.4 and §2.1.5.

2.1.2 Corpora for language comparison

For a cross-linguistic investigation settling on MPTs as the research method, the very first step is to find a suitable text to build a parallel corpus. In the following content, I will always apply the term of massively parallel corpus, or parallel corpus for convenience, to refer to the material resource of MPTs, i.e., a collection of an original text and its translations. However, this might lead to terminological confusion. As Aijmer (2008: 276), Kenning (2010: 487-488) and Levshina (2022a: 131-133) explain, parallel corpora are a subset of multilingual corpora, while multilingual corpora also include another type called comparable corpora. There is a need to elucidate the differences between these notions, given that the description of the corpus used for the current study involves two traits, i.e., parallelism and comparability. Both types of corpora are designed to serve cross-linguistic research, whereas the respective characteristics decide their corresponding adequate domains.

According to Aijmer (2008: 276), a PARALLEL CORPUS is composed of a source text and its translational equivalents. A parallel corpus does not have to incorporate a large quantity of translations. Instead, it can contain a translation in just one other language. That is, the size of a parallel corpus is not strictly stipulated. Regardless of the size of a parallel corpus, one of the most salient features of this kind of corpus is that its structure is usually well-aligned,

which means the correspondences of a segment — either on the level of sentence or of word — are clearly identifiable in other translations. This allows to observe differential realization of a certain linguistic phenomenon between different languages. In summary, although the source text of a parallel corpus is translated into various languages, the content always remains the same.

On the contrary, texts from a COMPARABLE CORPUS are different both in language and substance (Aijmer 2008: 276). The reason that these texts are assembled to establish a corpus is that they resemble each other with respect to a certain property, which vary in line with the research purposes such as text genres, topics, or the timespan. Since the content of texts is inconsistent, no alignment regarding sentences or words can be achieved within a comparable corpus. Therefore, a comparable corpus is hardly practical if the research aims particularly at the cross-linguistic performance of a lexical item or a linguistic construction. Like every coin has two sides, the collection of texts for a comparable corpus is relatively easier than for a parallel corpus, considering that the source text does not need not to be identical and it is practicable to employ already accessible corpora to solve newly posed linguistic questions.

Both kinds of corpora are helpful in regard to comparability, i.e., to compare linguistic structures across languages. According to Kenning (2010: 496), these two types of corpora can bring many benefits if they are complementarily utilized in different stages of an investigation, e.g., a fuller picture of the observed phenomena. Yet, when it comes to a specific word or a linguistic expression as the research interest, comparable corpora, despite what its name asserts, are less useful than a parallel corpus, since parallel corpora provide a consistent contextual situation for comparisons.

2.1.3 Sources of massively parallel corpus

The nature of some non-fiction official documents, such as international laws, e.g., the *Universal Declaration of Human Rights* from the United Nations, and annual reports from big companies, implies that they are qualified to serve as MPTs. Firstly, they are attainable in multiple languages. Second, the usage of these textual materials demands that the translations must be precise and rigorous. And finally, they are normally available free of charge and easy to access online.

Nevertheless, the conspicuous shortcoming of official documents is that the available scope and topics in this kind of textual resources are usually confined. In other words, the number of instantiations that might be of linguistic interest is scant. Many linguistic domains fail to be exhibited in such texts. Besides, the requirement of translational preciseness leads to the loss of natural language use to some extent. A research targeting colloquial expressions will be disappointed in this kind of text. Moreover, most official documents are seldomly translated into lesser-used or endangered languages. This interferes with the ambition of most cross-linguistic studies to discover the worldwide diversity.

Literary works are another popular candidate for MPTs. Two of the most famous ones are *Le Petit Prince* by Antoine de Saint-Exupéry (see e.g. Stolz et al. 2017) and J. K. Rowling's *Harry Potter* (see e.g. Stolz 2007). In comparison with official documents, literary works are much more widespread worldwide and the multiplicity of content is markedly richer. Their translational equivalents are supposed to more closely represent the idiomatic and customary usage of target languages. However, the biggest hurdle preventing literary works from being applied to linguistic studies is the constrained accessibility due to copyright issues. In addition, the subjective decisions of translators, such as deletion of certain expressions or paraphrasing of particular content, might impinge on the parallelism between translational equivalents.

Religious texts are also a suitable source for a massively parallel corpus. Bibles, as above mentioned, are already utilized for many cross-linguistic investigations. Given that this investigation applies Bible translations as data sources, a detailed discussion about them will be given in §3.1.2 and §3.1.3. In comparison with other sources, many Christian texts provide the greatest typological diversity (Levshina 2022a: 134), e.g., the Bible and pamphlets of Jehovah's witnesses, since they are translated into a huge amount of languages worldwide by missionaries or missionary organizations. Yet, considering different customs and translational difficulties, this linguistic advantage is not found in all religious collections, e.g., the Quran or the Tao Te Ching are typically not translated for religious purposes. Due to the nature of religious texts, the quality of translation is mostly guaranteed. However, like official documents, religious texts can only represent formal language style. Besides, lots of modern terms are unavailable in those often archaic written texts.

Apart from the three above-mentioned categories, subtitles of movies is considered to be another potential source of MPTs, as they are also available in multiple languages and are easily accessible. Especially for researchers who are interested in direct speech or idiomatic expression, movie subtitles would be a good choice, given that the majority of content is composed of dialogs between characters and the language is stylistically more vivid and emotional (Cysouw & Wälchli 2007: 97; Levshina 2017 Levshina 2022b: 184). However, like all other kinds of massively parallel texts, the translation of film subtitles is also influenced by the source language. Also, the number of available grammatical phenomena is restricted in movie subtitles.

2.1.4 Advantages of MPTs

First of all, the textual environment of MPTs contains the corresponding contexts alongside the studied items. The traditional data sources, such as reference grammars and dictionaries, provide only a compact description of an expression or the linguistic structure. The concrete function or a subtle usage of research objects is often only described coarsely. Even if investigators intend to have a more detailed view of a given topic, the shortage of concrete contexts will obstruct the progress. The lack of context entails that the concrete instantiations embedded contextually are untraceable. Notwithstanding, these specific instantiations play a vital role in identifying particular functions of a linguistic structure. In comparison to the traditional reference grammars, the size and content of MPTs guarantee the amount and the multiplicity of contexts. The contextually-embedded situations of research items are therewith found and extracted with ease.

The uniform contextually-embedded situations provided in MPTs ensure the second merit — the parallelism of the investigated items. A comparison across languages is meaningful only when the research entity is consistent all along. The identical context of translations from a parallel corpus warrants such consistency, because the same source text is translated into diverse languages. By means of comparison between these aligned translations, one can obtain similar as well as divergent linguistic expressions that different languages utilize for a certain concept. Via the observation of these results, it is feasible to distill some linguistic patterns in terms of a linguistic phenomenon.

Thirdly, the shared origin of investigated objects can help to retrieve the missing or changed content. Various factors can give rise to unavoidable lost content in language translations, especially before the digitalization of paper documents was realized, such as the long-term inappropriate preservation, irremediable damages attributed to wars, natural disasters, or even the intentional sabotage. The personal preferences or understandings of translators can also cause alterations in content. It leaves even bigger regret when the text is only available in a mono-language or already extinct languages. With this pity, a number of classical and archaic texts are unable to make a contribution to linguistic studies, even though they could be treasures for studying the diachronic development of languages. When a text has been already translated into multiple languages, translational equivalents can be used as cross-references and thus it is possible to recover the lost or changed parts in the source text or other languages' translations.

Last but not least, MPTs are user-friendly by supplying a sizeable amount of language information. The primitive motivation of writing a grammar is to document a language. In the majority of grammatical descriptions, grammarians normally interpret a linguistic phenomenon with a couple of examples. Most often, this already meets the need of readers. Nonetheless, if the users of grammars do not rest on the interpretation from grammarians but intend to conduct an investigation in line with their own interest, for instance, to explore potential connections between some linguistic phenomena, or to dig into a finer-grained grammatical structure, it is then devoid of information in traditional grammars. In this sense, MPTs can be construed as offering raw materials that flexibly conform to the individual request of researchers.

2.1.5 Limitations of MPTs

Despite the benefits that MPTs bring to linguistic research, one has to be aware of limitations when using this approach. The nature of MPTs *per se* must be at first clearly stated: a text, no matter how lengthy and sizeable it is, cannot stand for a language in whole, but is rather just a collection of instantiations. Therefore, strictly speaking, a research adopting MPTs actually yields results from DOCULECTS, i.e., language varieties attested in concrete documentation (Cysouw & Good 2013: 342). One should bear in mind that such results do not necessarily represent the complete language structure, but instead language performance in a specific and

limited way. Only instantiations attested in MPTs are available to be involved in the investigation. Even though a research question may cover different subdomains, some of them have to be neglected due to the deficiency of available instantiations in a text.

Secondly, divergent source languages and base texts can have an impact on translation, which may blemish parallelism. It should be noticed that not all textual materials would encounter this problem. Works of literature, especially those popular or contemporary, has only one commonly recognized origin in one language which then serves as the base text for all translations, such as *Harry Potter*. The translators of this kind of texts need not bother with the choice of an undisputed source text. Besides, the purpose of translating this kind of text is mostly nothing but for the spread of literature and to meet the urge of readers globally. It means that, except for some mild adjustments in terms of expression, the pivotal content in translations ought not to undergo a huge alteration. As the opposite type, translation of texts of antiquity or with religious purposes is strongly influenced by the source language and the original version. By way of example, Bible, as one of the most noteworthy religious collections, comprises translations in Hebrew, Aramaic, and Greek which are used by different religious communities in line with their own historical context and textual traditions (see de Vries 2007: 151-153). As the result, no such text can be ascertained as the common basis for all current available biblical translations. The miscellaneous source texts can account for differences in biblical translations in respect of the number of verses, content, and interpretations of translators.

The last disadvantage to be discussed here is the counterpart of the last point in §2.1.4. As a kind of primary sources, massively parallel texts are short of grammatical descriptions or any elaborate analyzes when they are utilized for linguistic studies. In this sense, a MPT can be seen purely as a warehouse of language materials. In spite of the merit of the great number of data, researchers must always resort to grammars and dictionaries and then, based on knowledge acquired from these references, translate and interpret the examples in MPTs by themselves. Especially for those lesser-described languages, it will cost a huge effort to understand the detail of raw texts. Since it is impossible to flawlessly master a language just by reading the grammar, it is a risk that investigators might make mistakes during giving their own interpretation of an unfamiliar language or a language-particular phenomenon.

2.2 Automatic approaches

2.2.1 General notes

Traditionally, the procedure to investigate a linguistic phenomenon is that linguists collect language data based on the research question and hypotheses, process them, and then bring the data under scrutiny. Such observation is typically accomplished manually. However, as the amount of available material becomes larger and larger and more and more statistical methods are developed, it is possible to count on automatic approaches to conduct some bulky and repetitive work and achieve reliable results. One goal of this investigation is to classify content interrogatives extracted from discourse and find out the prototypical context for each interrogative category. With this aim and considering the size of data, the first phase of data analysis is performed with computational means.

In this chapter, I will introduce the automatic approaches applied to analyze the categorization of content interrogatives and visualize the output. However, it is important to note that the computational approaches only serve as a tool to conduct the data classification. Subsequent manual interpretation is still of essence to obtain results. Given that the statistical methods itself fall outside of the research scope, only a concise introduction of the computational tools and the procedures will be presented in the following.

Some facts are worth heeding due to the nature of automatic methods. First, apart from the convenience for data exploration, it is frequently required that the investigators have to set cut-off values or choose a suitable distance function by themselves resting on their knowledge of data. This entails that the output is highly subject to the decisions made by analysts on every step and thus might vary even if just one of such option changes.

Another inescapable consideration is that, although the underlying theories already stipulate the ground on which the results are drawn, the automatic procedure normally will not give any explanation for its decision. In consequence, researchers should always make effort to comprehend the results and qualitatively interpret the data structure afterward. It requires that investigators should have a good knowledge of the data.

Regarding the automatic outcome, investigators may encounter a pitfall that some results do not make any sense in the actual data, but rather is an artefact of this computational method. After all, the automatic output are the product of algorithms and manually pre-set parameters. It requires therefore extra discretion during the interpretation.

Finally, it has to be emphasized that the choice of computational methods to tackle data *per se* is a subjective decision. It is a wrong question to ask whether an analytic method is 'right' or 'wrong' but rather it is just a matter of suitability with respect to a specific goal or a certain type of data (Kaufman & Rousseeuw 2005: 37). While taking recourse to an automatic approach, one should be alert to its restrictions and clarify them in the interpretation. Accordingly, the results obtained in this way should not be equated to a undoubtably robust picture of a linguistic domain, but they only present the outcome yielded with a certain analytic mean.

2.2.2 Cluster Analysis

There are several techniques that can be used to detect the structure within a dataset. For this study, I adopt CLUSTER ANALYSIS to do the work. This method aids in classifying a given set of objects *n* by means of assigning them into a number of clusters *k* and thereby describes the data structure. It enables investigators to identify possible patterns within data, especially datasets that are too large to be analyzed manually. Cluster analysis is an exploratory tool and is especially useful to unearth the potential patterns. As opposed to other common methods to analyze data, no *a priori* hypothesis is compulsory to be made before using this approach to group the data primitives (Divjak & Fieller 2014: 406). Conversely, the descriptions gained from cluster analysis boost the generation of hypotheses (Moisl 2015: 7).

According to the type of output, various clustering methods can be generally separated into HIERARCHICAL and NONHIERARCHICAL kinds (Moisl 2015: 156-157). Hierarchical methods group the data from bottom-up and create a multilevel dendrogram. Each data object is seen as a separate cluster at the beginning. If two clusters are similar enough, they are merged together to form a bigger cluster on the next level. This procedure is executed successively as long as it is tenable. The output of a hierarchical method is presented as a clustering tree.

On the contrary, nonhierarchical methods demand that the number of clusters should be preset. All data objects are then grouped into these clusters. A classical algorithm of nonhierarchical method is called *k*-means. Its idea is that a set of centers *k* is created and designated for each cluster as prototypes, which are called CENTROIDS, and then every data point is allocated into a cluster (Kaufman & Rousseeuw 2005: 38; Levshina 2015: 317). In this way, *k* clusters are obtained. The distance between the centroid and each member of the

same cluster is supposedly minimal in average so that data objects of a cluster can be considered related. Each object can only be assigned to one cluster. Objects from different clusters are assumed to be heterogeneous. However, if the *k* value is set too high for an operation, some groupings might not necessarily mirror valuable distinctions but are consequent on trivial elements or meaningless decisions of the program (cf. Everitt et al. 2011: 9).

2.2.3 Grouping data

In this investigation, the data are composed of content interrogatives collected from different languages. Thus, data are not numeric but rather categorical. A quantitative partitioning cannot be conducted directly for these data, since they are language-specific expressions and the comparable qualities across languages have to be identified at first. In this sense, the comparison can only be conducted among the interrogative contexts that are consistent across languages. The comparison should begin internally with a language. According to the use of content interrogatives, it can be recognized which contexts are similar in a language. This procedure is conducted within each sampled language. Then, the cross-linguistic similarities between contexts can be further identified. On this basis, a quantitative method is applicable to classify contexts into groups to present the cross-linguistic pattern.

Considering the traits of the data and the goal of this study, I choose the nonhierarchical method of PARTITIONING AROUND MEDOIDS (PAM) to perform the cluster analysis. This method is developed from *k*-means approaches. According to the (dis)similarity calculated between data objects, PAM generates a dissimilarity matrix as the basis of the clustering. Compared to other *k*-means approaches, PAM uses MEDOID instead of centroid to be the prototype of each cluster. Medoids locate in the center of clusters as well. Different from centroids, medoids are not abstract points created by the computational calculation but are rather represented by objects from the actual dataset. The distance between the medoid and other members of the same cluster stands for the dissimilarity between them. Hence, medoids are deemed to be optimal to depict the real distances between data points and can also alleviate the distraction from outliers (Kaufman & Rousseeuw 2005: 71-7; Moisl 2015: 187, 191). This process is executed with the function *pam*() in the package *qlcMatrix* in R.

2.2.4 Visualization

Although PAM suggests the classification of data, the outcome is presented in a relatively abstract form. Without a clear display, it is still difficult to manually identify the detailed data structure and the proximity between data objects. To ease the understanding and interpretation, it is practical to visualize the results in the graphical form with the approach of MULTIDIMENSIONAL SCALING (MDS). In the current survey, this step is realized through the function *lmap()* in the package *qlcVisualize* in R.

MDS is a statistical technique that exhibits the (dis)similarities between analytical objects in a multidimensional space. To be noticed, although it is usually not explicitly noted, MDS is actually a generic term that incorporates variants developed for different concrete purposes and performed with different procedures. In most linguistic studies applying this method, MDS refers to the version of the classic scaling (van der Klis & Tellings 2022: 4). Based on the (dis)similarities between data objects, like the underlying theories of PAM, MDS depicts the structure of a dataset in a graphical shape. This program uses dots to represent data entities and arranges them into a coordinate space. In accordance, the distances between dots on the output map are contingent on the (dis)similarity values between data entities.

Inherently, the distribution of data entities is multi-dimensional. Especially for the dataset encompassing a large set of information, the spatial representation of the structure will be complicated. In terms of this issue, MDS applies the dimensionality reduction technique to determine the optimal dimensionality for the representation of datasets with complex multi-dimensions. Normally, the most two or three significant dimensions are selected. As the result, the output of an MDS analysis can be plotted on a 2D or 3D map. The mathematical background will not be further introduced here, since it is not the focus of this study. A comprehensive explanation of the theories and practices of MDS can be found in Levshina (2015: 333-350) and van der Klis & Tellings (2022).

The closeness between points on an MDS map indicates the similarity between the corresponding linguistic objects. As opposed, the less the objects resemble, the further apart the corresponding dots locate from each other in the coordinate space (Everitt et al. 2011: 37). When the proximity of a set of points is structurally identifiable on an MDS map, which sometimes will be delimitated by adding contour lines, these points can be considered to constitute a cluster and the corresponding entities in the dataset might share a certain

linguistic property. Moreover, the points of a group are supposed to be contrasted with those of other groups to a certain degree, which reflects the commensurate disparity between the corresponding instantiations.

With the help of indications generated by MDS, researchers can discover distributional patterns of data entities and, in the further step, explain linguistic correlations within the dataset. Same as the approach PAM, MDS does not provide a description or interpretation of the grouping of data either. It just reveals the potential correlations among data points that assemble on the map (Wälchli & Cysouw 2012: 682).

3 Data

In this chapter, information about data of this investigation will be given. I will first introduce the data corpus in §3.1. Next, I will present the sampling strategy and the sample languages in §3.2. In §3.3, the procedure of data extraction will be described. The morphological processing of the extracted data will be illustrated in the following §3.4. Finally, a brief introduction of online repository of data will be given in §3.5.

3.1 Corpus

3.1.1 The Parallel Bible Corpus

For the current study, I use the Parallel Bible Corpus for data collection. As its name indicates, this corpus consists of translations of Bible texts in numerous languages. In the following, I will first give a brief introduction to the Parallel Bible Corpus. Then I will discuss strengths and potential problems of using Bible translations for cross-linguistic investigations.

The Parallel Bible Corpus is constructed by Mayer & Cysouw (2014) and is available in Github.[3] Currently, the corpus encompasses 2000 different biblical translations in 1460 languages varieties. Among all these translations, 54741 unique verses from the Old Testament and 7958 verses from the New Testament are found in total. For the current study, data are only collected from verses of the New Testament.

The original raw texts are extracted from accessible sources and then properly prepared so that the consistent format can facilitate the computational processing for further comparison. The texts are prepared in *.txt* files and all file names are structured with the pattern of 'ISO-x-bible-TRANSLATION'.[4] The abbreviation *ISO* refers to the language code of the translation in ISO 639-3 standard.[5] TRANSLATION is an optional disambiguating suffix. It is especially convenient if a language boasts more than one version.

As the result of the pre-processing, all texts in the corpus are tokenized as well as standardised in Unicode. Meanwhile, words are separated from punctuation and non-

[3] URL: https://github.com/cysouw/parralleltext. Due to copyright issues, the online repository of the Parallel Bible Corpus is not public. Access can be obtained by contacting Michael Cysouw.

[4] If a language has only one translational version, the file name will be 'ISO-x-bible'.

[5] Abbreviations of the sampled language are provided in Appendix A.

alphabetic symbols by means of interpolating spaces between them, as instantiated in Figure 3.2 below. This step was predominantly performed automatically. However, a manual check was also applied in order to find errors that might emerge during the automatic preparation. The manual check also examines the special usage of non-alphabetic symbols in some languages in which those symbols actually serve for the orthography or represent sounds in the language.

Translations in the Parallel Bible Corpus are composed of two parts, i.e., lines with metadata and text of content-related verses. At the head of each document, eleven lines provide the basic information about the corresponding translation. These lines are labeled by hash characters, meaning that they should be excluded from the core data. An example from the english translation *eng-x-bible-etheridge* is given in Figure 3.1.

```
# language_name:        English
# closest_ISO_639-3:    eng
# ISO_15924:            Latn
# year_short:           2010
# year_long:            1846: Four Gospels and the Apostolical Acts and Epistles<br>1849: Remaining Epistles and the Book of
Revelation<br>2010: published online
# vernacular_title:     The Peschito Syriac New Testament
# english_title:        The New Testament in English
# URL:                  http://sourceforge.net/projects/zefania-sharp/files/Bibles/ENG/Etheridge/
# copyright_short:      Public Domain
# copyright_long:       Source: http://www.peshito.com/<br>Rights: Public Domain<br>The Peschito Syriac New Testament by J.W. Etheridge,
comprising: The Syrian Churches: their Early History, Liturgies, and Literature. With a literal Translation of the Four Gospels, from the
Peschito, etc. 1846  and The Apostolical Acts and Epistles from the Peschito or Ancient Syriac: to which are added, the remaining epistles
and the Book of Revelation, after a later syrian text, etc. 1849.
# notes:
```

Figure 3.1: Metadata of the Bible translation *eng-x-bible-etheridge*

As exhibited in the Figure 3.1, the first and second line refer to the name of language and its ISO 639-3 code, which is congruent with the first unit of the file name *eng-x-bible-etheridge*. The subject of the third line ISO-15924 stands for the script of the text. For instance, this english text employs Latin alphabet. The next two lines tell the time when the text is published. The complete title of the Bible translation is given in the corresponding language as well as in English in the sixth and seventh line. Since the original texts are often gathered from online resources, it is necessary to provide the original web address, as showed in the eighth line. The URL is especially helpful if one needs to retrieve the original during the manual check or there are missing or problematic verses in the translation. Finally, the copyright information and notes are given in the last three lines.

Following the meta-features are verses with the actual biblical content. Every line in this chunk starts with a string of digits serving as the identification of the verse. It is divided from the text by a TAB. Figure 3.2 presents a glimpse of the format.

```
40001001    Buch der Abstammung Jesu Christi , des Sohnes Davids , des Sohnes Abrahams :
40001002    Von Abraham stammte Isaak , von Isaak stammte Jakob , von Jakob stammten Juda und seine Brüder .
40001003    Von Juda stammten Phares und Zara aus der Thamar , von Phares stammte Esrom , von Esrom stammte Aram .
40001004    Von Aram stammte Aminadab , von Aminadab stammte Naasson , von Naasson stammte Salmon .
```

Figure 3.2: Verses of the Bible translation *deu-x-bible-pattloch*

A notable trait of the format in this corpus is that all verses are consistently designated to a structured form of digits. The numeral identifiers are important to maintain the parallelism between translations. A string of digits can be split into three parts. The initial two digits stand for the number of the book to which the verse belongs.[6] The numbers of the books in the New Testament used for this study are ranged from 40 to 66. The chapter of the verse is then represented by the next three digits. The final three digits of the ID indicates the verse number. Take the verse ID 40001001 of the first line in Figure 3.2 as an example. This string refers to the first verse of the first chapter of the Gospel of Matthew in the New Testament.

Although the system of verse ID functions problem-free for the most part in this corpus, there still exists disorderliness in some translations. Occasionally, the content of multiple verses is merged into a single one. In this case, the ID strings of other involved verses are registered as usual, but the content part remains empty. Another situation is that a verse is completely nonexistent in a translation due to various reasons, such as different source texts or being missed out by translators. This time, the ID of this verse is absent in the corresponding text. It also occurs that some verse IDs in a translation are corrected by the compilers of the corpus, because different IDs are assigned to those verses in other translations. Yet, this case might not be an error or even a deliberate arrangement of the translator. However, it is necessary to alter those IDs in order to remain the parallelism between translations.

[6] A comprehensive list of books and the corresponding verse numbers can be found in Table 2 of Mayer & Cysouw (2014: 3162).

3.1.2 Advantages of Bible translations

It is not novel that Bible texts are used for language research and corpus-based investigations. Especially before electronic approaches were possible, Bible texts are one of the most important source of cross-linguistic corpora along with grammars and dictionaries. During the pre-electronic era, Bible texts have a long history to be used for producing concordances. With regard to the considerable supply of language material, Bible concordances can be seen as "one of the first pieces of corpus-based research with linguistic associations", as remarked by Kennedy (1998: 13). Meyer (2008: 1-2) lists some noteworthy contributions from this time. Already in the 13th century, Cardinal Hugo created a concordance of the Bible in Latin. Then in the 15th century, Isaac Nathan ben Kalonymus wrote a Hebrew concordance of the Bible. The concordance compiled by Alexander Cruden in the 18th century is regarded as one of the most extensive enterprises. Based on the King James Version of the Bible, this work contains astonishingly 2,370,000 words in total. The content not only consists of the common and proper nouns, but also includes function words and particular collocations. Besides, Cruden patiently points out the location of every entry in the Bible, while different forms of a word are separately listed as entries. These efforts reflect the groundbreaking value of Cruden's concordance for the linguistic investigations back at that period.

Several features of the Bible entail it being one of the most attractive linguistic source for hundreds of years. In terms of the available data, it provides a vast size of texts. The whole Bible incorporates 60 books with approximately 800,000 words in total (Christodouloupoulos & Steedman 2015: 377). Within the Parallel Bible Corpus, translations include 10707 verses on average (Mayer & Cysouw 2014: 3158). Although this number is exceeded by a lot of modern corpora that encompass different genres, it still outweighs most literary works.

Not only the size of data is undoubtedly huge. Given the religious purpose of the Bible, the number of languages into which the Bible is translated also considerably exceeds other textual material. Considering the languages globally, Bible has been partially translated into more than one-third of over 7000 living languages, while nearly 7% have a complete version of translation (Mayer & Cysouw 2014: 3159).

Along with the quantity, the variety of target languages is impressive with regard to the diversity of the Bible translational equivalents. The Bible is widely spread by missionaries to remote and isolated regions where the lesser-known or even endangered languages are

spoken. Missionary activities compelled the necessity to translate the Bible into local languages, which conduces to the documentation of these lesser-described languages. In contradiction to other parallel or comparable corpora of which the translation equivalents have a heavy bias towards major or well-known languages, a corpus of biblical texts contains sampled language more broadly and evenly distributed over the world.

Another strength of using the Bible as data is its potential for conducting investigations of different linguistics topics. Besides the general disciplines, such as syntax, morphology and semantic, the large and principled collection of biblical texts can also benefit investigations with some specific linguistic interests. For instance, the Bible is an invaluable resource for the study focusing on the diachronic development of languages. Especially for languages with a huge number of speakers, it is not uncommon that there is more than one translation spanning a long time. For example, there are 33 translations of English in the Parallel Bible Corpus and the year of compilation ranges from 1611 to 2013. These translations document the usage of language at different period and provide an excellent opportunity to observe how languages shift over time.

With the development and application of technology, researchers have found that the Bible provides a reliable frame for the establishment of a modern comparative corpus. According to the identification of book, chapter and verse, the content of the Bible across translations is clearly traceable. The computational processes help with the tokenization of texts and guarantee an alignment between the Bible translations, while the parallel structure of the Bible texts eases the automatic operations and enables more research possibilities. A powerful example of such research opportunities is the Parallel Bible Corpus described previously in §3.1.1.

3.1.3 Potential issues of Bible translations

Though Bible texts are viable for linguistic studies in many ways, their nature determines that we must take heed of the issues that might appear during the investigation and impact the results. A major concern is that the language of the Bible is antiquated. Indeed, the Bible as a work created thousands of years ago, its language and writing style cannot be representative of the modern usage. If the research goal is strictly the present-day language, one must be conscious of this limitation. In this case, researchers had better turn to other types of sources

or choose the biblical translations compiled in recent years if available. However, for many languages that are lesser-described and have few available resources, there is no better alternative than the Bible, considering its abundant information.

Another disadvantage related to this issue is that it may have a shortage of common and modern subject matters (cf. Resnik et al. 1999). Although there exist scenarios related to daily life in biblical texts, most topics and stories in the Bible are about religion. Besides, it is impossible that the Bible texts from archaic times discuss concepts in today's society.

Since Bible translations are constantly recruited in linguistic investigations, researchers should also give thought to the complication that arises from the translation process. Three main factors could have an influence on the translational outcome, i.e., source texts, methods of translation, and translators.

As shortly mentioned in the previous discussion about the disadvantages of MPTs, different religious communities acknowledge divergent versions of the Bible. Even within a religious group, there is a considerable variation. This issue is concluded by de Vries (2007) as TEXTUAL MULTIPLICITY (de Vries 2007: 151-153). Moreover, de Vries (2007: 153) puts forward another problem similarly related to the intricate Bible source, namely CANONICAL MULTIPLICITY. Given the various histories, beliefs and needs of different religious communities, the canonicity transmitted by the Bible versions diverges too. Its degree can vary from canonical and deuterocanonical to apocryphal. These two aspects set difficulties already at the inception stage of translation.

Last but not least, the TRANSLATIONAL MULTIPLICITY (de Vries 2007: 153-156) is a consequence of the subjective decision of translators as well. This issue does not solely happen to the Bible but is inevitably derived from the nature of translation. Regarding an expression in the source text, different translators could apply various translations in the target language. Individual language sense and convention decide that translators would omit or add elements during the translation. The simplification or explicitation of the content is sometimes conducted, as translators might see the necessity in a given cultural environment or consider

some particular groups of audience. All these translation changes result in that no direct equivalence between translations can be doubtlessly assured.[7]

Another facet related to the translation is the usage purposes. For religious communities and readers who mostly value the orthodox quality of the Bible, the translation should be strictly conducted word-to-word. However, the intention of a lot of translation works is rather to send on religious missions and promote the belief to the larger public. Hence, translators would turn to the method that places stress on making the content easier to be accepted. In contrast with the former approach, the latter puts more effort to preserve the sense of the text. Just as principles suggested by Nida & Taber (1982: 14-15) for the Bible translation, contextual consistency takes precedence over verbal consistency. In this way, given that the semantic scope of a word does not completely overlap with its concordance in another language, the sense-for-sense approach will definitely engender adjustment to expressions and affect the verbatim equivalence.

3.2 The Sample

3.2.1 General considerations

Cross-linguistic investigations have the ambition to explore and reveal the diversity of languages as extensively as possible. However, it is a good wish but also practically unrealistic to bring every language into a study. On the one hand, among all human languages, only a small portion of them have already been described. There are still a lot of languages awaiting proper documentation around the world. Many of them are even extinct before ever being known. On the other hand, descriptions of many lesser-used languages were either compiled a long time ago, i.e., it might be a daunting issue for today's investigators to adapt to the old grammatical tradition, or the quantity of grammatical phenomena is sparse in grammars, which leads to that they can hardly fulfil the need of research. Although there are also numerous languages with available and applicable references, a respectable number of

[7] According to Wälchli (2010: 333), it is difficult to attest the complete semantic identity in translational equivalents due to unavoidable changes in the translation. Actually, no one can say that translations of a text are strictly 'same' in any aspect. This reality decides what we are pursuing is that the analytical entities identified across languages resemble in terms of the meaning as much as possible. Only under this circumstance are they comparable. As further pointed out in the same paper, the contextually-embedded situations provided by well-organized parallel texts qualify the semantic similarity of entities to a great extent.

them cannot be involved in a linguistic survey. Therefore, it is necessary to select an attainable group of languages as the sample based on which we then can perform an investigation. Building up a dependable and representative sample at an economical cost becomes thus an important step of a cross-linguistic work.

For studies with different objects and goals, the sources and sizes of the sample vary. Therefore, it is essential at the beginning to choose an adequate strategy for the sampling. Three major types of sampling methods are commonly introduced in literature, i.e., variety sampling, probability sampling, and random sampling. The detailed descriptions of the corresponding research questions and precise sampling procedures can be found in Bell (1978), Dryer (1989), Perkins (1989, 2001), Rijkoff et al. (1993), Rijkoff & Bakker (1998), Maslova (2000) and Dahl (2001). Yet, considering that the sampling of the present study is largely dependent on the availability and quality[8] of texts in the Parallel Bible Corpus, I cannot strictly follow any of these well-stratified sampling techniques. Instead, I can only, for a practical reason, construct a sample based on what is accessible. Such a strategy is conventionally labeled CONVENIENCE SAMPLING (Song 2001: 20; Cysouw 2005a: 555; Bakker 2011: 106) with which researchers just take the obtainable and, of course, reliable data without any strict prerequisite.

In the course of the diversity of the sample, two criteria play a crucial role — the genetic relatedness of languages and their areal distribution. Although I have tried to maintain the equilibrium between the genetic relatedness and areal distance of the sampled languages as much as possible, all biases cannot be eliminated. Some language families are undeniably underrepresented. The main obstruction is on account of the bibliographic deficiency, which is reflected in three aspects in the current survey.

First, as mentioned above, little or even no grammatical documentation can be found for a plethora of languages, especially those spoken only by a small number of people or in isolated areas. The same situation is found in over 1400 languages of the Parallel Bible Corpus. Even

[8] The 'quality' here refers to the number of verses and the use of content interrogatives in the New Testament. As beforehand discussed, due to multiple factors, the translational outcome of the Bible is complicated and varied. It is not rare that a biblical translation loses a big amount of verses. Or, in terms of my research object, content questions are paraphrased during the translational process, which leads to the omission of content interrogatives. For example, the question *He asked 'when is the dinner?* can be reformulated as a declarative sentence *He asked the time of dinner*. In this case, the example will be regarded as unsuitable for this study and is excluded from the sampling.

though it would definitely boost the balance of genetic and areal parameters by including those languages in the sample, the deficiency of references impedes the possibility.

Secondly, the bibliographic bias is also in relation to the type of the data source of this investigation. The goal of the present study is to explore the diversity of content interrogatives under comparable circumstances. With this aim, the Massively Parallel Text is employed as the research method. The collection of content interrogative units should be conducted with the precondition that they are applied for the same context in different languages. Different from the typical cross-linguistic works that collect data straightly from grammars or wordlists, the extraction of content interrogatives for this research is performed directly from authentic texts, even in languages with which I am not familiar. Since the translational equivalents in the Parallel Bible Corpus are not glossed, i.e., they are just raw language materials, the quality and quantity of available grammatical information are pivotal for the correct extraction and analysis of research items. Many small and lesser-described languages have merely a grammar sketch, which is usually acceptable to solve many linguistic questions. Yet, the way in which the data are collected in this study decides that such a short grammatical description may be insufficient to discerning the rest of clause constituents in the clause. The precise analysis of content interrogatives will be inevitably constrained by such an information shortage. Therefore, many languages are lamentably disregarded from the sample.

The last factor associated with bibliographical issues during the sampling is the orthography of object languages. The writing system varies across languages. A sizeable amount of lesser-spoken languages only have one biblical translation in their own writing system, whereas the already scanty grammatical information about these languages is usually documented in English. It leads to the dilemma that such references cannot aid the reading of the original text. For this research, except for Chinese, Korean and Japanese whose orthography is familiar to me, I mainly choose translations that are available in Latin script, due to the convenience to read the script as well as to consult grammars and dictionaries. Owing to the conceivably huge consumption of time for reading the text, the unfamiliar languages are excluded from the sample.

Aside from bibliographical difficulties, another restriction of the sampling comes from the translations. Under certain circumstances or in some cultures, addressing a question directly may be seen as impolite. It is also illegitimate in some languages to inquire information by

means of asking questions. Instead, one would choose other ways to acquire absent knowledge. For example, in the scenario of greeting, one can first introduce oneself in order to imply that the other party should provide his/her identity information too. Another common situation in this regard is that speakers are prone to request information by addressing indirect questions, which causes the absence of content interrogatives. These issues are very language-dependent and culturally variable. During the translation, such factors must be taken into consideration. Translators may thus make adaptations in order to conform to the appropriate expression. It gives rise to the consequence that, despite the same context, content interrogatives are replaced or absent in some translations. If the translation of a certain language has gone through too many adjustments or omissions of content interrogatives, this language is not included in the sampling.

3.2.2 The sampling strategy of this study

On the basis of the convenience sampling, the strategy of this investigation tries to balance the genealogical and areal distribution of the sampled languages, despite the restricted number of applicable translations. Besides, it is also expected that a sampled language can exhibit linguistic peculiarities shared within its family.

However, it is tricky to decide which language is 'typical' enough to represent the whole language family. On the one hand, significant linguistic characteristics of a family cannot be wholly displayed by a single language. On the other hand, the degree of similarity between languages varies between different families. This means that in some families the member languages show a high resemblance, whereas in other families, especially those with a large size, there are still noticeable linguistic differences between related languages, even though they are classified into the same family.

Considering such a case, the sampling strategy of this study is based on language families, which can be labeled **family sampling**. In this spirit, the sampling starts with selecting various languages families. Then multiple languages were selected from each family into the sample. This approach intends to, on the one hand, balance cross-linguistic and family-internal diversity and, on the other hand, observe the internal pattern within each language family.

Family	Language
Sino-Tibetan (2)	Mandarin Chinese (cmn), Yue Chinese/Cantonese (yue)
Altaic (6)	Halh Mongolian (khk) , Gagauz (gag), Karakalpak (kaa), Turkish (tur), Uyghur (uig), Turkmen (tuk)
Eskimo-Aleut (2)	Central Yupik (esu), North Alaskan Inupiatun (esi)
Austro-Asiatic (4)	Car Nicobarese (caq), Eastern Bru (bru), Parauk (prk), Vietnamese (vie)
Austronesian (12)	Balantak (blz), Balinese (ban), Batak Karo (btx), Chamorro (cha), Ma'anyan (mhy), Madurese (mad), Makasar (mak), Acehnese (ace), Iban (iba), Indonesian (ind), Jarai (jra), Tagalog (tgl)
Khoe-Kwadi (1)	Nama (naq)
Niger-Congo (16)	Northern Kissi (kqs), Noon (snf), Maasina Fulfulde (ffm), Wolof (wol), Baoulé (bci), Igbo (ibo), Toro So Dogon (dts), Dii (dur), Northern Dagara (dgi), Ejagham (etu), Nomaande (lem), Masaaba (myx), Tharaka (thk), Rundi (run), Kagulu (kki), Nyanja (nya)
Uto-Aztecan (5)	Lowland Tarahumara (tac), Tetelcingo Nahuatl (nhg), Western Huasteca Nahuatl (nhw), El Nayar Cora (crn), Hopi (hop)
Eyak-Athabaskan (2)	Dogrib (dgr), Gwich'in (gwi)
Iroquoian (1)	Cherokee (chr)
Mayan (3)	Yucatec Maya (yua), Tabasco Chontal (chf), Chuj (cac)
Mixe-Zoque (4)	Highland Popoluca (poi), Francisco León Zoque (zos), Coatlán Mixe (mco), Totontepec Mixe (mto)
Tupian (3)	Sirionó (srq), Paraguayan Guaraní (gug), Tenharim-Parintintin-Diahoi (pah)
Quechuan (3)	Inga (inb), Ayacucho Quechua (quy), Huallaga Huánuco Quechua (qub)
Arawakan (4)	Yine (pib), Garifuna (cab), Parecís (pab), Machiguenga (mcb)
Maningrida (1)	Burarra (bvr)
Pama-Nyungan (2)	Kuku-Yalanji (gvn), Western Arrarnta (are)
Uralic (3)	North Saami (sme), Finnish (fin), Hungarian (hun)
Indo-European (12)	English (Eng), German (deu), Dutch (nld), Icelandic (isl), Danish (dan), Spanish (spa), Catalan (cat), Romanian (ron), Welsh (cym), Irish (gle), Czech (ces), Croatian (hrv)
Isolate (2)	Korean (kor), Japanese (jpn)

Table 3.1: The sampled languages

The sample of the investigation consists of 90 Bible texts in 88 languages. A list of the sampled languages is provided in Table 3.1 above. In order to enlarge and ease the selection of eligible contexts containing content interrogatives at the beginning of data collection, two translational versions have been chosen in German and English, respectively. Yet, these versions are established in different years. The two translations in German are published severally in 1951 and 2014, while the English translations are from 1976 and 2011. The intention is to mirror some shifts of content interrogative in these two languages.

This study is not confined to the use of content interrogatives in a confined version of time, not will there be any attempt to look at direct diachronic changes. The main reason is that in most cases there is no alternative for languages with only one translation. Although the source language may influence the translational outcome, such a concern is not taken into consideration during the sampling due to the scarcity of relevant information which sources were used in the preparation of each translation.

Compared to other cross-linguistic investigations, the sample scale of this study is obviously much smaller and limited. A major reason is that the manual extraction of content interrogative units is exceedingly time-consuming, which does not allow me to cover more languages within the limited time. Some biases in terms of language families and genealogical relationships are unavoidable.

3.3 Data collection and processing

3.3.1 Locating interrogative contexts

The data collection began with the selection of qualified contexts of content interrogatives from Bible translations. I chose seven translations in five languages to commence collecting useful instantiations. Besides four translations in English and German, as aforementioned in §3.2.2, the other three texts are written in Spanish, Mandarin, and Cantonese, respectively. The reason for the choice of these translational equivalents as the start point is that these languages are easier for me to read. Also, there are already numerous grammars and references in which the content interrogatives in these languages are thoroughly described.

Then, I located verses in which content interrogatives occur and identified the corresponding interrogative units. The targeted content interrogatives were manually entered in the second column between the verse ID in the first column and the actual textual content

in the third column. A TAB is applied to separate columns. A screenshot is given in Figure 3.3 to exemplify the format.

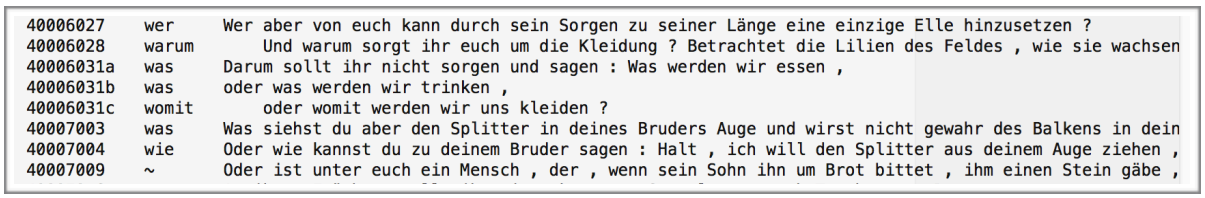

Figure 3.3: Format of the extracted data

Two situations presented in Figure 3.3 need an explanation. If a verse contains more than one content question, it will be manually divided into the corresponding number of sub-verse. The verse ID remains the same, which is only extra coded with a letter, as is demonstrated from the 40006031a to 40006031c.

Sometimes a translation does not use content questions in a particular context as the other sampled languages do. In this case, the symbol tilde is applied as the placeholder in the second column, which will be ignored during the automatic comparison. The verse 40007009 in Figure 3.3 shows such an instance.

3.3.2 Extracting interrogative contexts

After identifying contexts of content interrogatives based on the initial five languages, the next step is to extract the qualified contexts from translational equivalents in other sample languages. The consistent verse ID in the first column eases the work. According to the verse number of the selected contexts, the Unix shell automatically extracted the corresponding verses from the target translation and wrote them into a new document. Then, I repeated the manual identification of content interrogative units with the aid of grammars and literature. The well-processed documents in this step are named on the pattern of 'check-ISO-bible/ TRANSLATION' in order to be distinguished from the original raw texts.

I also adopted the statistical approach FAST-ALIGN (Dryer et al. 2013) to assist in the identification of content interrogatives, especially those in lesser-described languages. This procedure is executed in Unix shell. The idea of this technique is that based on a parallel text the algorithm identifies the corresponding words in a sentence between a pair of translational equivalents (Tiedemann 2011: 1, 59; Mayer & Cysouw 2012). However, every suggestion

made by Fast-Align is manually examined and decided whether to be accepted. A recent application about word alignment in the Parallel Bible Corpus can be seen in *ParCourE* (Imani et al. 2023).[9]

3.4 Morphological processing

3.4.1 Subdividing complex forms

After the steps presented in §3.3, the collection of content interrogatives from the sampled languages is basically finished. Before running the automatic comparison of data, they have to be polished, since content interrogatives in natural texts are morphologically complex to various degrees in different languages. Normally, grammars only provide the basic forms of content interrogative units, whereas in the real discourse they can turn up in a morphologically much more complicated structure.

For manual analysis, the compositionality of a linguistic construction is in most cases unproblematic. Based on their knowledge, trained linguists are able to analyze a complicated structure and pick out the crucial segments. However, unlike the human brain, the computational programs used for the current study only recognize the input objects as relevant or similar if they are orthographically identical in the form. That is, even though the base stem of two content interrogative units is the same, different elements that emerge in the construction might lead to that the automatic comparison fails to detect parallelism or similarity between input entities.

Regarding this issue, the solution is to split the complex structure of content interrogatives into reasonable pieces so that the computational program is able to read all elements of the construction one by one. On this basis, the program can then assesses the degree of resemblance between input entities. Therefore, the authentic instantiations of content interrogatives extracted from original texts must be first manually segmented into components.

[9] URL: http://parcoure.cis.lmu.de/

3.4.2 Morphological structures of content interrogatives

In previous research on content interrogatives, linguists have already realized their structural complexity. Mackenzie (2009: 1140) outlines several possible types of morphosyntactic relatedness referring to content interrogatives. An interrogative stem can present with inflectional affixes. For instance, in Parecís z*ale-nae* 'who.PL' consists of the base z*ale* 'who' and the plural suffix *-nae* (Brandão 2014: 159). Reduplication is another common strategy to produce a content interrogative based on other interrogative words, such as *wanja-wanja* 'when' in Kuku-Yalanji, which is derived from *wanja* 'where' (Patz 2002: 81). Derivation is attested frequently in building content interrogatives as well. An example is found in Ayacucho Quechua that *may-kama* 'how far' is composed of *may* 'where' and the suffix *-kama* 'until' (Zariquiey & Córdova 2008: 98, 101). A content interrogative can also be a multi-word expression. An example of this kind can be found in Northern Kisi *wɛ̀ɛ lɛ̀ɛlɔ̀ɔ̄* 'what time/when' (Childs 1995: 111), which is composed of two separate words.

Furthermore, etymological relatedness is a widespread phenomenon across languages in the world. Take interrogative words *wo*, *wohin* and *woher* in German as an example. *Wohin* 'whither' and *woher* 'whence' are diachronically connected to the basic question word *wo* 'where'. The directional specification is further indicated by *hin* and *her*. These two directional indications have already been fused with the stem word.[10] Compared to *where from* and *where to* in English, the compositionality of *woher* and *wohin* in German is less clearly perceived by speakers. Another extreme example is displayed by the question word *warum* 'why' in German. Historically, *warum* is composed of *wo* 'where' and *um* 'in order to'. Nevertheless, two components are fused in the modern standard German so that the word *warum* is no longer deemed polymorphemic but rather monomorphemic. The opposition between the diachronic origin and synchronic perception brings about the question to whether it is still meaningful to segment elements of a contemporarily simplex word with the intention to reflect its historical development, and if yes, to what extent the decomposition should be conducted.

Another issue might lead to the difficulty to understand interrogative construction. It is cross-linguistically recurrent that a part of a content interrogative unit originates from another

[10] Yet, it is still possible to ask with the expression *Wo gehst du hin?* 'Where are you going?' in which the directional indication *hin* is detached from the question word *wo*. The questions *wohin* and *wo […] hin* have no significant difference in terms of the meaning.

basic question word, whereas the rest element of this unit has no actual meaning nor a grammatical function. Such an element only serves to distinguish a word from others, i.e., its value is purely contrastive (Haspelmath & Sims 2010: 2). Aronoff (1976: 10) dubs this case as CRANBERRY MORPH, which Hölzl (2018: 77-78) has also encountered and made a discussion. The interrogative word *imanir* in Huallaga Quechua is an example in point (Dixon 2012: 383). This word is used to ask for reasons. The first part of *imanir* comes from the interrogative base *ima* 'what'. Notwithstanding, no grammatical function or semantic meaning is assigned to the subsequent constituent *nir*. This component is not attested in other words either. That is, the element *nir* only manifests itself in the interrogative *imanir*. The emergence of this kind of element might result from a diachronic derivation or a peculiar occurrence of a certain morpheme which is then fused with another constituent during historical changes.

Heine et al. (1991: 58) further put forward an even more tricky situation in which an interrogative word appear to be analyzable but in fact does not comprise any semantically meaningful morphemic element. For example, most question words in English share the initial letters *wh-*. Such a phenomenon is common worldwide. Another example can be found in Swahili. This language possesses two interrogative roots, i.e., *-ni* and *-pi*. Since the semantic meaning of these two roots is unknown, they cannot be simply equivalent to a basic interrogative word in the normal sense. The Swahili interrogative paradigm is constructed based on these two stems, such as *na-ni* 'who', *ni-ni* 'what', *wa-pi* 'when' and *vi-pi* 'how'. Heine et al. (1991) apply the term SUB-MORPHEMIC ELEMENTS to label this case.

The constituents introduced above are all serving to build content interrogatives. In other words, their existence is tightly associated with the interrogative expression. In many languages, if content interrogative units are extracted directly from the natural discourse, it is highly possible that they are morphologically marked with elements with other grammatical functions, such as focus and mood markers. Especially in languages that have agglutinative or polysynthetic morphology, basic interrogative forms can be encoded with multiple information by various linguistic components, as the content question in Central Alaskan Yupik given in (3.1).

(3.1) **Central Alaskan Yupik** (Miyaoka 2012: 446)

Ca-tur-ug-cit?

what-eat-DES-INT.2SG

'What do you (SG) want to eat?'

In such a complex structure, it is difficult to tell which part should be counted as exclusively formulating the inquiry intention, or in other words, the 'real' content interrogative. Not to mention in languages with deficient grammatical description. An example is found in Cysouw (2007) in which he discovers that the interrogative meaning is further expressed by the light verbs following the question word *tsica*. Another case is pointed out by Olawsky (2006: 816) that in Urarina the interpretation of the question word *ʤa*, which can both refer to human and non-human as a subject or object argument, is subject to the context, the transitivity of the involved verb or the presence of a focus marker. Hence, it is reckless to simply set aside the rest components of the compositional interrogative unit during the analysis. Besides, we cannot exclude the possibility that a language coins a unique construction for a certain interrogative context. To err on the side of caution, it might be necessary to take note of the information provided in context and preserve its corresponding grammatical markedness alongside the basic interrogative elements during the data processing.

Recapitulated from the complicated conditions above, it is not easy to find a perfect and unified solution to deal with the issue of morphological complexity of content interrogatives. In the previous studies about content interrogatives, many linguists are inclined to avoid the polymorphemic concern and only consider the unanalyzable basic interrogative forms as the research object. Mackenzie (2009: 1133) only reckons "the simple forms as true interrogative forms" in his research, and so is Hengeveld et al. (2012).[11] Dixon (2012: 383) does not see the possibility of morphologically analyzing the interrogative forms mentioned above in Huallaga Quechua, since "there are no other instances of [them]". Heine et al. (1991: 58) take up a

[11] The decision of these both linguists about "basic question words" is deemed by Hölzl (2018: 77) as "arbitrary". Their criteria is simply that a basic question word is a monomorphemic expression, whereas a compositional form is excluded from the basic group, without considering in what way the compositional structure is or what kind of semantic meaning and function the elements have.

similar position towards sub-morphemic elements and disregard them in the investigation, as "they are not productive parts of the morphological inventory of the languages concerned".

3.4.3 Morphological analysis

As argued above, possible relevant grammatical markedness with the basic interrogative elements were also preserved during the extraction of content interrogative units. This is conducted in the sense that those components stay in a tight morphosyntactic relation, such as the affixation of constituents conveying other grammatical information in Central Alaska Yupik in (3.1). However, if the descriptive grammar explicitly points out that two components, despite their loose syntactic positions in a clause or an expression, have to co-appear simultaneously in a given context, then they will be also extracted together as a complex structure, e.g., the construction *wo […] her* 'where […] from' in German.

All collected content interrogative units were segmented into meaningful or reasonable elements according to available grammatical information. The element that is deemed to be a basic content interrogative in the grammar is taken as the core of the whole unit. Other morphologically bound segments were separated and marked with a hyphen beforehand or afterwards according to their position relative to the core. If an interrogative unit is a multi-word construction, then segments were only divided by a space. Apart from words and affixes, symbols representing stress or tone were also separated from the interrogative core. Same as multi-word expressions, no hyphen was added to these symbols.

Commonly, allomorphs can appear in different phonological contexts. In this case, one form was chosen to be the unified shape to represent all occurrences in the morphological processing, since this does not change the interrogative meaning and can maintain the consistency of relevant occurrences in a language. Yet, it should be noticed that the choice of the unified form does not imply that this form is more 'standard' than others. Instead, it is rather a random pick. And in most cases, I adopt the form that is discussed the most frequently in the grammar.

In some languages, suppletive forms can be found in the interrogative paradigm. For instance, *kuka* in Finnish is the nominative form to ask for 'who' in the singular, which is actually inflected from the stem *kene-* (Karlsson 2008: 207-208). Except for the partitive case, the element *kene-* appears in all other forms in the singular paradigm, e.g., *kenen* in the

genitive case, *kenet* in the accusative case, and *kenessä* in the inessive case. In such a situation, the stem of the irregular form was extra given in the morphological analysis in order to identify the semantic relationship within a paradigm. So, *kuka* was processed as *kene kuka*.

3.5 Online repository

All data and results of this research are available in an online repository.[12] The structure of the repository will be introduced in this section.The original Bible translations are found in the Parallel Bible Corpus.

The documents with extracted verses and the attested content interrogatives, which are resulted from the step in §3.3.2, are found in the folder 'questionwords'. All files in this folder are named with the pattern of 'check-ISO-bible/TRANSLATION'.[13] All content interrogatives attested in each context in all translations are summarized in Table 'allQuestions'. Each row of 'allQuestions' demonstrates all language-specific content interrogatives extracted from a context, while in each column all content interrogatives collected from a translation can be found.

The results of the morphological analysis, as presented in §3.4.3, are stored in the folder 'morphology'. The files are named with the pattern of 'ISO.bible/TRANSLATION'. Similar to 'allQuestion', all morphologically processed content interrogatives are included in Table 'allMorphology'. Content of these two tables was input into the automatic program for data analysis.

The scripts for automatic programs introduced in §2.2 can be found in the folder 'scripts'.

[12] URL: https://github.com/yiyo06/ContentInterrogative

[13] Abbreviations can refer to §3.1.1 and Appendix A.

4 Clusters of content interrogatives

In this chapter, the results of the clustering of interrogative contexts will be presented. Each cluster is assigned a semantic label and its meaning will be interpreted. Crucially, the name given to each cluster is only for the convenience of description. The decision for a certain label does not involve any preference for a specific semantic theory. In §4.1, I will give the basic information about the classification and the primary six resulting clusters. From §4.2 to §4.7, the outcome of the internal grouping within each primary cluster, i.e., sub-clusters, will be discussed in detail.

4.1 Information about clustering

From each Bible translation, 413 interrogative contexts are selected. According to the similarity between interrogatives used in these contexts, the clustering algorithm Partitioning Around Medoids (Kaufman & Rousseeuw 2005; see §2.2.3) assigns them into different groups. Meanwhile, the algorithm evaluates the quality of each grouping, which is presented by the value called 'average silhouette width'. The higher the average silhouette width is, the better the clustering is supposed to be. The following Figure 4.1 shows the average silhouette width respectively for the number of clustering from two to 75. The best result of the grouping is found at six groups. This indicates that 413 contexts should be optimally separated into 6 primary clusters.

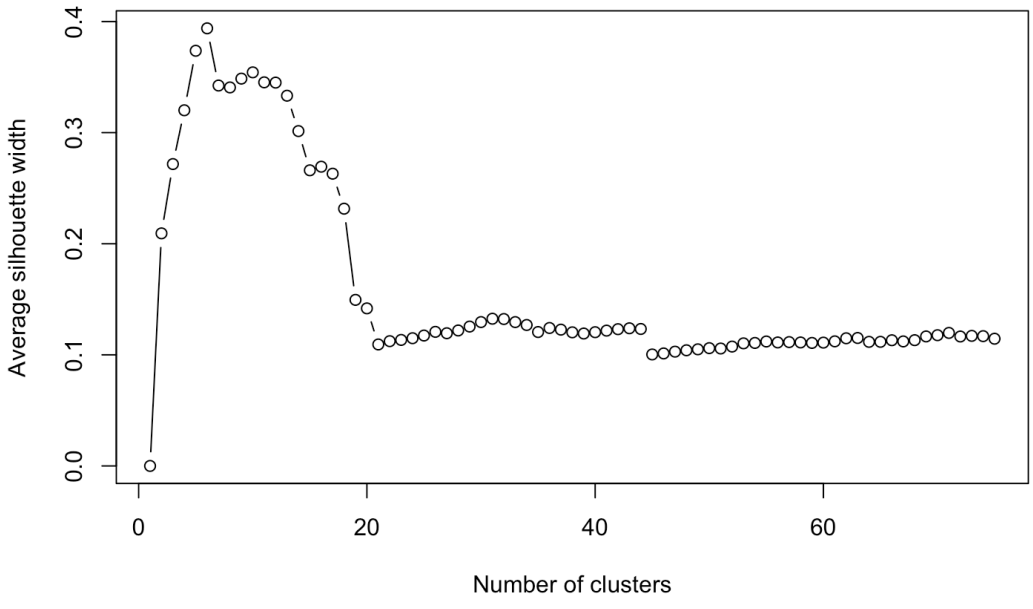

Figure 4.1: Results of clustering

Table 4.1 below provides the number of contexts and the average silhouette width of each primary cluster. The interrogative in English presented in Table 4.1 is attested in the contexts with the highest silhouette width of each group, as given in the last two columns. This value indicates that these contexts are the most representative of the corresponding cluster.

Cluster	Number of contexts	Average silhouette width	Typical context	English Interrogative
1	12	0.33	40025039	*when*
2	33	0.37	43001038b	*where*
3	98	0.48	41005030	*who*
4	125	0.28	44021033b	*what*
5	89	0.51	42006002	*why*
6	56	0.33	45010014c	*how*

Table 4.1: Information about six primary clusters

As can be seen in Table 4.1, the best average silhouette width is found at the fifth cluster, which indicates that this cluster is highly compact. In contrast, the fourth cluster has the lowest value, which indicates the most internal diversity. The fourth cluster is also the biggest one with 125 interrogative contexts, whereas the first cluster is the smallest with 12 contexts. According to the attested interrogatives and the content of contexts, I propose the following labels for these six primary clusters:

- TIME (see §4.2),
- PLACE (see§4.3)
- PERSON (see §4.4)
- THING (see §4.5)
- INTENTION (see §4.6)
- MANNER/EXTENT (see §4.7)

In order to further detect the internal structure of each cluster, a second level of grouping is conducted. As a result, a different number of sub-clusters are identified within each primary group. They will be elaborated in the corresponding sections.

4.2 Cluster of TIME

The present chapter will discuss the first primary cluster of the TIME category. Firstly, some general information about this cluster will be given in §4.2.1. Then, the following sections will elaborate three identified sub-clusters within this group with examples from sampled languages. They are arranged as follows:

- §4.2.2 — TIME.GENERAL
- §4.2.3 — DURATION.FUTURE
- §4.2.4 — DURATION.PAST

4.2.1 Overview

Compared to the other five primary clusters, the first cluster is considerably smaller with only twelve interrogative contexts. The average silhouette width of this cluster is 0.33, as shown in Figure 4.1 in §4.1. Table 4.2 below gives the IDs of verses belonging to this cluster and the silhouette width of each interrogative context in the second and third column. Next to the quality of clustering, interrogatives in English, German, and Mandarin extracted from translations *eng-x-bible-common*, *deu-x-bible-newworld*, and *zho-x-bible-contemp* are shown in the next three columns. Considering that almost all content interrogatives used in German, English, and Mandarin pertain to temporal concepts, I name this cluster **TIME**.

Figure 4.2 below shows the distribution of interrogative contexts within this cluster with symbols indicating encodings in English. As plotted above, it can be clearly seen that the internal structure of the cluster TIME comprises three groupings of data points. The first group is found at the top-left in the graph. Data points in this group represent the contexts mostly asked with the question word *when* in English. Four contexts constitute the second group located at the top-right. They are all encoded with the interrogative construction *how long* in English. The third group consists of only one context at the bottom in Figure 4.2. This context is also marked with the interrogative *how long* in English.

Such a clear distribution is also quantitively confirmed by the second level of clustering, as shown in Figure 4.3. The optimal result implies three sub-clusters within this cluster. The following sections will discuss these sub-clusters and their respective typical interrogative contexts in detail.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	40025039	0.47769	*when*	*wann*	什麼時候
2	40025038	0.46305	*when*	*wann*	NA
3	40025037	0.45841	*when*	*wann*	什麼時候
4	42021007a	0.42909	*when*	*wann*	什麼時候
5	43006025	0.41942	*when*	*wann*	什麼時候
6	42017020	0.41546	*when*	*wann*	什麼時候
7	66006010a	0.29175	*how long*	*bis wann*	多久
8	41009019b	0.24218	*how long*	*wie lange*	多久
9	42009041a	0.24004	*how long*	*wie lange*	多久
10	43010024	0.23643	*how long*	*wie lange*	什麼時候
11	45004010	0.17150	*how*	*welchen*	怎麼
12	41009021	0.17037	*how long*	*wie lange*	多久

Table 4.2: Verse selection of cluster TIME

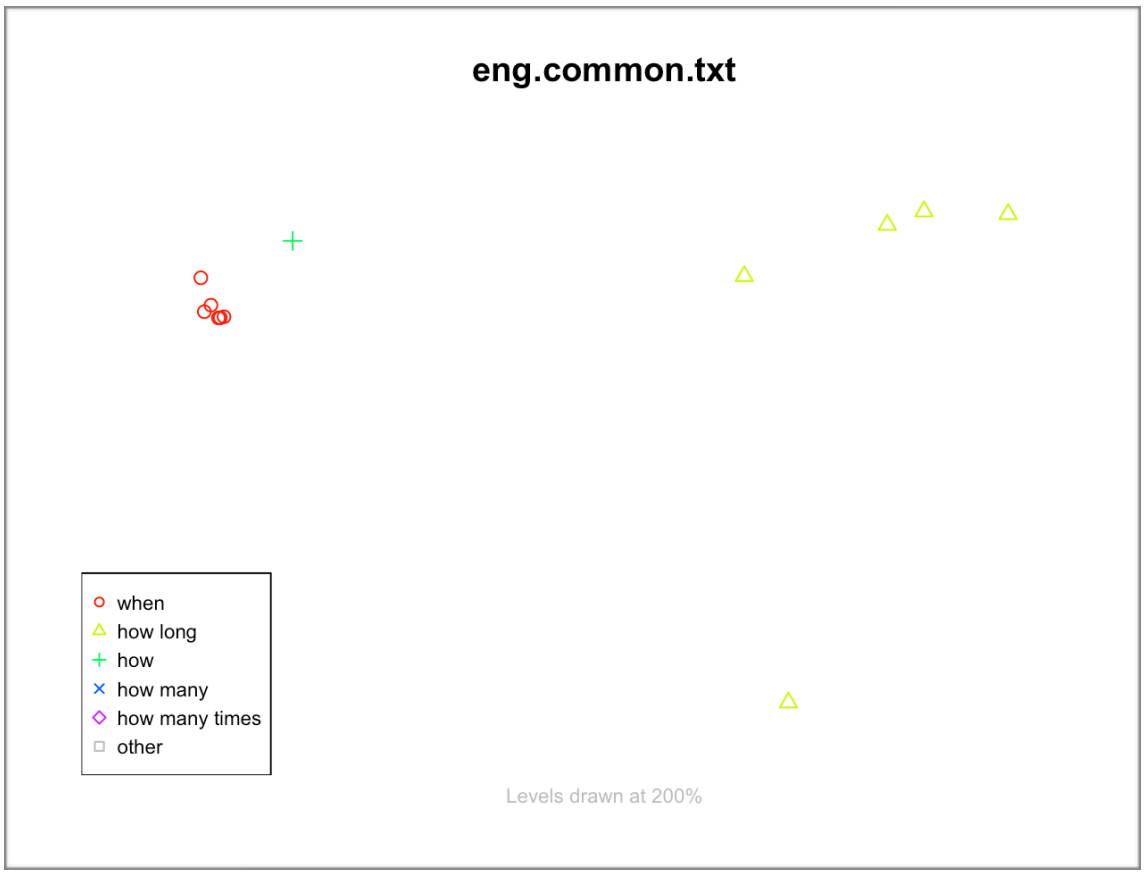

Figure 4.2: MDS plot of cluster TIME

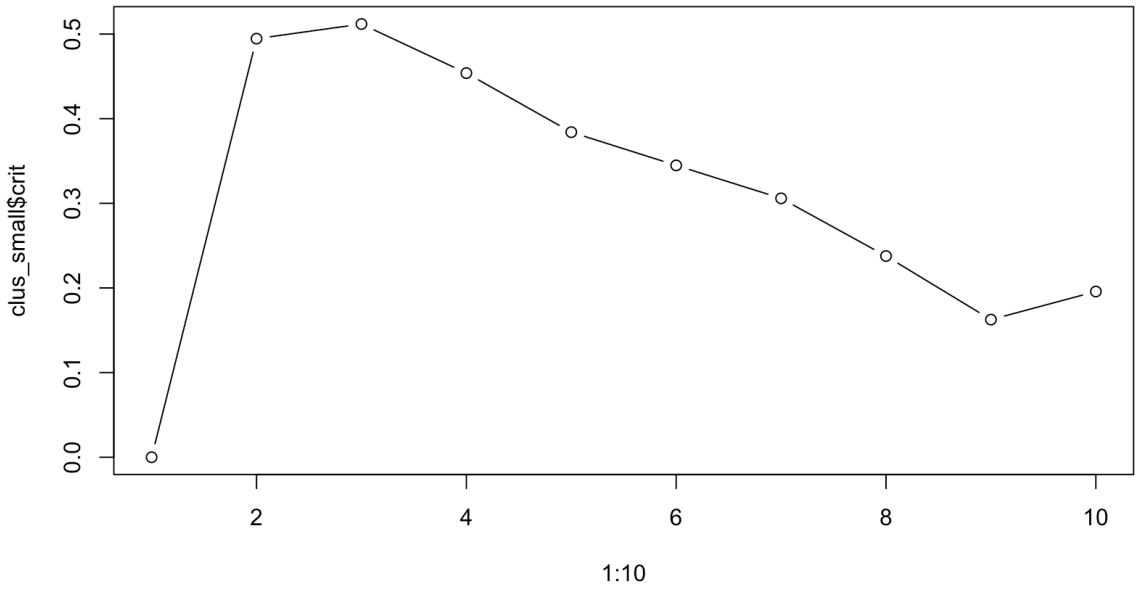

Figure 4.3: Suggested sub-clusters of TIME

4.2.2 TIME.GENERAL

According to the internal grouping, six contexts are assigned to the first sub-cluster.[14] Since most of them are encoded with the general temporal interrogative *when* in English, the label **TIME.GENERAL** is given to this sub-cluster. Let us see the first four verses with the highest silhouette width in (4.1) below.

(4.1) *eng-x-bible-common*

40025039	***When** did we see you sick or in prison and visit you?*
40025038	***When** did we see you as a stranger and welcome you, or naked and give you clothes to wear?*
40025037	*Lord, **when** did we see you hungry and feed you, or thirsty and give you a drink?*
42021007a	*They asked him, "Teacher , **when** will these things happen?"*

[14] Even though the context in 45004010 is also assigned to this sub-cluster, many languages do not choose the temporal interrogative to ask this question but apply the interrogative from the category MANNER. Thus, this context is manually excluded.

The first three verses are consecutive in the text and the corresponding questions refer to similar content. In these contexts, someone wants to know the time of an event that happened in the past. As widely known, question words in English do not formally distinguish tense when they are used to ask for the general time. German and Mandarin show the same grammatical trait as English in terms of temporal content interrogatives.

However, some languages in the sample have distinct temporal interrogative forms for different tenses. For instance, according to van der Berg & Busenitz (2012: 204), there does not exist a general form to inquire about temporal information in Balantak, an Austronesian language spoken in the island Sulawesi, which is opposed to English. Instead, this language has two question words *ipi* and *maripi*, which respectively refer to a future or a past event. For these three contexts, Balantak adopts *maripi*, which underlines the temporal aspect PAST.

This semantic difference is not just found in Balantak. Some other sampling languages also utilize a specific interrogative form for a past event. Gwich'in, an Eyak-Athabaskan language spoken in Canada and the United States, applies *nijin dąį'* in these contexts. This form is exclusively indicative of the past tense (Peter 1979: 143). Central Yupik from the Eskimo-Aleut family also distinguishes the past tense by using the form *qangvaq-* in these questions (Miyaoka 2012: 300). Therefore, thanks to the grammatical evidence provided in these languages, it can be considered that contexts extracted from 40025039, 40025038, and 40025037 refer particularly to questions asking for the time of a past event. Correspondingly, they represent a temporal subdomain called **TIME.GENERAL.PAST**.

In contrast, the context in 42021007a presents a situation in which the event time of happenings in the future is inquired about. Whether it refers to the exact clock time of a day or a general date is not told in the text. Among all grammars used for this study, I did not find any information about an interrogative form exclusively for the clock time in the future or the past. It appears that languages neglect the tense of a clock time and tend to use a common coding for this temporal notion. Nevertheless, given that the clock time is a relatively modern concept, it is nearly impossible that it would come up in the Bible.

If we only observe the usage of content interrogatives in English, German and Mandarin in 42021007a, as provided in Table 4.2 in §4.2.1, there is no particular grammatical form or markedness indicating the future tense. The reason is the same as already illustrated above — these languages do not formally distinguish the tense in the interrogative. The exact temporal

location of an event can be interpreted through the inflection of verbal predicates, the usage of temporal adverbs or the contextual information.

However, there are also languages in the sample that apply particular interrogative forms for TIME questions in the future tense. Some of them are given in Table 4.3 below. Compare the content interrogatives used in 42021007a to indicate FUTURE and 40025039 for PAST extracted from the following five languages.

	FUTURE	**PAST**
Parecís	*xoana xowakaite*	*xoana xowaka*
Batak Karo	*ndigan*	*ndiganai*
Balantak	*ipi*	*maripi*
Gwich'in	*nijin ji'*	*nijin dąį'*
Central Yupik	*qaku*	*qangvaq*

Table 4.3: Interrogatives for FUTURE and PAST

The first row in Table 4.3 presents the interrogatives in Parecís, an Arawakan language spoken in Brazil. The formal difference lies in that, based on the construction marking PAST, the enclitic *ite* is added to indicate FUTURE. The function of this enclitic is to mark the future tense (Brandão 2014: 79). An opposite situation is found in Batak Karo, an Austronesian language spoken in Indonesia, i.e., the past tense is formally marked. According to Woollams (1996: 225), *ndigan* used in 42021007a is a general question word for 'when', whereas the morphologically more complicated form *ndiganai* in 40025039 is used exclusively to ask for time in the past. Similarly, Balantak seems to construct the question word *maripi* indicating the past tense based on *ipi* querying the time in the future (van der Berg & Busenitz 2012: 204).

Gwich'in has a general question word *nijin* 'when/where'. On this basis, the future and past tense are explicitly expressed by combining *nijin* with the independent word *ji'* or *dąį'*, respectively (Peter 1979: 143). In order to distinguish questions asking for future and past events, Central Yupik employs two separate forms *qaku* and *qangvaq*. These forms share structural similarity and are very possibly derivationally related (Miyaoka 2012: 300).

Except Parecís, occurrences presented in Table 4.3 are also attested in 42017020. Given these content interrogatives are particularly applied to the future tense in these languages, the contexts of 42021007a and 42017020 can be considered representative for the inquiry about the time of an event in the future. The subdomain that this context pertains to is thus dubbed **TIME.GENERAL.FUTURE**.

Due to the similar content of the first three typical contexts, we shall take the fifth verse from the list to enrich the content of representatives, as given in (4.2) below.

(4.2) *eng-x-bible-common*

43006025 *When they found him on the other side of the lake, they asked him, " Rabbi , **when** did you get here?"*

Same as the first three selected verses, the question in (4.2) aims at the time of an event in the past. When observing interrogatives used in English, German and Mandarin, as shown in the fifth rows in Table 4.2 above, they are also identical to those for TIME.GENERAL.PAST. However, formal differences can be seen in some other sampled languages. For example, Cantonese, although related with Mandarin, uses *gei²si⁴* 幾時 'what time' (lit. *how many time*) in 43006025, whereas *sam⁶mo¹si⁴hau⁶* 甚麼時候 'when' (lit. *what ime*) is found in TIME.GENERAL.PAST. Table 4.4 below lists the other four languages demonstrating the same case along with a comparison with TIME.GENERAL.PAST (40025038).

	'what time'	**TIME.GENERAL.PAST**
Balinese	*kalinapi*	*dipidanke*
Chuj	*janic'*	*b'a'n̈i*
Parauk	*yam mawx*	*lai mawx*
Huallaga Huánuco Quechua	*imay örataj*	*imaytaj*

Table 4.4: Interrogatives indicating 'what time'

In Balinese from the Austronesian family, the form *kalinapi* attested in 43006025 is composed of *kali* 'time' and *napi* 'what', while *pidan* is a basic question word to ask for the

time in the past (Shadeg 2014: 55, 90, 281). In Chuj, a Mayan language spoken in Guatemala and Mexico, *jan-* is an interrogative root meaning ‘how many’. On this basis, *janic’* is built to query ‘what time’ (Hopkins 2012: 96). The way in Chuj to construct the interrogative for ‘what time’ is interestingly identical to Cantonese, as just mentioned above, i.e., taking the interrogative form for QUANTITY as the root of the structure ‘what time’.

Maxw in Parauk, an Austronesian language also named Wa and spoken in Myanmar and China, is normally employed to ask ‘who’, while it can also be a component of other interrogative constructions, including those for temporal questions. No difference is explained between *lai maxw* and *yam mawx* in the grammar written by Ma (2012: 46, 104-105) in English, according to which both forms refer to ‘when’. However, in the grammar written in Mandarin (Zhao & Chen 2005: 52), *lai mawx* is translated as *jǐshí* 几时, which is equal to *gei²si⁴* 幾時 in Cantonese but only written in simplified characters, whereas *yam mawx* means ‘when’ and corresponds to *sam⁶mo¹si⁴hau⁶* 甚麼時候 in Cantonese. That is, compared to Cantonese, Parauk displays a reverse interrogative usage in 43006025 and 40025038.[15]

Another kind of interrogative formation is found in the Quechuan language Huallaga Huánuco Quechua in which *imay* is a general question word for ‘when’. For the context in 43006025, this language adopts a combination *imay örataj*. In this construction, the appearance of *öra* ‘time’ seems to emphasize a more exact time (Weber 1989: 329).

However, this context cannot be simply considered the question for ‘what o’clock’, since many languages have another unique form to ask for this kind of information. Besides, as above mentioned, it is unlikely that a clock time is queried in the Bible. Because no corresponding answer is given in the following verses, the exact temporal information that is asked in 43006025 cannot be confirmed.

Yet, the special forms used in some languages just discussed could be a sign of a subtle semantic subdomain. Structurally, *kalinapi* in Balinese, *imay örataj* in Huallaga Huánuco Quechua and *gei²si⁴* 幾時 in Cantonese are both composed of a basic interrogative form and a noun meaning ‘time’. From a syntactic perspective, the interrogative base is attributively

[15] However, this could also be an issue of inequivalent translation. Basically, there is no such a usage of *jǐshí* 几时 in standard Mandarin. The interrogative *shénmeshíhòu* 什么时候 to ask for ‘when’ in Mandarin is a compositional construction comprising ‘what’ and ‘time’, which cannot be simply equated to the question *what time* in English. The translators in Mandarin might intend to describe the fine semantic difference between *lai maxw* and *yam maxw*. Nevertheless, the corresponding translation may be inaccurate and confusing.

combined with a noun and thus serves as a limiter. This function is similar to *which*, *what kinds of* or *how many* in English. Such a kind of constructions reflects an intention to define or limit the noun and expresses the meaning of selection. In this sense, the semantic subdomain presented in 43006025 might refer to a limited span and thus could be regarded as **TIME.SELECTION**.

4.2.3 DURATION.FUTURE

According to the result of the second level clustering, the four verses in row 7 to 11 in Table 4.2 above are assigned to this sub-cluster. They are given in (4.3) below.

(4.3) *eng-x-bible-common*

66006010a	*Holy and true Master,* ***how long*** *will you wait before you pass judgment?*
41009019b	***How long*** *will I put up with you?*
42009041a	*You faithless and crooked generation,* ***how long*** *will I be with you?*
43010024	***How long*** *will you test our patience?*

As shown in (4.3), these contexts present situations in which the duration of an action or a state is inquired about. In the English translation, the future tense in all three examples entails that the actions or states involved do not occur synchronically with the interrogative utterance. Instead, while the questions are addressed, the actions do not happen yet. It is also possible that the actions have begun and will endure after the interrogative utterance. The duration between the interrogative utterance and the end of the actions is queried. Thus, the semantic property of this sub-cluster refers specifically to the whole temporal length of an action or a state that will last to or occur in the future. In this sense, questions of this cluster can also be understood as asking for 'until when' the action or the state will remain. To tell apart this semantic trait from those discussed in §4.2.2, this sub-cluster is labeled **DURATION.FUTURE**.

Compared to other basic interrogative categories, the way to ask for the duration in the future is only sparingly described in grammars. For some languages, the corresponding interrogative forms are even completely absent in grammatical references. This complicates

the identification of the complete interrogative construction for DURATION.FUTURE in the raw text. However, it is still noticeable that many languages tend to utilize distinct interrogatives for the period in the future and general temporal information. Various constructional patterns can be observed among sampled languages and they will be summarized in the following.

Based on the attested forms in this sub-cluster, there are two main morphological types of interrogatives for DURATION.FUTURE. Firstly, in some languages there exists a special interrogative that serves particularly for DURATION.FUTURE. Such a form is morphologically independent of any other basic interrogatives and is sometimes related to other grammatical categories. For instance, in 41009010b Croatian applies the word *dokle* which is formally dissimilar from interrogatives for other temporal concepts, e.g., *kada* 'when' attested in TIME.GENERAL (Browne & Alt 2004: 36). *Dokle* denotes 'until when' or 'how long' in a question, while it can also serve as the temporal conjunction 'until' to lead a subordinating clause.

Yet, this kind of discrete interrogative is only scarcely attested. Predominantly, the interrogative for DURATION.FUTURE is morphologically compositional. In terms of the meaning of the components, several constructional patterns are identified with high frequency among sampled languages.

The first common kind of composition for DURATION.FUTURE is related to the general temporal interrogative 'when'. As described before, the semantic property of DURATION.FUTURE can also be interpreted as the span of an action or a state lasting until a certain moment. This way to query the temporal duration is attested in many languages in the world. The corresponding interrogative is then normally composed of elements meaning 'until' and 'when'.Table 4.5 below gives five examples showing the pattern '**until when**' collected from 42009041a. The constituent indicating 'until' are marked in bold. As a comparison, interrogatives in the corresponding languages for TIME.GENERAL.PAST from 40025038 are also provided (Durie 1985: 165, 259; Wheeler & Yates 1999: 263, 489; Bolles & Bolles 2019: 20, 42; Estigarribia 2020: 111).

The second typical construction is a combination with the category QUANTITY. Six examples are displayed in the following Table 4.6. Interrogative constructions used in 42009041a are presented in the first column, while the second column shows how the corresponding languages ask for information about QUANTITY.

	‘until when’	‘when’
Acehnese	***sampoe*** *án pajan*	*pajan*
Catalan	***fins*** *quan*	*quan*
Korean	언제까지	언제
Yucatec Maya	***tac*** *ba'ax kiin*	*ba'ax kiin*
Paraguayan Guaraní	*araka'e* ***peve*** *piko*	*araka'e piko*

Table 4.5: Examples for ‘until when’-construction

	how many/much time’	‘how much/many’
Mandarin	多久	多少
Indonesian	*berapa lama*	*berapa*
Chamorro	*kuåntos tiempo*	*kuånto*
Toro So Dogon	*waaru yagɔ baa*	*yagɔ baa*
Welsh	*am faint*	*faint*
Turkish	*ne kadar*	*ne kadar*

Table 4.6: Examples for ‘how much time’-construction

To inquiry about the temporal span of an action or a state in the future, speakers of languages belonging to this type appear to count the amount of time ahead and thus erect the corresponding formulation based on the interrogative meaning ‘how many/much’. The rest in the construction varies across languages. For example, as can be seen in Table 4.6, Mandarin combines *duō* 多 ‘how many/much’ with *jiǔ* 久 ‘long’, which is the same as the structure in Indonesian (Sneddon 1996: 222). Instead, Chamorro from the Austronesian family chooses *tiempo* ‘time’ cooperating with *kuåntos* ‘how many’ (Chung 2020: 495). This kind of formation is also found in Toro So Dogon, a Niger-Congo language spoken in Mali and Burkina Faso (Heath 2017: 284). Welsh applies a different pattern in which the preposition *am* ‘for’ occurs with the question word *faint* for QUANTITY together (King 2003: 270). Besides, it is also possible that a language encodes the meaning ‘how much’ and the temporal duration with an identical form, as Turkish (tur) presents in the last row in Table 4.6 (Göksel &

Kerslake 2005: 264). Here, all these examples are categorized into the pattern '**how much time**' for the sake of parallelism.

The third frequently attested construction for DURATION.FUTURE is composed of the basic interrogative meaning 'how' and an element meaning 'length' or 'long'. This structure is especially recurrently adopted by languages of the Indo-European family, e.g., English and German. Table 4.7 below gives four other examples of the pattern '**how long**'.

	'how long'	**'how'**
Dutch	*hoelang*	*hoe*
Icelandic	*hversu lengi*	*hversu*
Finnish	*kuinka kauan*	*kuinka*
Czech	*jak dlouho*	*jak*

Table 4.7: Examples for 'how long'-construction

Aside from these three patterns, there are also other lesser-frequent strategies to establish interrogative constructions for the temporal length in the future. For instance, in Hungarian the form *meddig* is lexicalized from *mi* 'what' and now refers exclusively to 'how long' (Előry 2007: 1137). Car Nicobarese, an Austro-Asiatic language spoken in India, and Irish have a set of interrogative roots based on which forms for different categories are generated. In Car Nicobarese, the basic root *i-* is combined with the morpheme *rô-* indicating 'length' to shape the construction *i rôòten* extracted from 42009041a (Braine 1970: 204, 207; Whitehead 1925: xliii). Similarly, the form *cád fad* in Irish collected from the same context incorporates the interrogative root *cád* and *fad* meaning 'length' (Stenson 2008b: 13).

4.2.4 DURATION.PAST

In the second level of clustering, the context in 41009021 is conspicuously separated from the other contexts of TIME, as the corresponding data point is placed solely at the bottom in Figure 4.2 above. The content of this verse is illustrated in (4.4).

(4.4) *eng-x-bible-common*

41009021 *Jesus asked his father, "**How long** has this been going on?" He said , "Since he was a child."*

Similar to DURATION.FUTURE, the context in 41009021 also refers to a question with the intention to obtain the duration of a state. The interrogative coding used in English for 41009021 is *how long*, which is the same as those for DURATION.FUTURE. However, the tense of this context discloses the semantic uniqueness. Different from the future tense in DURATION.FUTURE, the present perfect continuous tense occurs in the utterance in 41009021. This denotes that the inquired state has started in the past, already before the interrogative utterance occurs, and it still remains. Therefore, the expected answer to this question is actually the period in which the state has hitherto lasted. In order to differentiate the semantic property of this sub-cluster from DURATION.FUTURE, the name **DURATION.PAST** is employed. To sum up, we can understand the different intentions of DURATION.FUTURE and DURATION.PAST as that the former is indicative of the meaning 'until when', as analyzed in the last section, whereas the latter wants to ask 'since when' an action or state has already been in this condition.

For the bygone length of an action or a state in the past, most sampled languages tend to utilize the same interrogative construction as those for DURATION.FUTURE. For example, as seen in Table 4.2 above, none of English, German, and Mandarin apply any special morphological markedness in 41009021. Nevertheless, this semantic differentiation still receives attention in some languages and is then formally distinguished. This is realized through several morphological structures. For instance, particular markers, e.g., suffixes indicating the past tense, can be added to the interrogative for DURATION.PAST. Besides, as illustrated in the last section, the construction for DURATION.FUTURE is commonly derived from the basic question word 'when' across languages. Now, this kind of interrogative formation also serves quite often for the period in the past. In this case, constructions of DURATION.FUTURE and DURATION.PAST are normally composed of elements meaning 'until when' and 'since when', respectively.

However, it is possible that languages develop a unique form for DURATION.PAST. Table 4.8 below gives interrogatives specifically for DURATION.PAST in six sampled languages. To serve as a comparison, forms for DURATION.FUTURE extracted from 66006010a in corresponding languages are also provided.

As can be seen in Table 4.8, Garifuna from the Arawakan family and Spanish adopt dissimilar forms for these two semantic subdomains. In Garifuna, the construction *átiri dan*, literally translated as 'how much time', is used to for the temporal span in the past. In 66006010a, *darísan ídame* is applied for DURATION.FUTURE. In this composition, *darís* means 'until' followed by the question marker *=san*, while *ída* serves as the basic question word 'how' combined with the tense aspect enclitic *=me* for distant future, which jointly means 'when' (Haurholm-Larsen 2016: 175-176; Gonzaléz 2012: 52).

	DURATION.PAST	**DURATION.FUTURE**
Garifuna	*átiri dan*	*darísan ídame*
Spanish	*cuánto tiempo*	*hasta cuándo*
Maasina Fulfulde	*gila mande*	*faa mande*
Japanese	いつごろから	いつまで
Croatian	*otkad*	*kad*
North Alaskan Inupiatun	*qaŋaaglaan*	*qanutun*

Table 4.8: Comparison between DURATION.PAST and DURATION.FUTURE

A similar compositional pattern can be found in Spanish. For DURATION.PAST, this language utilizes the combination *cuánto tiempo.* The literal meaning this structure is 'how much/many time'. In contrast, the expression *hasta cuándo* for DURATION.FUTURE is composed of 'until' and 'when'.

The next three languages in Table 4.8 exhibit the second morphological possibility mentioned previously, i.e., interrogatives for DURATION.PAST and DURATION.FUTURE share or are derived from a common component which is normally indicative of 'when'. In Maasina Fulfulde, a Niger-Congo language spoken in Mali, the question word *mande* 'when' is respectively combined with *gila* 'since' and *faa* 'until' for the duration in the past and future (Breeedveld 1995: 252, 486). Japanese applies the same kind of interrogative construction.

The basic question word *itsu* いつ 'when' is combined with *made* まで 'until' to inquire about DURATION.FUTURE, while it comes up with *kara* から 'since' for DURATION.PAST (Makino & Tsutsui 2008: 25).[16] In Croatian, there exists a form *otkad* specifically for the question 'from when'. This word is discernibly related to the general question word *kad* 'when' (Leskien 1976: 405).

Finally, North Alaskan Inupiatun, an Eskimo-Aleut language spoken in Alaska, also formally tells apart these two temporal subdomains. To ask for the duration in the past, this morphosyntactically highly complex language uses the form *qaŋaaglaan*, which, according to Seiler (2012: 164), refers particularly to 'how long ago' or a temporal question in the past tense. Conversely, *qanutun* appearing in DURATION.FUTURE is a general expression meaning 'for how long' or 'how much' (Seiler 2012: 164).

4.2.5 Summary

This chapter depicted the cluster of TIME. The internal structure of this cluster is clearly identifiable. Three main subgroups are found. The first sub-cluster TIME.GENERAL discussed in §4.2.2 contains contexts in which questions for a common temporal concept are addressed. By observing the particular interrogative codings used in sampled languages, a more subtle semantic differentiation is manually drawn within this sub-cluster, while three smaller sets are further established, i.e., TIME.GENERAL.PAST for an event happening in the past, TIME.GENERAL.FUTURE for an event in the future and TIME.SELECTION indicating a limited period. Only a very few sampled languages mark these three semantic features with specific interrogative codings. These codings are usually morphologically associated, i.e., sharing the same derivational stem or element. A situation in which languages have different and morphologically irrelevant structures to encode these three temporal concepts is not attested in this sample. Most sampled languages have just one general form to enquiry about temporal information without signaling the tense. Therefore, the algorithm is unable to automatically classify the corresponding contexts into individual sub-clusters.

---

[16] Between いつ 'when' and から 'since' stands the component ごろ meaning 'about', whose appearance is not obligatory for obtaining information about the duration in the past. The translator applied ごろ here probably with the intention to formulate the inquiry more precisely, or to soften the tone in order to express politeness.

The sub-cluster DURATION.FUTURE described in §4.2.3 is about the temporal length of an event that is still going on or will remain, which can be understood as ‘until when’ in English. In contrast, §4.2.4 illustrated the sub-cluster DURATION.PAST with a context acquiring the already passing-by time, which is equated to ‘since when’ in English. In terms of structural formation, despite the unique forms specifically for these two meanings in some languages, the interrogative codings attested in these two sub-clusters are compositional and derived from other interrogative categories. The main sources are TIME.GENERAL, as ‘until when’ and ‘since when’ in English, QUANTITY, i.e., ‘how many/much time’, and MANNER, i.e., ‘how long’.

4.3 Cluster of PLACE

In this chapter, the cluster composed of contexts related to locational concepts will be elaborated. Firstly, a general introduction of this cluster will be given in §4.3.1. In terms of spatial interrogatives, Stolz et al. (2017) provide a paradigm with three main aspects, i.e., PLACE, GOAL and SOURCE. On this basis, more subtle distinctions are observed among the sub-clusters of this group. These sub-clusters will be presented in the following sections:

- §4.3.2 — PLACE.EVENT
- §4.3.3 — PLACE.OBJECT.SG
- §4.3.4 — PLACE.OBJECT.PL
- §4.3.5 — PLACE.GOAL
- §4.3.6 — PLACE.FROM.ORIGIN
- §4.3.7 — PLACE.FROM.SOURCE

Finally, the way that languages encode questions of two subgroups related to the direction ‘from’, i.e., ORIGIN and SOURCE, will be summarized in §4.3.8.

4.3.1 Overview

This cluster comprises 33 interrogative contexts representing the semantic domain **PLACE** in total. As can be seen in Table 4.1 in §4.1, the average silhouette width of this cluster, which

indicates the clustering quality, lies at 0.37. In order to provide an approximate impression of this cluster, the following Table 4.9 presents a selection of contexts allocated to this cluster. Along with the respective silhouette width of each context, the IDs of corresponding verses and interrogatives used in translations in English, German, and Mandarin are also given.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	43001038b	0.49838	*where*	*wo*	哪裡
2	43011034	0.49079	*where*	*wo*	哪裡
3	43009012	0.48571	*where*	*wo*	哪裡
4	40002002	0.48393	*where*	*wo*	哪裡
5	42022009	0.47967	*where*	*wo*	哪裡
6	42017037	0.46958	*where*	*wo*	哪裡
7	40013054a	0.39803	*where*	*woher*	哪裡
8	41006002a	0.38674	*where*	*woher*	哪裡
9	40015033	0.38424	*where*	*woher*	哪裡
10	43013036	0.38403	*where*	*wohin*	哪裡
11	43016005	0.37612	*where*	*wohin*	哪裡
12	44007049b	0.24564	*where*	*welches*	哪裡

Table 4.9: Verse selection of cluster PLACE

Based on the (dis-)similarities between these interrogative contexts, MDS plots the internal structure of this cluster, as can be seen in Figure 4.4 below. In this Figure, symbols reflect the use of interrogatives in English for cluster PLACE. Most contexts are encoded with the question word *where* in the translation *eng-x-bible-common*. Yet, despite the same interrogative in English, the distribution of data points indicates that there exist semantic differences across these contexts, which leads to diverse interrogative forms across other sampled languages. Roughly observed, there are three areas with points diffusing in the space, i.e., at the top-left, bottom-middle, and right. Within points gathering at the bottom-middle, we can further identify small groupings.

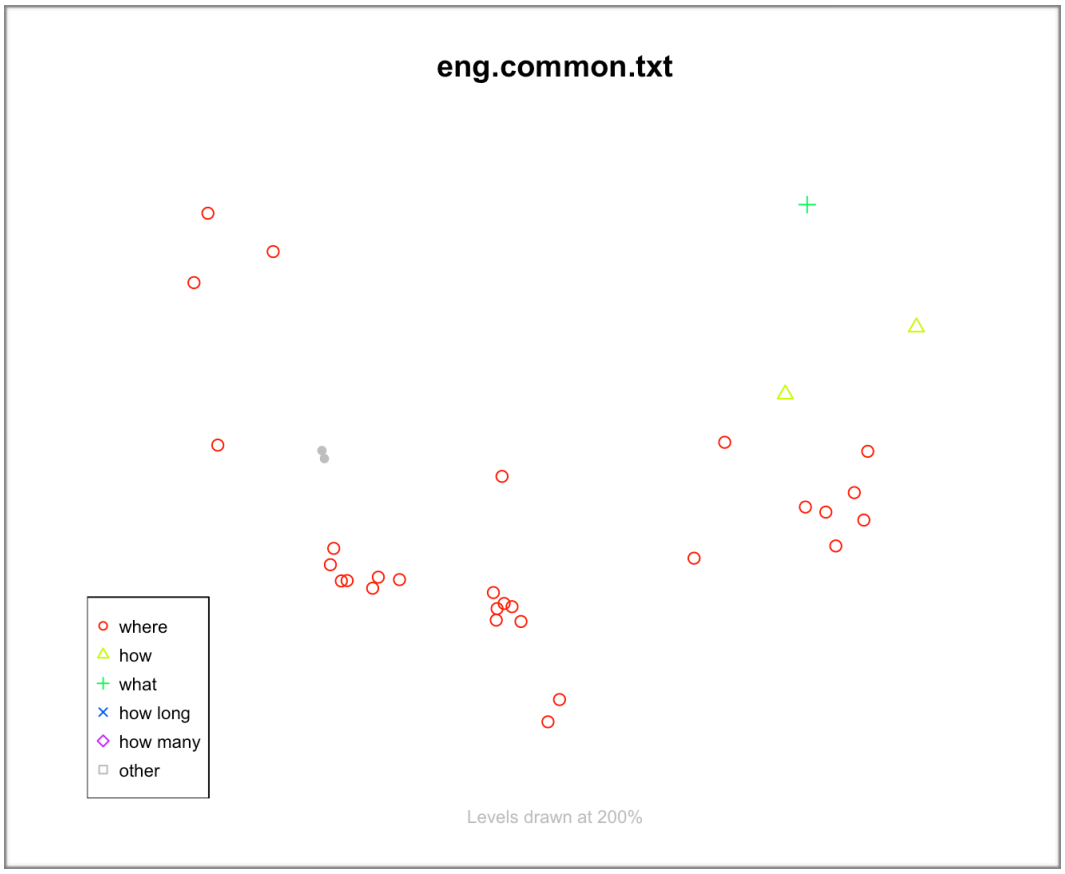

Figure 4.4: MDS plot of cluster PLACE (English)

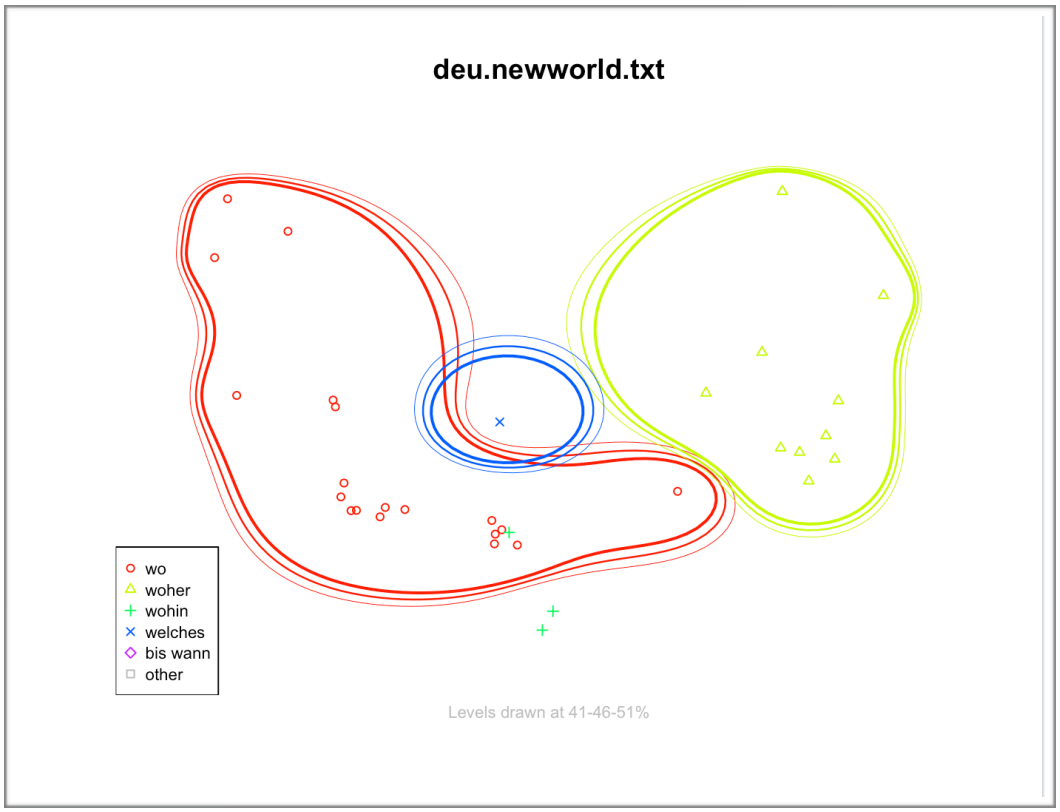

Figure 4.5: MDS plot of cluster PLACE (German)

If we take question words in German as the reference, some linguistic hints can already be drawn, as shown in Figure 4.5 above. Symbols and lines in Figure 4.5 represent interrogatives applied in the German translation *deu-x-bible-newworld*. Data points at the right in the graph are all green triangles referring to the question word *woher* 'where from', while read points indicating *wo* 'where' assemble at the left side. The finer groupings at the bottom are still unrecognizable. Nevertheless, it is clear to tell that there are three contexts marked by a green cross. In these contexts, the question word *wohin* 'where to' is applied. In this sense, we can expect that the exact locational meaning of contexts of PLACE can be recognized via analyzing interrogatives used in different languages.

For a statistically more reliable internal classification, a second level of clustering within this cluster is conducted. The result is shown in Figure 4.6 below. As can be read in this figure, the optimal result of the second level clustering lies in 18. That is, the program suggests that 33 contexts can be further divided into 18 groups. It is expected that more subtle semantic subdomains that are encoded with spatial interrogatives can be identified based on this automatic classification. The following §4.3.2 to §4.3.7 will elaborate the interpretable sub-clusters.

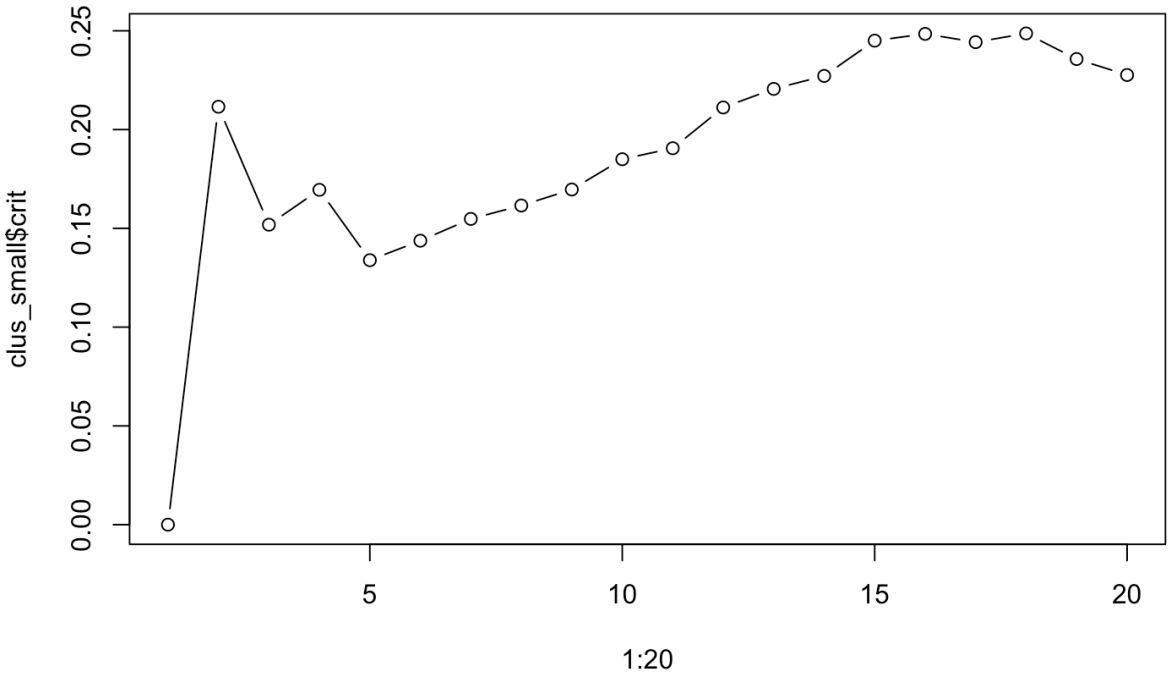

Figure 4.6: Suggested sub-clusters of PLACE

4.3.2 PLACE.EVENT

The first sub-cluster to be elaborated in this section incorporates five contexts, as given in (4.5). Except for the incomplete content in 42017037, the intentions of the other four questions are all about obtaining the stative location of an event or a happening. In these contexts, predicates are verbs describing an event that occur in an unchanged location. The question word *where* functions adverbially in the clause and supplements the spatial detail of the event. Given these traits, this sub-cluster is named **PLACE.EVENT**.

(4.5) *eng-x-bible-common*

43001038b	*"Rabbi (which is translated Teacher), **where** are you staying?"*
43011034	*He asked, "**Where** have you laid him?"*
42022009	*They said to him, "**Where** do you want us to prepare it?"*
40002004	*He gathered all the chief priests and the legal experts and asked them **where** the Christ was to be born.*
42017037	*The disciples asked, "**Where**, Lord?"*

	Chuj	Chamorro	Dogrib	Gwich'in	Parecís	Tenharim-Parintintin-Diahoi
43001038b	*i'ajm til*	*månu nai*	*edı̨ı̨*	*nijin*	*aliyako*	*mome*
43011034	*'ajtil*	*månu nai*	*edı̨ı̨*	*nijin*	*aliyo*	*mome*
42022009	*'ajtil*	*månu*	*edı̨ı̨*	*nijin*	*alyako*	*mome*
43008019	*'ajtil*	*månu nai*	*edı̨ı̨*	*nijin*	*aliyekoa*	*mome*
42017037	*'ajtil*	*amånu*	*edı̨ı̨*	*nijin*	*alyako*	*mome*

Table 4.10: Interrogatives of PLACE.EVENT

Remarkably, 43001038b and 43011034 have the highest silhouette width of the cluster PLACE. This indicates that these two questions are representative of the complete cluster. For PLACE.EVENT, most sampled languages choose the basic interrogative form for PLACE provided in grammar. To give a glimpse of the interrogative usage in PLACE.EVENT and, more

importantly, to serve as a comparison for the sub-clusters to be discussed in the next sections, Table 4.10 above lists the question constructions in six languages with accessible grammatical descriptions (Hopkins 2012: 1, 256; Chung 2020: 492; Saxon & Siemens 1996: 25; Alexander & Alexander 2011: 22; Brandão 2014: 331; Betts 2012: 12).

4.3.3 PLACE.OBJECT.SG

As a result of the second level of clustering, four contexts are assigned to the second sub-cluster. The corresponding verses are given in (4.6) below.

(4.6) *eng-x-bible-common*

43009012 *They asked, "**Where** is this man?"*
40002002 *They asked, "**Where** is the newborn king of the Jews?"*
43007011 *They kept asking, "**Where** is he?"*
43008019 *They asked him, "**Where** is your Father?"*

The scenarios of contexts in this sub-cluster are about asking for the specific location of a person. In the English translation, the verses shown above all contain a predicative clause in which the third-person singular copula *is* links the question word *where* to the subject. Different from verbs denoting concrete actions or dynamic motions, the use of a copulative predicate underlines a stative and unchanged existence of an animate subject. For this reason, I name this cluster **PLACE.OBJECT.SG**.

For this sub-cluster, English, German, and Mandarin all adopt the general locational question word meaning 'where', as can be seen in Table 4.9 above. These three languages, as most other sampled languages, do not formally differentiate the semantic subdomain PLACE.OBJECT.SG from PLACE.EVENT analyzed in §4.3.3. In this case, the singular facet of the subject is normally coded in the predicate, e.g., *is* in English or *ist* in German, or simply plays no role syntactically, e.g., Mandarin. Yet, there are also languages in the sample that utilize a distinctive interrogative to inquire about this kind of location information. Some of them are given in the following Table 4.11. A comparison of form can be drawn by observing the preceding Table 4.10.

	Chuj	Chamorro	Dogrib	Gwich'in	Parecís	Tenharim-Parintintin-Diahoi
43009012	*aj 'ay*	*mångge*	*welaedì*	*adaajii*	*aliyo*	*mahã*
40002002	*aj 'ay*	*månu nai*	*welaedì*	*adaajii*	*aliyako*	*mome*
43007011	*ay*	*mångge*	NA	*adaajii*	NA	*mahã*
43008019	*ajtil 'ay*	*mångge*	*edı̨ı̨*	*adaajii*	*aliyo*	*mahã*

Table 4.11: Interrogatives of PLACE.OBJECT.SG

In Chuj, *aj 'ay* is a part of the interrogative phrase *p'aj 'ay* meaning ‘where is it?’ (Hopkins 2012: 1). In Chamorro, the phrase *månu nai gaigi* ‘where is?’ is contracted to the form *mångge* (Topping & Dungca 1973: 235-236; Chung 2020: 420). Dogrib, an Eyak-Athabaskan language spoken in Canada, distinguishes the meaning ‘where is (he/she/it)?’ by applying the particular coding *welaedì* (Saxon & Siemens 1996: 229). Gwich’in employs the form *adaajii* for all these four contexts. The part *daajii* of this construction refers to ‘where (a person)’ (Alexander & Alexander 2011: 37). According to Brandão (2014: 333-334), Parecís uses the structure *aliyo* exclusively to convey the meaning ‘where is’. A special grammatical trait of this form is that it can cooperate alone with a nominal predicate without using the copula. Tenharim-Parintintin-Diahoi, a Tupian language spoken in Brazil, also presents a peculiar form *mahã* to ask for ‘where is it?’ (Pease 1968: 64).

4.3.4 PLACE.OBJECT.PL

The next sub-cluster constitutes only two contexts and the corresponding content can be seen in the following (4.7).

(4.7) *eng-x-bible-common*

43008010 *Jesus stood up and said to her, “Woman, **where** are they? Is there no one to condemn you?”*

42017017 *Jesus replied, “Weren’t ten cleansed? **Where** are the other nine?”*

Similar to the last sub-cluster PLACE.OBJECT.SG, the questions in this group pertain to the non-dynamic position of an animate subject with a copula serving as the predicate. However, the subjects of these two contexts are both in the plural and the corresponding copula in English is changed to *are*. Considering this semantic similarity as well as the subtle difference relating to the quantity of the involved subject, the label **PLACE.OBJECT.PL** is assigned to this sub-cluster.

In terms of the usage of interrogative, PLACE.OBJECT.PL bears a resemblance to PLACE.OBJECT.SG as well. That is, the majority of sampled languages do not create a particular form for the query information about this domain. The following Table 4.12 presents some special interrogatives for PLACE.OBJECT.PL. Serving as a comparison, forms used in 40002002 belonging to PLACE.OBJECT.SG in the corresponding languages are also given in the last row.

	Dogrib	**Burarra**	**Yine**	**Kagulu**	**Thakara**
43008010	*gılaedì*	*yina aburr-gaya*	*ginakatkana*	*wahoki*	*n'ata*
42017017	*gılaedì*	*yina aburr-gaya*	*ginakatkana*	*wowahe*	*barîkû*
PLACE.OBJECT.SG	*welaedì*	*yina an-gaya*	*ginaklu*	*hoki*	*arî kû*

Table 4.12: Examples of PLACE.OBJECT.PL

As described in the last section, Dogrib applies *welaedì* for the domain PLACE.OBJECT.SG. In contexts of PLACE.OBJECT.PL, this language takes the second part of the construction, i.e., *-aedì*, and then combines it with the 3rd person plural prefix *gı-* to yield the interrogative specifically for the stative location of the subject in the plural (Tłı̨chǫ Community Services Agency 2007: 67). This strategy, i.e., adding an affix for plurality to another interrogative, is commonly attested among languages marking count distinctions in form. Another example is provided by Burarra, an Australian Aboriginal language spoken by the Burarra people. In this language, the structure *yina…gaya* denotes the meaning of 'where, which of known possibilities' (Green 1987: 73). According to the number of referents, an appropriate prefix can be added in front of *gaya*. For contexts in 43008010 and 42017017, the prefix *aburr-* to express 3rd person plural is used (Glasgow & Glasgow 2011). In the same way, Yine, an

Arawakan language spoken in Peru, marks the plurality in the locational interrogative *ginaka* by attaching the plural suffix *-na* (Hanson 2010: 48, 325).

Kagulu and Thakara, two languages from the Niger-Congo family, are grammatically noted for the noun class system with different prefixes. For contexts of PLACE.OBJECT.PL, Kagulu applies *wa-*, an agreement prefix for the class indicating the plurality, as shown in Table 4.3.4 (Petzell 2008: 49, 90).[17] Similarly, the locational interrogative stem *-riku* in Thakara is attached to the pronominal prefix *ba-*, which marks the subject in the plural (Lindblom 1914: 20, 22).

4.3.5 PLACE.GOAL

According to the internal classification, this sub-cluster is composed of three interrogative contexts, as can be read in (4.8).

(4.8) *eng-x-bible-common*

43013036	"L*ord, **where** are you going?"*
43016005	*None of you ask me, '**Where** are you going?'*
43007035	***Where** does he intend to go that we can't find him?*

Compared to other previously analyzed sub-clusters, speakers of these three contexts intend to know the destination of the hearer or another person. Unlike the copula or stative verbs serving as the predicate in the last three sub-clusters, the motion verb 'go' is attested in all three questions in (4.8). This implies that the position of the actor is going to be changed, or more specifically, he/she is moving away from the original position. To be noticed, the motion of departure in all three contexts either takes place simultaneously with the interrogative utterance or is about to happen after the question is addressed, which can be inferred from the present continuous tense in the clause or the use of the verb *intend* in

[17] The appearance of *wowahe* in verse 42017017 might be in virtue of a contraction of the interrogative comprising the verb *kuwa*/uwa 'be' (Petzell 2008: 177). The locational question word *hoki* 'where' can occur in form of a clitic *-hi* or *-he*. In this case, the interrogative for the domain PLACE.OBJECT.PL is generated as *wa-uwa-he* (pl-be-where). Yet, as the rule of vowel coalescence in Kagulu notes, when the low vowel /a/ co-occurs with the high vowel /u/, it leads to the mid vowel /o/ (Petzell 2008: 43). Therefore, *wauwahe* should be transformed into *wowahe*.

43007035. In order to emphasize the dynamic state and the queried goal in this sub-cluster, it is labeled **PLACE.GOAL**. Table 4.13 below lists some special interrogative forms to ask for such a destination. As for comparison, interrogatives used for PLACE.EVENT (43001038b) in these example languages are also provided.

To mark the motion *go*, there are generally two ways in which the sampled languages build the interrogative. Firstly, prepositions or affixes indicating the direction 'to' can be added to the basic interrogative meaning 'where'. A typical example is *wohin* in German, which is composed of the question word *wo* 'where' and the adverb *hin* 'to'. Other examples are found in the first four languages in Table 4.13. In Ayacucho Quechua, a Quechuan language spoken in Peru, *may* functions as the interrogative stem for spatial concepts and is normally combined with various suffixes (Zariquiey & Córdova 2008: 101). In all three contexts about GOAL, *may* is followed by the suffix *-ta*, which signifies the goal when the actor is a human being, whereas the suffix *-pi* expressing a stative location shows up in the question for PLACE.EVENT (Parker 1969: 40). In Korean, when *lo* 로 serves as a directional postposition, it marks the direction of moving away of the agent. Similar to *-pi* in Ayacucho Quechua, *e* 에 appearing in 43001038b indicates the unchanged place where something happens (Hoppmann 2011: 217).

The next two languages from the Austronesian family in Table 4.13 also apply the same strategy as Korean, i.e., combining the directional preposition meaning 'to' with the general question word 'where' to yield the construction 'to where'. In Batak Karo and Ma'anyan, the direction 'to' is marked by *ku* (Woollams 1996: 153, 225) and *ma* (Sundermann 1912: 232), respectively.

	Ayacucho Quechua	**Korean**	**Batak Karo**	**Ma'anyan**	**Turkish**	**Finnish**
43013036	*maytataq*	어디로	*ku ja*	*maawe*	*nereye*	*minne*
43016005	*maytataq*	어디로	*ku ja*	*maawe*	*nereye*	*minne*
43007035	*maytaya*	어디로	*ku ja*	*maawe*	*nereye*	*minne*
PLACE.EVENT	*maypitaq*	어디에	*i ja*	*hang awe*	*nerede*	*missä*

Table 4.13: Examples for PLACE.GOAL

The second structure of the interrogative for PLACE.GOAL is also derived from the basic form for PLACE. Languages of this kind then mark the semantic difference by means of case suffixes. For instance, based on *nere*-, a fundamental pronoun referring to 'where', Turkish applies the dative case marker *-(y)e* to denote the destination (Göksel & Kerslake 2005: 436), as shown in Table 4.13. In some other languages, it is already difficult to separate the derivational structure into meaningful segments. An example is *minne* in Finnish, an interrogative particularly to ask for GOAL and a synonym of *mihin*. The subtle differentiation lies in that the denotation of *minne* is less precise than *mihin*. According to Karlsson (2008: 113-114), *mihin* is inflected from the question word *mikä* 'what, which' and is related to the illative case.

As seen before, languages that formally mark the semantic property of destination are inclined to build the interrogative for PLACE.GOAL on the basis of another question word, mostly the basic form 'where'. It is rare across languages that a form is specifically coined for the inquiry about GOAL. Noticeably, Acehnese, an Austronesian language spoken in Indonesia, provides a unique interrogative only for this kind of information, as shown in the following Table 4.14.

	Acehnese	**Croatian**
43013036	*ho*	*kamo*
43016005	*ho*	*kamo*
43007035	*ho*	*kamo*
PLACE.EVENT	*dipat*	*gdje*

Table 4.14: Unique forms for PLACE.GOAL

In Acehnese, the question word *pat* can refer to both 'where' and 'where from', while it is not allowed in the context of 'where to'. To ask for the destination, the form *ho* should be used, which appears morphologically unrelated to *pat* (Durie 1985: 152). A similar case can also be found in Croatian. In this language, the interrogative *kamo* for GOAL is independent from *gdje* 'where'. Yet, unlike the morphological uniqueness of *ho* in Achenese, *kamo* shares resemblance with other words relating to the meaning 'to somewhere', e.g., *ovamo* 'to here', *tamo* 'to there' and *onamo* 'to there' (Brown & Alt 2004: 36).

4.3.6 PLACE.FROM.ORIGIN

The sub-cluster discussed in this section comprises two contexts and they are presented in following (4.9).

(4.9) *eng-x-bible-common*

66007013b *and **where** did they come from?*

43019009 *He went back into the residence and spoke to Jesus, "**Where** are you from?"*

(4.10) *deu-x-bible-schlachter*

43019009 *und ging wieder in das Amthaus hinein und sprach zu Jesus: **Woher** bist du?*

In the two contexts in (4.9), speakers address a question in order to obtain the start point of another person. In English, the direction is not marked in the question word *where*. As the strategy to denote the direction, the preposition *from* occurs in both questions. Especially for 43019009 in which the predicate is not a motion verb, as *come* in 66007013b, but the copula *are*, the use of *from* is obligatory to disambiguate the dynamic condition from a stative event. Such a composition of a copula and a preposition indicating the direction is not attested in PLACE.GOAL. All three examples of PLACE.GOAL use the verb *go* to explicitly indicate the dynamic locational change.

In contrast, German provides a different case. This language formally distinguishes interrogatives for *wo* 'where', *wohin* 'where to', and *woher* 'where from'. In this sense, no additional directional device is needed in the following content, even when the copula solely serves as the predicate, as 43019009 in German translation *deu-x-bible-schlachter* in (4.10) above. For this sub-cluster, the name **PLACE.FROM.ORIGIN** is assigned.

In terms of the form of interrogative, many languages in the sample apply the same strategy as in English, i.e., an identical structure for both 'where' and 'where from'. Similar to the formation for GOAL, interrogative for ORIGIN is mostly related to or derived from the basic

form meaning ‘where’ as well. Six examples are given in Table 4.15 below. To provide a comparison with the subdomain GOAL and EVENT, examples here are selected from languages already mentioned in §4.3.6.[18]

For information about ORIGIN, Ayacucho Quechua adds the ablative suffix *-manta* ‘from’ to the locational stem *may* (Zariquiey & Córdova 2008: 96). In a similar way, Korean applies the postposition *seo* 서 indicating the start point based on the stem *eodi* 어디 for PLACE (Hoppmann 2011: 217). In Batak Karo, the interrogative used for contexts 66007013b and 43019009 is composed of three parts — *i* ‘at’, *ja* ‘where’, and *nari* ‘from’ (Woollams 1996: 152, 225). Interestingly, different from the preposition *ku* indicating ‘to’, the direction ‘from’ is marked via the postposition *nari* which should be placed after the stem *ja*. Ma'anyan also chooses the preposition *teka* to express the direction ‘from’ (Sundermann 1912: 232).

	Ayacucho Quechua	**Korean**	**Batak Karo**	**Ma'anyan**	**Turkish**	**Finnish**
66007013b	*maymantata*	어디서	*i ja nari*	*teka awe*	*nereden*	*mistä*
43019009	*maymantam*	어디서	*i ja nari*	*teka awe*	*neredensin*	*mistä*
PLACE.GOAL	*maytataq*	어디로	*ku ja*	*maawe*	*nereye*	*minne*
PLACE.EVENT	*maypitaq*	어디에	*i ja*	*hang awe*	*nerede*	*missä*

Table 4.15: Examples for PLACE.FROM.ORIGIN

Turkish and Finnish exemplify how languages with an ample case system differentiate the interrogative of each locational subdomain. Opposed to the dative case *-(y)e* denoting GOAL, Turkish uses the ablative case suffix *-den* for ORIGIN (Göksel & Kerslake 2005: 45). In Finnish, when the elative suffix *-stä* ‘out of inside’ is combined to the question word *mika* ‘what, which’, the resulting form *mistä* is indicative of the start place (Karlsson 2008: 113-114, 168).

Among all sampled language, only Croatian provides a seemingly idiosyncratic form *odakle* for ORIGIN (Alexander & Elias-Bursać 2006: 48). Compared to the interrogative *gdje* ‘where’ and *kamo* ‘where to’ in this language, as given in Table 4.14 above, no morphological

[18] Examples are found in 43013036 for PLACE.GOAL and 43001038b for PLACE.EVENT.

connection can be observed between these three forms. And I have not been able to find any information about their etymological origin.

4.3.7 PLACE.FROM.SOURCE

Finally, three interrogative contexts are assigned to the last sub-cluster of the PLACE category and they are given in (4.11).

(4.11) *eng-x-bible-common*

40013054a	***Where*** *did he get this wisdom?*
41006002a	*Many who heard him were surprised. "**Where** did this man get all this?"*
40013056	*And his sisters, aren't they here with us ? **Where** did this man get all this?*

These three contexts have very similar content in which the speaker wants to know the source of something abstract. In 40013054a, the inquired object is wisdom. From the context around 41006002a and 40013056, it can be inferred that the involved items in these two questions refer to skills or abilities, respectively. Unlike all other sub-clusters discussed above, the targeted answer to these questions is not about a real place or direction in space, but the way of someone being wise or capable. In the semantic perspective, this is more closer to the domain of the MANNER category that is normally encoded with the question word *how* in English. Thus, to some extent, these questions can also be regarded as rhetorical. However, despite the abstract facet of the content, English as well as many other sampled languages still uses the locational interrogative meaning 'where' or 'where from' to build questions in (4.11). Considering that the predicate of these three contexts is the motion verb *get* that normally implies a change from the previous circumstance, I label this sub-cluster **PLACE.FROM.SOURCE**.

However, not every language applies the locational interrogative for PLACE.FROM.SOURCE. Instead, different variations are observed. Western Arrarnta, a Pama-Nyungan language spoken in Australia, and Ayacucho Quechua respectively adopt *nthakin* and *imaynanpi*, which

should be translated as ‘how’ in English (Strehlow et al. 2018: 294; Soto-Ruiz 1979: 63). *Imaynanpi* in Ayacucho Quechua expresses additionally the surprise of the speaker. Dogrib is morphologically highly complex. For the first and the third context in this sub-cluster, this language applies the expression *dànıt’à* ‘how is it?/how come?/why is it?’ and *dànìghǫ* ‘why/ for what reason?’ (Saxon & Siemens 1996: 12). Also for these two contexts, Cherokee, an endangered Iroquoian language of the Cherokee people, utilizes *gado* ‘what’ of the THING category (Montgomery-Anderson 2008: 481). The use of these interrogatives might result from different interpretation and expression of the SOURCE domain. Formulations like *Why does he get this wisdom?* or *How is it that he gets this wisdom?* will not mislead the perception of information about SOURCE.

Interrogatives referring to PERSON are also attested in some languages for SOURCE. For 40013054a, *ma’g̃a* ‘who’ is employed in Tenharim-Parintintin-Diahoi (Betts 2012: 159). This language then applies *marã* ‘how’ and *maraname* ‘how is it’ for the next two contexts (Betts 2012: 158). Kuku Yalanji, a Pama-Nyungan language spoken in Australia, marks the first and the third context with *wanyanamun*, a question word meaning ‘who’ in possessive, while *wanyu* ‘what’ appears in the second context (Patz 2002: 79). Another mixed use of interrogatives is found in Western Huasteca Nahuatl from the Uto-Aztecan family. This language adopts *ajqueya* ‘who’ for the first two contexts and *canque* ‘where' for the last question (Beller & Beller 1977: 221-222). However, because of the deficiency of grammatical information, how interrogatives pertaining to the PERSON domain function in questions about SOURCE cannot be explained yet.

In summary, we can tell that the formation of this sub-cluster is much more complicated than the other kinds of PLACE. Unlike other sub-clusters discussed before in this chapter whose content can be well determined through some specialized markers, the choice of interrogatives for PLACE.FROM.SOURCE largely depends on how speakers want to formulate the question. Nevertheless, it can be concluded that the majority of sampled languages tend to relate the information about SOURCE to the PLACE category, even though the intention of questions is not really about a specific locational concept, as noted at the beginning of this section. It remains as an open question why SOURCE is interpreted as being close to the PLACE category in so many languages.

4.3.8 Internal structure of PLACE.FROM

Summarized the last two sections, forms used for the sub-clusters PLACE.FROM, i.e., ORIGIN and SOURCE, are mostly derived from the basic locational interrogative that normally occurs in questions about EVENT. For the sub-cluster ORIGIN, only Croatian provides a question word *odakle* that is unrelated to other locational interrogatives. So far, I have not found a case that a language creates a specialized form just for the sub-cluster SOURCE. All sampled languages recruit interrogatives from other categories for the subdomain SOURCE. In terms of the codings for PLACE.FROM, three major types are identified within all occurrences in the sampled languages.

Firstly, languages do not distinguish between the stative place and the direction 'from'. The same interrogative is then also applicable to the subdomain SOURCE. Among the sample, 33 languages adopt this usage. English is representative of this type. As an example, questions respectively belonging to EVENT (43001038b), SOURCE (40013054a) and ORIGIN (66007013b) in three languages are shown in (4.12) with interrogative constructions in boldface or being underlined.

(4.12) EVENT = SOURCE = ORIGIN

a. Mandarin

EVENT 老師, 你 住在 哪裡*?*

SOURCE 這 個 人 從 哪裡 得到 如此 的 智慧*?*

ORIGIN 他們 從 哪裡 來*?*

b. Northern Dagara

EVENT *Arabi i, -a yɛ̂r-bir ŋa pɛr ɪ Wul-wul-kara-**nyɪnɛ** n'a fv kpɪɛr?*

SOURCE ***Nyɪnɛ** n'a v páw a yã-bãwfv ŋa?*

ORIGIN *ɛ́ **nyɪnɛ** n'a bɛ yi wa ?*

c. Parauk

EVENT *Rabi eue , ot Maix **dee mawx**?*

SOURCE *Pui in pon pingnya cuyi mai siawp sibrawm khankix khaing **dee mawx** lie?*

ORIGIN *mai kix kaoh khaing **dee mawx**?*

The second common coding for PLACE.FROM is represented by German. In languages of this kind, the interrogative for stative location is distinct from the one for dynamic origin. The form used for ORIGIN is also adopted in contexts of SOURCE. In the sample, 31 languages demonstrate this pattern. Examples from two languages are given in (4.13).

(4.13) EVENT ≠ SOURCE = ORIGIN

a. German

EVENT *Rabbi (das heißt übersetzt: Lehrer), **wo** wohnst du?*

SOURCE ***Woher** hat dieser solche Weisheit?*

ORIGIN *und **woher** sind sie gekommen?*

b. Central Yupik

EVENT ***Nani** uitalarcit Rabbai?*

SOURCE ***Naken-mi**[19] una yuk elitmek mat'umek pinga?*

ORIGIN ***naken-llu**[20] tekitat?*

c. Icelandic

EVENT *Rabbí (það þýðir meistari), **hvar** dvelst þú?*

SOURCE ***Hvaðan** kemur honum þessi speki og kraftaverkin?*

ORIGIN *og **hvaðan** eru þeir komnir?*

[19] According to Miyaoka (2012: 272), *-mi* is a locational suffix for singular.

[20] According to Miyaoka (2012: 21), *-llu* is an enclitic meaning ‘and’.

The common ground shared by languages belonging to the first two types is the parallel between ORIGIN and SOURCE. This pattern prevails in most sampled languages. As the opposite, languages of the third type encode SOURCE and EVENT with the same marking, whereas questions about ORIGIN take the interrogative specifically meaning 'where from'. In the Spanish translation chosen for this study, such a case is found in which *dónde* 'where' is used for the first and the third verse referring to SOURCE in (4.11) above, while the interrogative construction for 'where from' is *de dónde*. In addition to Spanish, another nine languages pertain to this kind. In (4.14), the same contexts as in (4.12) and (4.13) in three languages are given serving as examples.

(4.14) EVENT = SOURCE ≠ ORIGIN

a. Spanish

EVENT	*¿**dónde** estás alojado?*
SOURCE	*¿**Dónde** consiguió este hombre esta sabiduríat?*
ORIGIN	*¿*[…] *y **de dónde** vinieron?*

b. Irish

EVENT	***Cá** bhfuil cónaí ort?*
SOURCE	***Cá** bhfuair sé seo?*
ORIGIN	*agus **cad as** ar tháinig siad?*

c. Iban

EVENT	*Rabbi, "(reti nya, Pengajar,) " **dini** alai Nuan diau?*
SOURCE	***Dini** alai orang tu bulih penemu-dalam tu , enggau kereja ajih tu?*
ORIGIN	*lalu **ari ni** penatai sida?*

Apart from these three main types, there are other two possibilities attested in the sample. Languages might apply the identical form for EVENT and ORIGIN, whereas the interrogative of other semantic categories are borrowed for SOURCE. Examples of this kind, i.e., EVENT = ORIGIN ≠ SOURCE, are Dogrib and Cherokee presented in the last section. In the second minor

type, EVENT, ORIGIN, and SOURCE are all differently marked in form. Kuku Yalanji and Ayacucho Quechua belong to this group.

To be noticed, there are also some sampled languages that encode questions for PLACE.FROM with very mixed forms in the Bible translation. In this case, it is difficult to properly categorize them into any type just mentioned. Table 4.16 below summarizes languages that belong to these five types.

Type	Language
EVENT = SOURCE = ORIGIN	Baoulé, Eastern Bru, Chuj, El Nayar Cora, Northern Dagara, Toro So Dogon, Dii, English, Ejagham, Maasina Fulfulde, Igbo, Japanese, Kagulu, Northern Kissi, Nomaande, Makasar, Coatlán Mixe, Totontepec Mixe, Masaaba, Nama, Yine, Highland Popoluca, Parauk, Rundi, Noon, Lowland Tarahumara, Tharaka, Wolof, Yucatec Maya, Mandarin, Francisco León Zoque
EVENT ≠ SOURCE = ORIGIN	Acehnese, Batak Karo, Burarra, Car Nicobarese, Catalan, Tabasco Chontal, Danish, German, North Alaskan Inupiatun, Central Yupik, Finnish, Gagauz, Paraguayan Guaraní, Gwich'in, Hopi, Croatian, Hungarian, Icelandic, Jarai, Karakalpak, Halh Mongolian, Korean, Madurese, Ma'anyan, Tetelcingo Nahuatl, Romanian, Tagalog, Turkmen, Turkish, Uyguhr, Vietnamese
EVENT = SOURCE ≠ ORIGIN	Balinese, Garifuna, Czech, Chamorro, Welsh, Irish, Iban, Inga, Dutch, Spanish
EVENT = ORIGIN ≠ SOURCE	Cherokee, Dogrib, Machiguenga, Western Huasteca Nahuatl, Tenharim-Parintintin-Diahoi
EVENT ≠ SOURCE ≠ ORIGIN	Western Arrarnta, Kuku-Yalanji, Ayacucho Quechua

Table 4.16: Codings types of PLACE.FROM

4.3.9 Summary

This chapter elaborated on the cluster of PLACE and its internal structure. The first sub-cluster PLACE.EVENT discussed in §4.3.2 contains the questions asking for the stative location of an occurrence. In all contexts of this sub-cluster, the predicates are verbs indicating a concrete motion happening in an unchanged site. As the counterpart, some languages apply a form exclusively referring to 'be where', which involves no specific action but the stative location of the referent. In this case, the copula verb *BE* serves as the predicate in the questions. For this meaning, two sub-clusters, i.e., PLACE.OBJECT.SG and PLACE.OBJECT.PL, are identified with a numeral difference in §4.3.3 and §4.3.4.

Three sub-clusters related to dynamic motions are found through the algorithm. Correspondingly, several sampled languages mark the interrogative codings with different devices indicating the direction. In §4.3.5, the sub-cluster PLACE.GOAL is composed of contexts acquiring someone's destination. The reverse direction, i.e., the start point, is queried in the sub-cluster PLACE.FROM.ORIGIN presented in §4.3.6.

Noticeably, the algorithm suggested another sub-cluster related to the direction 'from', i.e., PLACE.FROM.SOURCE in §4.3.7. The inquiry made in these contexts is not about the location but the manner of being a state. The interrogatives attested in this subgroup show a mixture of different categories, e.g., PLACE.EVENT, PLACE.FROM.ORIGIN, REASON and MANNER. Given that the morphological structure of interrogative codings displays a salient similarity between PLACE.EVENT, PLACE.FROM.ORIGIN and PLACE.FROM SOURCE, a summary of this facet is given in §4.3.8. Yet, since no dedicated interrogative coding is found for PLACE.FROM.SOURCE, it is dubious whether it can represent a cross-linguistic comparative concept, as noted in §1.3.2 before. This issue will be further discussed in Chapter 6 later.

4.4 Cluster of PERSON

The current chapter will present the cluster of the PERSON category. The basic information about this cluster will be first given in §4.4.1. Compared to the last two clusters TIME and PLACE, more finer subdomains are identified in this group. In the subsequent sections, the following sub-clusters will be discussed:

- §4.4.2 — PERSON. ROLE
  - §4.4.2.1 — PERSON.ROLE.AGENT
  - §4.4.2.2 — PERSON.ROLE.PATIENT
  - §4.4.2.3 — PERSON.ROLE.RECIPIENT
  - §4.4.2.4 — PERSON.ROLE.GOAL
- §4.4.3 — PERSON.ASCRIPTION
- §4.4.4 — PERSON.IDENTITY
  - §4.4.4.1 — PERSON.IDENTITY.2SG
  - §4.4.4.2 — PERSON.IDENTITY.3SG
  - §4.4.4.3 — PERSON.IDENTITY.PL

- §4.4.5 — PERSON.SELECTION
- §4.4.6 — PERSON.POSSESSOR
- §4.4.7 — PERSON.KIND

4.4.1 Overview

The third cluster represents the semantic domain PERSON. It is composed of 98 interrogative contexts. The average silhouette width of this cluster is 0.48, which is slightly better than the PLACE cluster, as seen in Figure 4.1 presented previously. To gain a quick impression of this cluster, Table 4.17 gives information about some selected interrogative contexts.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	41005030	0.63642	*who*	*wer*	誰
2	42008045	0.63502	*who*	*wer*	誰
3	43007020	0.62809	*who*	*wer*	誰
4	45011034b	0.62433	*who*	*wer*	誰
5	44008034	0.44081	*whom*	*wem*	誰
6	42022027	0.40908	*which one*	*wer*	哪
7	43012038b	0.38810	*whom*	*wem*	誰
8	42020024a	0.35824	*whose*	*wessen*	誰的
9	40022020a	0.34684	*whose*	*wessen*	誰的
10	42010036b	0.32171	*which one*	*wer*	誰
11	42007042	0.25873	*which*	*welcher*	哪一
12	40008027	0.05737	*what kind of*	*wer*	什麼

Table 4.17: Verse selection of cluster PERSON

Figure 4.7 below draws the distribution of data points representing interrogative contexts of cluster PERSON. Unlike the clearly separated structure of the last cluster PLACE, the majority of data points of this cluster assemble at the left-center in the graph. Only around ten dots are loosely spread from the center to the right. Yet, when we observe the vertical dimension, a more clear stratification among dots appears at the left. Just as in the horizontal dimension, there is also a gathering of points at the center vertically relative to which some small

groupings can be recognized at the top and bottom. The various symbols marking different question words used in the English translation *eng-x-bible-common* reveal that points located at the bottom mainly refer to interrogative contexts encoded with *whom* and *whose*, whereas most other contexts, whose points crowd at the center, are asked with the question word *who*.

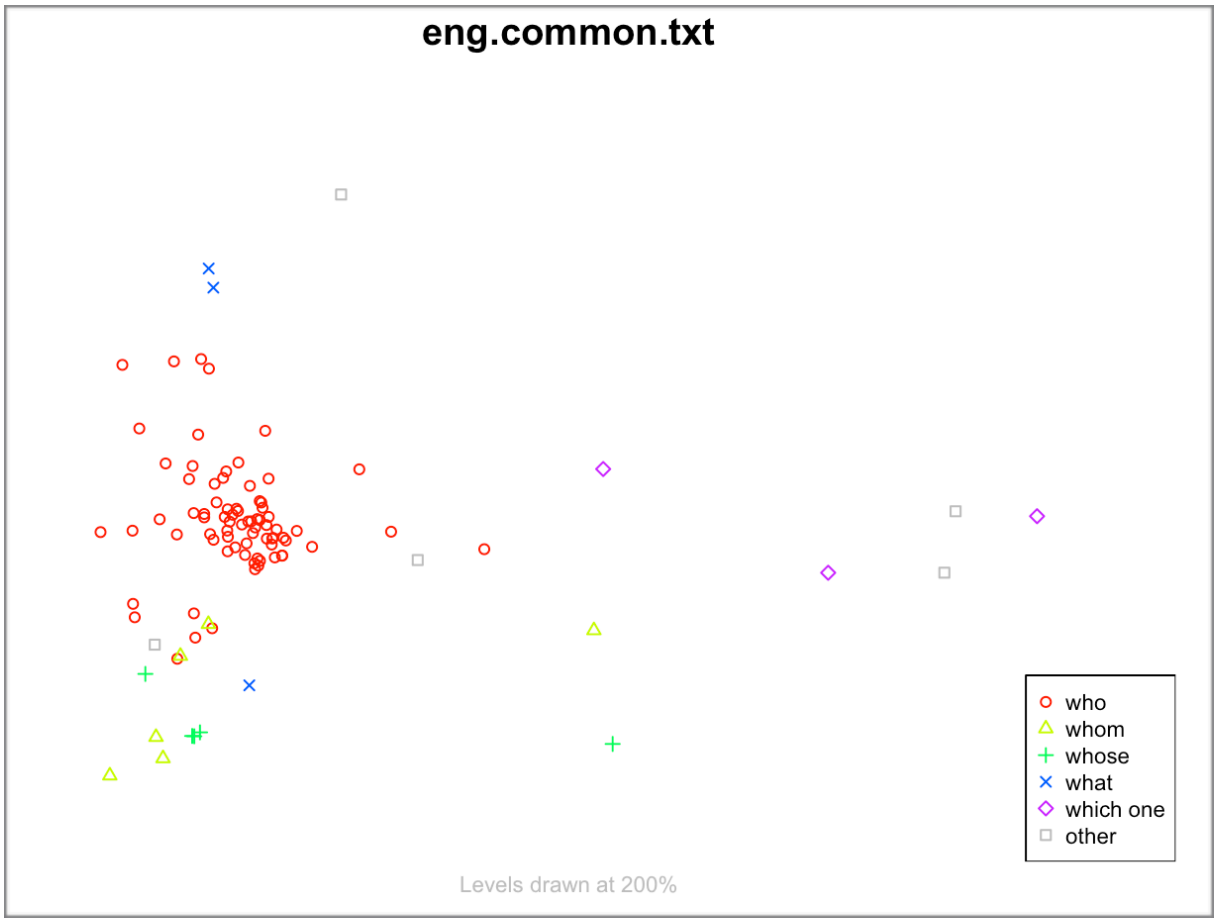

Figure 4.7: MDS plot of cluster PERSON

Figure 4.8 below shows the result of the second level of clustering of cluster PERSON. 98 contexts are optimally divided into two sub-clusters. They are respectively composed of 93 and 5 contexts. The five contexts that are assigned to the second sub-cluster are all encoded with *which* or *which one*, which refers to the semantic concept of SELECTION. In contrast compared to these five questions, other 93 interrogative contexts are so similar across sampled languages that the program is still unable to further classify them into smaller groups. Therefore, results with more sub-clusters should be taken into account.

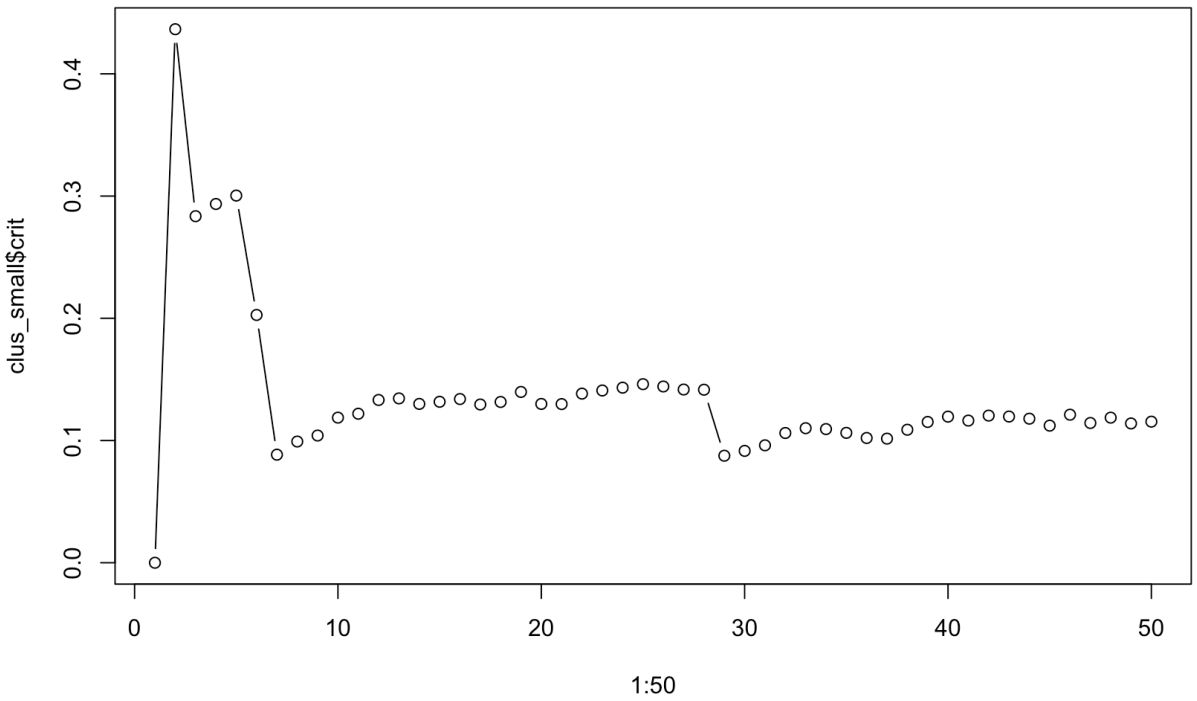

Figure 4.8: Suggested sub-clusters of PERSON

The suboptimal grouping, as can be seen in Figure 4.8, is found with five sub-clusters. This time, in addition to the one related to SELECTION (see §4.4.5 PERSON.SELECTION), two more sub-clusters are identifiable. They contain six contexts encoded with the question word *whose* in English (see §4.4.6 PERSON.POSSESSOR) and 3 contexts asked with *what* or *what kind of* (see §4.4.7 PERSON.KIND), respectively. Nevertheless, there are still two sub-clusters of which the semantic trait is difficult to recognize. In this case, clustering with a higher number of results, e.g., 19, 25, and 40 sub-clusters, will also be considered. These details will be discussed in §4.4.2 to §4.4.4 first.

4.4.2 PERSON.ROLE

As noted above, there are two uninterpretable subgroups after the secondary optimal clustering. For these contexts, many languages just utilize a single interrogative. By means of further sub-clustering, some of them are separated into smaller subsets. In the following, the eight following groups are identifiable: PERSON.ROLE.AGENT, PERSON.ROLE.PATIENT, PERSON.ROLE.GOAL, PERSON.ROLE.RECIPIENT, PERSON.IDENTITY.2SG, PERSON.IDENTITY.3SG, PERSON.IDENTITY.PL, and PERSON.ASCRIPTION. This section will depict the first three sub-

clusters related to semantic roles. The sub-cluster PERSON.ASCRIPTION will be elaborated in §4.4.3. Finally, a discussion about three sub-cluster related to IDENTITY is found in §4.4.4.

4.4.2.1 PERSON.ROLE.AGENT

The first group is composed of 22 contexts.[21] Five of them with the highest silhouette width are provided in (4.15).

(4.15) *eng-x-bible-common*

41005030	***Who** touched my clothes?*
42008045	***Who** touched me?*
43007020	***Who** wants to kill you?*
42022064	***Who** hit you?*
42012014	*Man, **who** appointed me as judge or referee between you and your brother?*

In these six questions, the question word *who* in English syntactically serves as the subject in the clause. From a semantic perspective, subjects in these contexts actively perform an action that has an impact on someone else. The predicates in these questions are correspondingly transitive verbs. Therefore, questioners of contexts in this sub-cluster want to obtain information about the agent of an event. Consequently, this sub-cluster is designated **PERSON.ROLE.AGENT**.

	Western Arrarnta	Kuku-Yalanji	Hopi	Turkish
41005030	*ngunhalama*	*wanjungku*	*hak*	*kim*
42008045	*ngulama*	*wanjungku*	*hak*	*kim*
42022064	*ngunhalama*	*wanjungku*	*hak*	*kim*

Table 4.18: Examples for PERSON.ROLE.AGENT

[21] 24 contexts were assigned to this sub-cluster. However, the contexts in 43012034a and 43013025 do not present a transitive usage. Thus, they are manually removed.

To ask for the agent of an event, most sampled languages tend to apply the general interrogative of the PERSON category. However, it is also not rare that languages formally draw a distinction between the agent and other personal roles with different morphological strategies. Table 4.18 above provides four examples from the sampled languages.

Western Arrarnta differentiates the subject of an intransitive verb and a transitive verb. According to Strehlow et al. (2018: 288), this language utilizes the form *ngu(nha)* for the former meaning, whereas the element *-la* is applied for the subject of a transitive verb. Kuku-Yalanji is a language with an ergative-absolutive alignment. That is, the agent of a transitive verb is marked with the ergative case, while the subject of an intransitive verb and the object of a transitive verb take the absolutive case. In contexts of AGENT, this language chooses the interrogative *wanjungku* in the ergative case to build the question (Patz 2002: 79).

Hopi, a Uto-Aztecan language spoken in the United States, employs different interrogatives to mark the number and syntactic function of the targeted entity. For this sub-cluster, Hopi adopts the form *hak* indicating a singular subject (Kalectaca & Langacker1978: 108). Turkish has an extensive case system. In this language, the subject of a transitive verb is marked with the nominative case, e.g., *kim* used in this sub-cluster, which is the same as the subject of an intransitive verb (van Schaaik 2020: 64). In contrast, other arguments of a transitive verb might take the dative or accusative case.

4.4.2.2 PERSON.ROLE.PATIENT

The second subgroup comprises two similar questions, as given in (4.16). In both contexts, questioners intend to know the person that the addressees are searching for. Syntactically, the question word holds the place of the object of a transitive verb in the clause. In opposition to the semantic role of agent in the last sub-cluster, the inquired human referent does not deliberately perform the action but is the target of the action. In this sense, this sub-cluster is given the label **PERSON.ROLE.PATIENT**.

As seen in (4.16), English uses the same question word *who* in this utterance. However, interrogatives in these questions are inflected in many other sampled languages for case. Table 4.19 below provides eight examples. Interrogatives used for PERSON.ROLE.AGENT in 41005030 are also listed for comparison.

(4.16) *eng-x-bible-common*

43018004 *Jesus knew everything that was to happen to him, so he went out and asked, "**Who** are you looking for?"*

43020015b *"**Who** are you looking for?" Thinking he was the gardener , she replied, "Sir , if you have carried him away, tell me where you have put him and I will get him."*

	German	Hungarian	Turkish	Colombian Inga	Korean	North Alaskan	North Saami	Finnish
43020015b	*wen*	*kit*	*kimi*	*pitatak*	누구를	*kiña*	*gean*	*ketä*
43018004	*wen*	*kit*	*kimi*	*pitatak*	누구를	*kiña*	*gean*	*ketä*
AGENT	*wer*	*kicsoda*	*kim*	*pitak*	누구	*kia*	*gii*	*kuka*

Table 4.19: Examples for PERSON.ROLE.PATIENT

In German and Hungarian, the basic question word for PERSON takes the accusative case in these two contexts. Correspondingly, they are changed to *wen* and *kit*, respectively (Kenesei et al. 1998: 192). Turkish and Colombian Inga, a Quechuan language spoken in Colombia, also mark the general question word for PERSON with the accusative suffix, i.e., *-mi* (van Schaaik 2020: 64) and *-ta* (Levinsohn & Galeano L. 1981: 76-77), to express the targeted entity of the action in this sub-cluster. Korean applies a different morphological strategy. This language combines *nugu* 누구 'who' with the postposition *leul* 를 indicating the accusative case to denote the object in the clause (Hoppmann 2011: 217). In North Alaskan Inupiatun, the interrogative *kiña* in these two questions is marked with the absolutive case, which refers to the object of a transitive verb as well as the subject of an intransitive verb (Lanz 2010: 126).

However, the functions of cases are not always parallel between languages. In North Saami from the Uralic family, the basic question word *gii* 'who' are inflected to the genitive case *gean* in both contexts. According to Aikio & Ylikoski (2010: 44-45), the genitive case in North Saami normally encodes the object of a transitive verb that semantically undergoes the performance conducted by the agent. In contrast, the question word *ketä* in Finnish is in the partitive case to express an object with indefinite quantity (Karlsson 2008: 134, 207).

4.4.2.3 PERSON.ROLE.RECIPIENT

This group is composed of two questions presented in (4.17). In these two questions, verbs denote an emotional performance towards the inquired human referent. In this sense, the queried information is not about a patient of an action but a recipient of an emotion. Thus, this sub-cluster is called **PERSON.ROLE.RECIPIENT**. To be noticed, the label of recipient used in this sub-cluster is slightly different from the classic definition in meaning. A recipient normally refers to the goal of a changed possession of a concrete entity, whereas an emotion is abstract. Nevertheless, a recipient of either a concrete substance or abstract emotion usually serves as the indirect object in the clause, which is differently marked from the transitive subject and direct object. In English, the question word follows the preposition *against* and *with* and thus appears in the form *whom*. Table 4.20 provides some other examples are given. Comparison can be made with the interrogatives used for PERSON.ROLE.AGENT (41005030).

(4.17) *eng-x-bible-common*

58003018 *And **against whom** did he swear that they would never enter his rest, if not against the ones who were disobedient?*

58003017 *And **with whom** was God angry for forty years?*

	North Saami	**Finnish**	**German**
58003018	*geaidda*	*keille*	*welchen*
58003017	*geaidda*	*keille*	*welchen*
AGENT	*gii*	*kuka*	*wer*

Table 4.20: Examples for PERSON.ROLE.RECIPIENT

North Saami takes the form *geaidda* indicating the illative case in the plural for both questions. According to Aikio & Ylikoski (2010: 51, 70), one of the basic functions of the illative case is to denote recipients or beneficiaries. Yet, it is not given in this grammar whether this language distinguishes the recipient of a concrete entity from an abstract emotion. Finnish applies *keille* in the allative case in the plural. As the grammar notes

(Karlsson 2008: 181), the allative case is used to mark the movement 'towards a surface' or 'to someone'. Since a recipient can be counted as a special kind of goal of a changed state, the usage of the allative case in Finnish is understandable. German does not mark the direction 'to' in the interrogative *welchen* 'which' appearing in these two questions. However, this form is indicative of the dative case in the plural, which is decided by verbs *zürnen* 'be angry with' and *schwören* 'swear' and indicates a syntactic difference from a typical direct object of an action.

4.4.2.4 PERSON.ROLE.GOAL

The next group **PERSON.ROLE.GOAL** contains two contexts presented in (4.18). The actions taking place in both questions imply a directional change either between referents or in terms of location. The speakers of these questions want to acquire the goal of such a change. In English, the locational question word *where* is used in the first context. However, by reading the surrounding text and translations in other languages, it can be inferred that the goal of the action *go* does not refer to a specific place, but a person. In this sense, the denotation of *where* in this question is equal to *to whom*. Similar to the last group, the information about GOAL is normally expressed with case markers in sampled languages. See examples given in Table 4.21.

(4.18) *eng-x-bible-common*

43006068 *Lord , **where** would we go?*

43012038b *To **whom** is the arm of the Lord fully revealed?*

	German	Turkish	Korean	North Alaskan Inupiatun	Finnish	Kuku-Yalanji
43012038b	*wem*	*kime*	누구에게	*kimun*	*kenen*	NA
43006068	*wem*	*kime*	누구에게로	*kimun*	*kenen*	*wanyanda*
AGENT	*wer*	*kim*	누구	*kia*	*kuka*	*wanjungku*

Table 4.21: Examples for PERSON.ROLE.GOAL

In contrast to the accusative case used for the direct object of a transitive verb, the question word *wem* in German and *kime* in Turkish are both in the dative case (van Schaaik 2020: 64). In Korean, the dative postposition *ege* 에게 meaning 'from…to' is added after *nugu* 누구 'who' (Hoppmann 2011: 217). North Alaskan Inupiatun combines the singular allative suffix *-mun* meaning 'to' with the interrogative stem *-ki* for PERSON (Lanz 2010: 126). To mark the inquired person as the goal of an action, the corresponding question word *kenen* is in the singular genitive case in Finnish (Karlsson 2008: 207). The final example language Kuku-Yalanji provides the form *wanyanda* in the locative case in this sub-cluster (Patz 2002: 79).

4.4.3 PERSON.ASCRIPTION

The following three contexts given in (4.19) are assigned into another sub-cluster. These questions are about how a person is said or thought by other people. Despite the copula construction occurring in the expression in English, the targeted referent should be reckoned as the topic of the action *say*. Given this semantic characteristic, the label **PERSON.ASCRIPTION** is given to this sub-cluster. In English and most sampled languages, no formal difference is shown in interrogatives for these contexts. The general question word for PERSON is applied. However, the action *say* is marked in the interrogative construction in some languages. Four examples are found in the sample, as shown in Table 4.22 below.

(4.19) *eng-x-bible-common*

40016013 ***Who*** *do people say the Human One is?*
41008029b ***Who*** *do you say that I am?*
42009018 ***Who*** *do the crowds say that I am?*

	North Saami	Korean	German	Hungarian
40016013	*geanin*	누구라고	*wen*	*kinek*
42009018	*geanin*	누구라고	*wen*	*kinek*
AGENT	*gii*	누구	*wer*	*kim*

Table 4.22: Examples for PERSON.ASCRIPTION

In North Saami, *gii* ‘who' is inflected to *geanin* in the essive case. When this case is used in a transitive clause, its function in this language is to denote what an entity is thought or said to be. Besides, it can also express what an entity is changed into as a result of an action (Aikio & Ylikoski 2010: 57, 70). In Korean, the element *lago* 라고 is used to emphasize the preceding content. In this sense, it can be translated as ‘is said/called/thought as’.

In some other languages, although there is no unique form for ASCRIPTION, as in North Saami and Korean, the interrogative is inflected to the case demanded by the motion ‘consider/said to be’. In the translation in German, the verb *halten* is combined with the preposition *für* to express ‘consider, hold for’. As following *für*, *wer* ‘who’ is demanded to be changed into *wen* in the accusative case. Instead, Hungarian applies *kinek* in the dative case (Kenesei et al. 1998: 192).

4.4.4 PERSON.IDENTITY

The following three sub-clusters show a similar syntactical expression. In English, the copula *BE* serves as the predicate and links the question word and the subject. No verb denoting a specific action is found in these questions. The queried information is about someone’s identity. Therefore, these sub-clusters are labeled **PERSON.IDENTITY**.

The majority of sampled languages do not have a distinctive form for PERSON.IDENTITY. For this subdomain, the identical interrogative for AGENT is frequently used. However, some languages show special interrogative constructions in these contexts. The first type is to mark the basic interrogative form ‘who’ as the predicate in the clause, which then can be translated as ‘be who’. In Toro So Dogon, the basic form *aa* ‘who’ can be followed by the clitic *y*= ‘it is’[22] to express the predicate function (Heath 2017: 281). A similar strategy is also found in Hopi. In this language, *haki* ‘who’ serves as the predicate of the phrase, while the shortened form *hak* ‘who’ is applied specifically to denote the subject in a question (Kalectaca & Langacker 1978: 108). The second frequent way to distinguish the interrogative meaning ‘be who’ is the application of case marking. For instance, in Highland Popoluca, a Mixe-Zoque language spoken in Mexico, the basic interrogative *i̱*[23] takes the absolutive case marker *mi*= or

[22] In the Bible translation of this language used for this study, this clitic is written as *i*=.

[23] In the grammar the form is written as *ii*.

i= when the interrogative coding functions as the predicate in a non-verbal clause (de Jong Boudreault 2009: 189; Elson & Gutiérrez G. 1999: 159).

The main reason that the clustering generates three separate subgroups related to IDENTITY lies in the morphological markedness of person and number. In some languages, these two grammatical facets are overtly expressed with corresponding markers in interrogatives. As follows, contexts respectively belonging to these three sub-clusters and examples are presented.

4.4.4.1 PERSON.IDENTITY.2SG

Seven contexts all denote the second person in the singular and thus are classified into the same group, as given in (4.20). The following Table 4.23 provides five languages with a special marking.

(4.20) *eng-x-bible-common*

43021012	*None of the disciples could bring themselves to ask him, "**Who** are you?"*
43001022a	*They asked, "**Who** are you?"*
43008025a	*"**Who** are you?" they asked.*
44022008	*I answered, "**Who** are you , Lord?"*
59004012	*But you who judge your neighbor, **who** are you?*
43008053	*He died and the prophets died, so **who** do you make yourself out to be?*
43001021	*They asked him, "Then **who** are you? Are you Elijah?"*

	Burarra	**Gagauz**	**Garifuna**	**Makasar**	**Highland Popoluca**
43021012	*ny-yingay*	*kimsin*	*cátei*	*inaiki'*	*mi̲i̲apaap*
43001022a	*ny-yinga*	*kimsin*	*cátabuti*	*inaiko*	*mi̲i̲apaap*
43008025a	*ny-yingiya*	*kimsin*	*cátabuti*	*inaiko*	*mi̲i̲apaap*
AGENT	*ana-nga*	*kim*	*ca*	*inai*	*i̲*

Table 4.23: Examples for PERSON.IDENTITY.2SG

Burarra, a Maningrida language spoken in Australia, possesses a full system of prefixes indicating person and number. For this sub-cluster, *ny-* refers to the subject of the second person in the singular (Glasgow & Glasgow 2011). For the second person, Gagauz from the Turkic family combines the suffix *-sin* with the basic interrogative *kim* 'who (Schulze 2002: 784). Garifuna from the Arawakan family also uses the suffix *-bu* for the second person in these questions (Haurholm-Larsen 2016: 83). The enclitic *=ko/=ki'* in Makasar, an Austronesian language spoken in Indonesia, and the proclitic *mi=* in Highland Popoluca are not only indicative of the second person but also denote the absolutive case (Jukes 2016: 143; de Jong Boudreault 2009: 168, 189). This confirms the role of the questioned human referent as the subject of an intransitive verb.

4.4.4.2 PERSON.IDENTITY.3SG

This group contains four contexts in which the interrogatives are marked for the third person in the singular. Content and examples are given in (4.21) and Table 24, respectively.

(4.21) *eng-x-bible-common*

40021010 ***Who*** *is this?*

42009009a ***Who*** *am I hearing about?*

42007049 ***Who*** *is this person that even forgives sins?*

42005021a ***Who*** *is this who insults God?*

	Burarra	**Gagauz**	**Uyghur**
40021010	*ana-nga*	*kimdir*	*kimdu*
42009009a	*ana-nga*	*kim*	*kimdu*
42007049	*ana-nga*	*kimdir*	*kimdu*
AGENT	*ana-nga*	*kim*	*kim*

Table 4.24: Examples for PERSON.IDENTITY.3SG

In these contexts, Burarra utilizes the prefix *ana-* to mark the third person in the singular (Glasgow & Glasgow 2011). The suffix *-dir* in Gagauz and *-du* in Uyghur are both indicative of the third person (Schulze 2002: 784; Engesæth & Yakup & Dwyer 2009: 191).

4.4.4.3 PERSON.IDENTITY.PL

The last subgroups of the PERSON.IDENTITY domain is composed of three contexts, as given in (4.22). In these questions, the queried referents are in the plural. To be noticed, although the expression of 44019015 is the same as IDENTITY.2SG in English, the plurality of the referent can be inferred through contextual information and translations in other languages. Table 4.25 displays six sampled languages providing markers for plurality.

(4.22) *eng-x-bible-common*

40012048b ***Who*** *are my brothers?*

66007013a ***Who*** *are these people wearing white robes?*

44019015 *I know Jesus and I'm familiar with Paul , but* ***who*** *are you?*

	Burarra	**Nomaande**	**Rundi**	**Spanish**	**Icelandic**	**North Saami**
40012048b	*aburr-ngay*	*báányé*	*ba nde*	*quiénes*	*hverjir*	*geat*
66007013a	*aburr-ngiya*	*báányé*	*ba nde*	*quiénes*	*hverjir*	*geat*
44019015	*nyiburr-nga*	*báányé*	*ba nde*	*quiénes*	*hverjir*	*geat*
AGENT	*ana-nga*	*aányé*	*nde*	*quién*	*hver*	*gii*

Table 4.25: Examples for PERSON.IDENTITY.PL

For the first two questions, Burarra applies the prefix *aburr-* for the third person in the plural. The subtle difference of the question in 44019015 is marked in this language by the prefix *nyiburr-* indicating the excluded first or second person in the plural (Glasgow & Glasgow 2011). Nomaande, a Niger-Congo language spoken in Cameroon, places the marker *bá-* of the noun class for plurality before the question word *aányé* 'who' (Wilkendorf 1998: 12). Rundi, another language from the Niger-Congo family, uses a compositional structure *ba*

nde to mark 'who' in the plural (Cox 1975: 114). Spanish differentiates *quién* 'who (one person)' and *quiénes* 'who (more than one person)'. The latter is found in this sub-cluster. In Icelandic and North Saami, *hverjir* and *geat* are both in the nominative case in the plural (Neijmann 2001: 82; Aikio & Ylikoski 2010: 70).

4.4.5 PERSON.SELECTION

Five contexts are assigned to this sub-cluster, as given in (4.23) below.

(4.23) *eng-x-bible-common*

42022027	*So **which one** is greater, the one who is seated at the table or the one who serves at the table?*
42010036b	***Which one** of these three was a neighbor to the man who encountered thieves?*
42007042	*When they couldn't pay, the lender forgave the debts of them both . **Which** of them will love him more?*
40027021	*The governor said, "**Which** of the two do you want me to release to you?"*
40021031	***Which one** of these two did his father's will?*

In the contexts presented above, the speakers address questions about a person as well. Yet, different from AGENT discussed in §4.4.2.1, a set of possibilities is provided in the question and the addressee should pick one of these options as the answer. In other words, the answer to this kind of question is limited to a certain range. In English, as shown in (4.23), the interrogative is no longer the general person *who*, but the question word *which* or the combination *which one* explicitly referring to the meaning of SELECTION. Thus, this sub-cluster is tagged as **PERSON.SELECTION**.

In terms of the interrogative coding of PERSON.SELECTION, the structure in English is not attested in every language in the sample. Instead, there exists a mixture of different forms, which resembles the situation in sub-cluster PLACE.FROM.SOURCE described in §4.3.7. For

PERSON.SELECTION, sampled languages normally employ the interrogative from either the PERSON or SELECTION category. In the following, some main types will be illustrated.

The first type of coding is represented by languages from the sample that apply the general interrogative for the PERSON category to mark PERSON.SELECTION. It seems that these languages do not formally differentiate the information about PERSON.SELECTION from other general PERSON questions. In this category, there are two further situations. For some languages, the information about the category SELECTION is inquired with the interrogative for PERSON and THING according to the grammatical description. No form is specifically created to express SELECTION. In this case, it appears that the information about SELECTION is not explicitly marked, no matter whether the targeted object is animate or inanimate. Sampled languages belonging to this kind are, for instance, Totontepec Mixe and Western Arrarnta. In terms of Totontepec Mixe, a Mixe-Zoque language spoken in Mexico, the entry for *cuál* 'which' in the dictionary compiled by Schoenhals & Schoenhals (1965: 191) refers to *pa̱n* 'who' and *ti* 'what'. For Western Arrarnta, Strehlow (1942: 186) translates *nguna*/*ngula* as 'who, which, what' and *iwuna*/*iwula* as 'what, which', whereas he did no give an interrogative exclusively indicating 'which'. Besides, for some languages the description of questions about SELECTION is just unavailable in grammar, for example El Nayar Cora and Lowland Tarahumara, two Uto-Aztecan languages in the sample.

Another situation of the first type is that languages choose the general interrogative of the PERSON category for PERSON.SELECTION, even though there exists a specialized form indicating a general selection according to grammatical references. However, it is not clear whether the PERSON.SELECTION subdomain is grammatically not encoded in the form in those languages, or information about PERSON.SELECTION is beyond the valid scope of the form particularly referring to SELECTION. Also, the subjective decision of the translator can be one of the resulting factors. Table 4.26 below gives some examples of this case (Tran & Tran 2007: 31; Schachter & Otanes1983: 506; Emenanjo 2015: 390; Breedveld 1995: 486; Brandão 2014: 331).

The second type of coding for PERSON.SELECTION demonstrates the opposite of the last pattern. Like the interrogative *which* and *which one* in English used in (4.23), languages of this type all utilize the specialized form for category SELECTION in this sub-cluster. Table 4.27 shows a comparison between PERSON.SELECTION and PERSON.ROLE.AGENT in five exampled

languages (Miller 2017: 10; Woollams 1996: 225; Miyaoka 2013: 447, 451; Timyan 1977: 153; Jukes 2006: 352-354).

	Vietnamese	Tagalog	Igbo	Maasina Fulfulde	Parecís
420022027	*ai*	*sino*	*ònye*	*homo*	*xala*
42010036b	*ai*	*sino*	*ònye*	*homo*	*xala*
AGENT	*ai*	*sino*	*ònye*	*homo*	*xala*

Table 4.26: 'who'-type of PERSON.SELECTION

	Eastern Bru	Batak Karo	Central Yupik	Baoulé	Makasar
420022027	*aléq*	*apai*	*naliak*	*ônin*	*kereanga*
42010036b	*aléq*	*apai*	*naliak*	*ônin*	*kereanga*
AGENT	*noau*	*ise*	*kia*	*wan*	*inai*

Table 4.27: 'which'-type of PERSON.SELECTION

Unlike the strict separation of codings exhibited in the last two kinds, the demarcation between general PERSON and PERSON.SELECTION is not rigidly drawn in some other languages. Instead, the use of forms is quite flexible. The following Table 4.28 provides a glimpse of the mixed interrogatives in five languages (Heath 2017: 280, 285; Smith 2017: 602; Davies 2010: 88; Hagman 1977: 50-52; Lindblom 1914: 19-20).

	Toro So Dogon	Ma'anyan	Madurese	Nama	Tharaka
420022027	*aai*	*hie*	*sapa*	*tariba*	*n'ûû*
42010036b	*inɛ aai*	*sa awe*	*kemma*	*mâb*	*n'ûrîkû*
42007042	*yagɔ*	*hie*	*sapa*	*tarib*	*n'ûrîkû*
40027021	*inɛ aa*	*hie*	*kemma*	*mâb*	*n'ûû*
40021031	*yagɔi*	*saawe*	*kemma*	*mâba*	*n'ûrîkû*

Table 4.28: Mixed type of PERSON.SELECTION

Table 4.29 below summarizes the sampled languages of each coding type of PERSON.SELECTION.

Type	Language
'who'	Balantak, Burarra, Chuj, Chamorro, Cherokee, El Nayar Cora, Danish, Dogrib, Dii, Maasina Fulfulde, Paraguayan Guaraní, Gwich'in, Igbo, Jarai, Korean, Machiguenga, Dutch, Parecís, Yine, Tagalog, Vietnamese, Wolof
'which'	Baoulé, Eastern Bru, Batak Karo, Tabasco Chontal, Northern Dagara, English, North Alaskan Inupiatun, Central Yupik, Irish, Hopi, Inga, Karakalpak, Northern Kissi, Makasar, Tenharim-Parintintin-Diahoi, Noon, Turkmen, Turkish, Francisco León Zoque
mix	Acehnese, Balinese, Car Nicobarese, Catalan, Czech, Welsh, German, Toro So Dogon, Ejagham, Finnish, Gagauz, Kuku-Yalanji, Croatian, Hungarian, Iban, Indonesian, Icelandic, Japanese, Halh Mongolian, Nomaande, Madurese, Coatlán Mixe, Ma'anyan, Masaaba, Nama, Tetelcingo Nahuatl, Western Huasteca Nahuatl, Nyanja, Highland Popoluca, Huallaga Huánuco Quechua, Romanian, Rundi, North Saami, Spanish, Sirionó, Tharaka, Uyguhr, Yucatec Maya, Mandarin

Table 4.29: Coding types of PERSON.SELECTION

It is common that languages do not formally differentiate the meaning 'which person' from 'which thing'. Just like English, the question word *which* is apt for both meanings. However, it is not alway this case. Some languages mark the animateness in the interrogative of SELECTION. As an example, Tenharim-Parintintin-Diahoi uses the question word *manamo* particularly meaning 'which person' (Pease 1968: 68). In contrast, this language has the interrogative *ma'g̃a* 'who' and *ma/mahã* that serves as a general question word meaning 'what/which' (Betts 2012: 158-159). Noon, a Niger-Congo language spoken in Senegal, displays a similar situation in which the prefix *y-*, an agreement marker for the animate singular, is combined with the selective interrogative stem *-iida* to mark 'which person' (Soukka 1999: 83, 136). Thus, the whole construction *yiida* attested in all five

contexts is indicative of a selection of an animate object in singular. On the contrary, the interrogative *wiida* with the prefix *w-* indicating an inanimate object is applied for 'which thing'.

Apart from the hint related to animateness, forms in some languages reveal the scale of options. That is, a choice made from two alternatives is inquired with a different interrogative from a selection with more than two options. See the two sampled languages in Table 4.30 below.

	North Alaskan Inupiatun	**Japanese**
420022027	*nalliak*	どちら
42010036b	*nalliat*	だれ
42007042	*nalliakta*	どちら
40027021	*nalliak*	どちら
40021031	*nalliakta*	どちら

Table 4.30: Special forms for PERSON.SELECTION

North Alaskan Inupiatun and Japanese are two examples that distinguish a selection made between two and many alternatives. North Alaskan Inupiatun adopts *nalliak* for two options and *nalliat* meaning 'which one of these (all)', respectively (Seiler 2012: 126). Similarly, in Japanese *dochira* どちら is used when there are only two options, whereas *dore* どれ indicates more than two possibilities (Bunt 2003: 229-230). Yet, interestingly, in this translation *dare* だれ 'who' is applied in the context for more than two options. With the evidence given in these two languages, it can be inferred that there exists a slight semantic difference between 42010036b and other contexts. That is, the question in 42010036b denotes a selection made among more than two options, while there are only two possibilities in the other four contexts. In this sense, the context in 42010036b represents the semantic subdomain **PERSON.SELECTION.MULTIPLE**, while other contexts in this sub-cluster take the tag **PERSON.SELECTION.TWO**.

4.4.6 PERSON.POSSESSOR

The sub-cluster presented in this section comprises six questions. The corresponding contexts are given in (4.24).

(4.24) *eng-x-bible-common*

42020024a	***Whose** image (and inscription does it have on it)?*
40022020a	***Whose** image (and inscription is this)?*
40022042b	***Whose** son is he?*
42011019	*If I throw out demons by the authority of Beelzebul, then by **whose** authority do your followers throw them out?*
42020033	*In the resurrection, **whose** wife will she be?*
44004007c	*or in **what** name did you do this?*

The contexts shown in (4.24) are questions about the belonging of an inanimate entity or the relationship between people. In English, the question word *whose* occurring in the first five contexts is morphologically connected to the general personal form *who* and syntactically precedes the noun. It is exclusively used to ask for the possession and substitutes the place of the possessor in the question. In this case, the label **PERSON.POSSESSOR** is assigned to this sub-cluster.

	German	Finnish	Icelandic	Romanian	North Saami
42020024a	*wessen*	*kenen*	*hvers*	*cui*	*gean*
40022042b	*wessen*	*kenen*	*hvers*	*cui*	*gean*
AGENT	*wer*	*kuka*	*hver*	*cine*	*gii*

Table 4.31: Inflection for PERSON.POSSESSOR

To enquire about information about POSSESSOR, languages have various strategies to build the interrogative construction. The first common type is represented by languages with a declension mechanism. In these languages, the basic interrogative meaning ‘who’ is inflected to the possessive or genitive case to refer to POSSESSOR. Table 4.31 above provides examples

from five languages of this kind with a comparison with the form used for AGENT (Karlsson 2008: 207; Neijmann 2001: 82; Sarlin 2014: 165; Aikio & Ylikoski 2010: 70)

As can be seen in Table 4.31, it is difficult to morphologically decompose the inflected structure for POSSESSOR into smaller elements. Apart from this form, possession can be marked by adding the possessive or genitive suffix to the general form meaning 'who' in some languages. Table 4.32 below shows three examples of this type. The corresponding suffixes are marked in bold (van Schaaik 2020: 64; Parker 1969: 40; Strehlow 1942: 187).

	Turkish	**Ayacucho Quechua**	**Western Arrarnta**
42020024a	*kim**in***	*pi**pa***	*ngunha**ka***
40022042b	*kim**in***	*pi**pa***	*ngunha**ka***
PERSON.AGENT	*kim*	*pitaq*	*ngunhalama*

Table 4.32: Possessive suffixes for PERSON.POSSESSOR

Possessor can be also expressed via the combination of the interrogative and a pre- or postposition. In this case, the construction is normally literally translated as 'of whom'. The following Table 4.33 gives three sampled languages exhibiting this morphological method (Wheeler & Yates & Dols 1999: 108; Tran & Tran 2007: 39; Hoppmann 2011: 216). The corresponding preposition is marked in bold. In Korean, the postposition *ui* 의 to mark the genitive or possessive is used.

	Catalan	**Vietnamese**	**Korean**
42020024a	***de** qui*	***của** ai*	누구**의**
40022042b	***de** qui*	***của** ai*	누구**의**
PERSON.AGENT	*qui*	*ai*	누구

Table 4.33: Prepositional construction for PERSON.POSSESSOR

In Mandarin and Japanese, no morphological change is conducted in the basic interrogative. Rather, possession is expressed by means of linking the question word meaning 'who' to the noun with a possessive particle. In questions classified into this sub-cluster, the

structure *shuíde* 誰的 and *dareno* だれの are attested respectively in Mandarin and Japanese. In these two forms, the element *de* 的 and *no* の function as the possessive particle (Bunt 2003: 226).

In some languages, the linking between the interrogative and the noun is realized through a possessive pronoun. In this way, the belonging of the object is inquired about. Three examples are given in Table 4.34.

	Noon	**Dii**	**Toro So Dogon**
42020024a	*cuuɓa*	*nóo woolá*	*aa mɔi*
42020033	*yuu ɓa*	*nóo woo*	*aa mɔi*
PERSON.AGENT	*ɓa*	*nóo*	*inɛ aa*

Table 4.34: Possessive pronominal construction for PERSON.POSSESSOR

Noon applies the construction *-uu ɓa* to ask for the possessor. In this structure, *-uu* is a stem to form a so-called appropriative pronoun. Its function is to replace the head noun in a genitive construction. With a corresponding agreement marker preceding this stem, e.g., *c-* for the first person in plural and *y-* for an animate object in singular, as can be seen in Table 4.34, the appropriative pronoun is then combined with the basic person interrogative *ɓa* in contexts related to possession (Soukka 1999: 140, 159).

In Dii, a Niger-Congo language spoken in Cameroon, the personal interrogative *nóo* is placed before the third person possessive root *woo* meaning 'its' to denote the animate possession (Bohnhoff 2010: 92, 312). In a similar way, the possessor of an alienable noun is formulated via the postnominal *mɔ* in Toro So Dogon. To ask for the possessor in this sub-cluster, *mɔ* is first combined with *-i* 'it is' and then follows the question word *aa* 'who' (Heath 2017: 114-115, 280).

In the types discussed before, it is easy to identify the morphological or etymological closeness between the forms for POSSESSOR and other general PERSON questions. However, in Czech and Croatian presented in Table 4.35 below, such a formal relationship is not crystal and no available information has been found in grammar so far (Naughton 2005: 101; Browne & Alt 2004: 58).

	Czech	Croatian
42020024a	*čí*	*čiji*
40022042b	*čí*	*čiji*
AGENT	*kdo*	*tko*

Table 4.35: Special forms for PERSON.POSSESSOR

4.4.7 PERSON.KIND

This group incorporates three contexts, as given in (4.25) below.

(4.25) *eng-x-bible-common*

46003005b ***What** is Paul?*

46003005a *After all, **what** is Apollos?*

40008027 ***What kind of** person is this? Even the winds and the lake obey him!*

It is conspicuous that, instead of the question word *who*, English uses *what* or *what kind of* in these three contexts, although the subject of the question is also a human being. This implies a semantic distinction existing between this group and other sub-clusters belonging to the PERSON category. When we read the surrounding text, it can be deduced that these three questions are not asking for someone's name or personal identity, but the social status or type. For instance, 46003005a and 46003005b are sequent in text and the answer to them is *They are servants who helped you to believe* in the translation *eng-x-bible-common*. For 40008027, no answer is given in subsequent verses. According to the contextual information, the content is about people who were so amazed by the control of Jesus over winds and lakes that they then addressed such a question. In this sense, it can be seen as rhetorical. However, this question targets at the performance or characteristic of the human referent. Given the choice of interrogatives in English, this sub-cluster is named **PERSON.KIND**. The difference between PERSON.KIND and PERSON.SELECTION is that the former refers to the 'type' of someone, whereas the latter designates a specific human referent.

With regard to the coding for PERSON.KIND, it also shows a mixture of interrogatives from different categories. Some languages, e.g., Noon, Dutch, Gagauz and Jarai, do not mark the

domain of PERSON.KIND in form and just apply the general interrogative of PERSON in this sub-cluster. Table 4.36 below provides a comparison of these languages (Soukka 1999: 155; Donaldson 2008: 103; Ulutaş 2014: 88; Jensen 2014: 94).

	Noon	**Dutch**	**Gagauz**	**Jarai**
46003005b	*ɓa*	*wie*	*kim*	*hlơi*
46003005a	*ɓa*	*wie*	*kim*	*hlơi*
40008027	*ɓa*	*wie*	*kimdir*	*hlơi*
AGENT	*ɓa*	*wie*	*kim*	*hlơi*

Table 4.36: 'who' for PERSON.KIND

Conversely, there are also a number of sampled languages utilizing the interrogative of THING to denote PERSON.SELECTION. Four examples are presented in Table 4.37 (Engel et al. 1987: 351; Emenanjo 2015: 390; Timyan 1977: 113).

	Mandarin	**Francisco Leon Zoque**	**Igbo**	**Baoulé**
46003005b	什麼	*tiyø*	*gini*	*nzu*
46003005a	什麼	*tiyø*	*gini*	*nzu*
40008027	什麼	*tiyø*	*gini*	*nzu*
AGENT	誰	*i'is*	*ònye*	*wan*

Table 4.37: 'what' for PERSON.KIND

Remarkably, in the context of 40008027 a special structure meaning 'what kind of' is attested in some languages, just like English presented in (4.25) above. See examples shown in the following Table 4.38. In German, Ejagham from the Niger-Congo family and Romanian, such a structure is composed of multiple elements. The construction *was für* in German comprises *was* 'what' and the preposition *für* 'for' to inquire about a choice or a type without a fixed set of options. Ejagham combines *bha(aghé)* 'which' and the noun *ekpak* 'kind/type' for this question (Watters 1981: 340). The composition of *ce fel de* in Romanian is similar to *what kind of* in English (Sarlin 2014: 167). Besides, Spanish in the sample also provides the form *qué clase de* with the same compositional pattern in this context.

	German	Ejagham	Romanian	Khalkha	Burarra
46003005b	*was*	*énê*	*cine*	*hen*	*ngu-ngiya*
46003005a	*was*	*énê*	*cine*	*hen*	*an-ngiya*
40008027	*was für*	*bhaghé ekpak*	*ce fel de*	*yamar*	*an-guyinmiya*

Table 4.38: 'what kind of' for PERSON.KIND

In contrast, a morphologically unanalyzable element is found in Khalkha, an Altaic language spoken in Mongalia, and Burarra for questions of PERSON.KIND. According to Janhunen (2003: 111), the question word *yamar* refers particularly to 'what kind of' in Khalkha, while this language adopts *ali* 'which' for SELECTION and *hen* 'who' for PERSON. Burarra owns an interrogative stem *-guyinmiya* exclusively meaning 'what kind' before which a suitable agreement marker should be added (Glasgow & Glasgow 2011).

4.4.8 Summary

In this chapter, the cluster PERSON and its internal classification are discussed. According to the semantic role that the referent plays in the clause, four sub-clusters are identified, i.e., PERSON.ROLE.AGENT, PERSON.ROLE.PATIENT, PERSON.ROLE.RECIPIENT and PERSON.ROLE.GOAL. They are subsumed under PERSON.ROLE presented in §4.4.2. A special subgroup PERSON.ASCRIPTION is then described in §4.4.3. For the belonging contexts, two languages have a unique form to query how the human referent is thought or said by others.

Three sub-clusters with contexts asking for the identity are discussed in §4.4.4, i.e., PERSON.IDENTITY.2SG, PERSON.IDENTITY.3SG and PERSON.IDENTITY.PL. The shared trait of these sub-clusters is that no verb denoting a concrete action is found as the predicate in the clause. Different markedness for person and number leads to the grouping of these three sub-clusters.

The sub-cluster PERSON.SELECTION is elaborated in §4.4.5. For questions of this sub-cluster, a limited set of choices is provided from which the answer is expected to be given. In terms of two and multiple options, some languages differently mark the interrogatives. Thus, two finer subsets are manually identified, i.e., PERSON.SELECTION TWO and PERSON.SELECTION.MULTIPLE.

The sub-cluster PERSON.POSSESSOR is presented in §4.4.6. As the label indicates, contexts of this subgroup inquire about the belongingness between the human referent and the other person or object. The last sub-cluster PERSON.KIND is illustrated in §4.4.7. In the corresponding contexts, the social status or type of the referent is queried.

4.5 Cluster of THING

In this chapter, the cluster containing contexts of the category THING will be presented. In §4.5.1, an overview of this cluster will be provided. Similar to the PERSON cluster, the THING cluster also has a sophisticated internal structure. The identified sub-clusters will be described in sections as follows:

- §4.5.2 — THING.PATIENT
- §4.5.3 — THING.THINK
- §4.5.4 — THING.DO
- §4.5.5 — THING.SAY
- §4.5.6 — THING.HAPPEN
- §4.5.7 — THING.SELECTION
- §4.5.8 — THING.KIND
- §4.5.9 — QUANTITY.MASS

4.5.1 Overview

The fourth primary cluster is substantially bigger than the last three and incorporates 125 contexts in total. Yet, according to Figure 4.1 in §4.1, the average silhouette width of this cluster is 0.28, which is the lowest compared to the other five groups. The following Table 4.39 presents some selected contexts as a sketch. Considering that most attested interrogatives refer to substance, the label **THING** is given to this cluster.

The following Figure 4.9 exhibits the internal structure of the THING cluster. Similar to the last cluster PERSON, data points in Figure 4.9 are primarily located on the left. The distribution within the dots on the left has not revealed any clear subgrouping yet. A few green triangles are far from the majority and stay at the right-bottom of the graph. They represent the contexts asked with the question word *which* in English. At the center bottom, two contexts are

encoded with the question word *how much* in English. Furthermore, three blue diagonal crosses standing for the construction *what kind of* and four green crosses for *why* are blended with the symbols for *what* at the top of the graph.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	44021033b	0.46504	*what*	*was*	什麼
2	41006024	0.46173	*what*	*worum*	什麼
3	41010051	0.45962	*what*	*was*	什麼
4	40020032	0.45391	*what*	*was*	什麼
5	40020021	0.44628	*what*	*was*	什麼
6	59002016b	0.39441	*what*	*welchem*	什麼
7	46004021	0.24794	*which*	*was*	什麼
8	42001066	0.23147	*what*	*was*	怎樣
9	41004030a	0.21599	*what*	*womit*	什麼
10	42004036	0.17902	*what kind of*	*was für*	怎麼
11	41012028	0.04514	*which*	*welches*	哪一
12	44010021	0.04141	*why*	*was*	什麼

Table 4.39: Verse selection of cluster THING

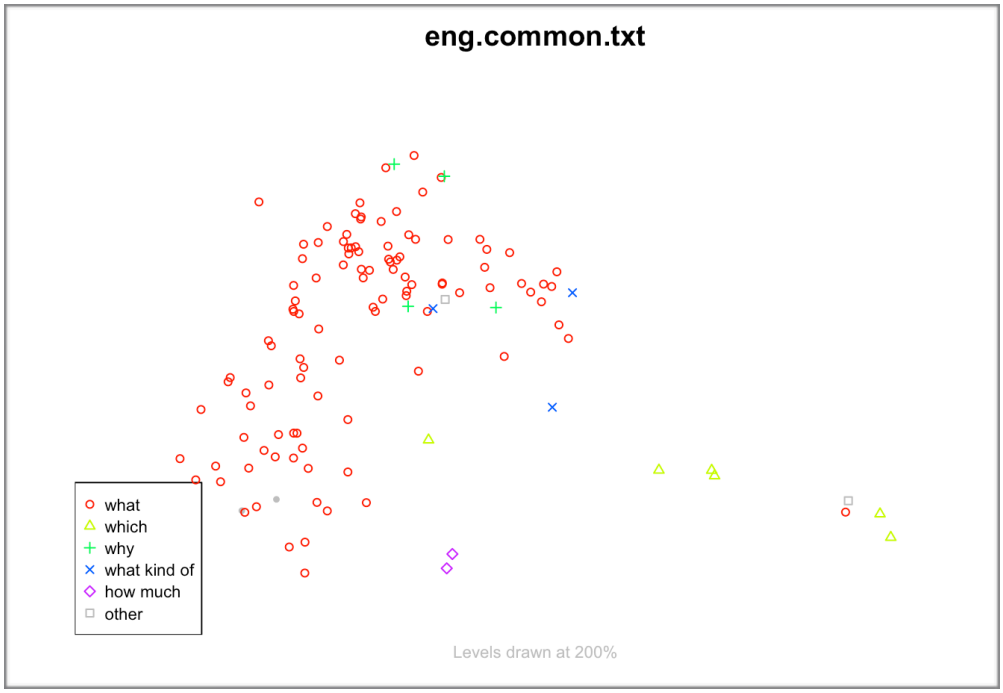

Figure 4.9: MDS plot of cluster THING

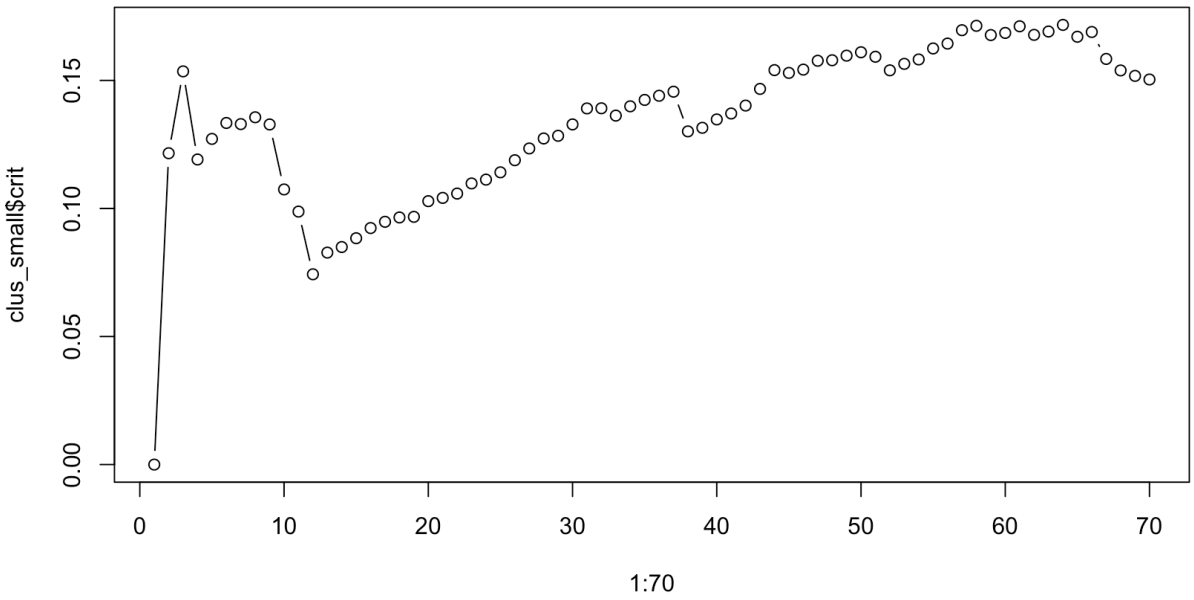

Figure 4.10: Suggested sub-clusters of THING

Figure 4.10 above presents the outcome of the second level of clustering of THING. The first interesting grouping is found at three sub-clusters with 68, 50, and 7 contexts, respectively. For the subgroup with seven contexts, the question word *which* is recurrently used in the English translation, which tells the connection to the concept of SELECTION (see §4.5.7 THING.SELECTION).

As the next optimal result, 125 contexts are assigned into eight subgroups. Six of them can be clearly interpreted. They are subgroups, respectively, with eight questions querying *what kind of* or *what* (see §4.5.8 THING.KIND), four questions with the similar content about *what sign*, two questions encoded with *how much* (see §4.5.9 QUANTITY.MASS), and ten questions related to the action *think* or *see*. Besides, the before-mentioned subgroup referring to SELECTION is further subdivided. One is specifically indicative of a selection made from two options, while the other one is about a selection with an unlimited set of possibilities.

Yet, there are still two huge subgroups with 58 and 36 contexts whose meanings are still unidentifiable. Thus, adopting the same practice for PERSON, clusterings in Figure 4.10 with a high number of results, e.g., 37 and 64 sub-clusters, will be also taken into consideration. The findings will be discussed in §4.5.2 to §4.5.6.

4.5.2 THING.PATIENT

In the grouping with 37 sub-clusters, 24 contexts are classified as one group. Some of them with the highest silhouette width are displayed below in (4.26) as examples.

(4.26) *eng-x-bible-common*

41006024	***What** should I ask for?*
40020021	***What** do you want?*
43001038a	***What** are you looking for?*
41010036	***What** do you want me to do for you?*
40006031b	***What** are we going to drink?*
40019020	*The young man replied, "I've kept all these. **What** am I still missing?"*
40006031a	*Therefore, don't worry and say, "**What** are we going to eat?"*
40019027	***What** will we have?*

The speakers of these questions intend to obtain an object or an event that is influenced by the action performed by the agent and denoted by the predicate. Syntactically, the inquired referent serves as the object of a transitive verb. From the perspective of the semantic role, the question word, i.e., *what* in English, substitutes the position of the patient. Therefore, this sub-cluster is named **THING.PATIENT**.

In terms of the coding, the majority of the sampled languages apply a basic or general form meaning 'what'. In some other languages, the patient role is reflected by the special marking on the basic form or inflection. The following Table 4.40 lists five examples.

	North Alaskan Inupiatun	Central Yupik	Hungarian	North Saami	Hopi
41006024	*sumik*	*camek*	*mit*	*maid*	*hihta*
40020021	*sumik*	*camek*	*mit*	*maid*	*hihta*
41010036	*sumik*	*camek*	*mit*	*maid*	*hihta*
40006031a	*sumik*	*camek*	*mit*	*maid*	*hihta*

Table 4.40: Examples for THING.PATIENT

North Alaskan Inupiatun and Central Yupik demonstrate the strategy of adopting a suffix to mark the patient role. In the former language, the basic interrogative stem for the THING category is *su-*. According to Seiler (2012: 18, 200), the function of the singular suffix *-mik* is to denote the so-called modalis case that signifies an instrument or a non-specific object of a transitive verb. Central Yupik displays a similar pattern. On the basis of the interrogative stem *ca-* 'what', the singular suffix *-mek* is used to mark the ablative-modalis case. This case is used to designate a component conducting the patient function or theme function in the clause (Miyaoka 2012: 596, 629).

In Hungarian, *mit* is inflected from the basic question word *mi* 'what' for the accusative case (Rounds 2001: 93). The form *maid* in North Saami marks the genitive-accusative case of the base *mii* 'what' (Aikio & Ylikoski 2010: 43, 70). The last example is found in Hopi. This language formally differentiates complements in a clause as well as their number. The form *hihta* attested in THING.PATIENT refers to the object of a transitive verb in both singular and plural (Kalectaca & Langacker 1978: 109).

4.5.3 THING.THINK

This sub-cluster is composed of six contexts, as given in (4.27) below.

(4.27) *eng-x-bible-common*

40022042a	***What** do you think about the Christ?*
40026066	*"**What** do you think?" And they answered, "He deserves to die!"*
40021028	*"**What** do you think?"*
43011056	***What** do you think? He won't come to the festival, will he?*
40017025a	*But when they came into the house, Jesus spoke to Peter first. "**What** do you think , Simon?"*
40018012	***What** do you think?*

The content of these six questions is highly similar. The intention of the questioners is to ask for addressees' opinion or view about someone or something. In English, the predicate of all contexts is the action *think*. Thus, the label **THING.THINK** is assigned to this subgroup.

In many sampled languages, e.g., English and German, no specialized marking is found in interrogatives. The general form for 'what' is used for these contexts. However, it is noticeable that 40 sampled languages employ the interrogative meaning 'how' for some or all of these six questions. Table 4.41 below provides examples in five languages and gives a comparison to the form meaning 'what'.

	Balantak	**Tabasco Chontal**	**Nama**	**Mandarin**	**Korean**
40022042a	*koi upa*	*cache'da*	*mati*	怎樣	어떻게
40026066	*koi upa*	*cache'*	*mati*	怎樣	어떻게
40021028	*koi upa*	*cache'da*	*mati*	怎樣	어떻게
43011056	*koi upa*	*cache'*	*mati*	怎樣	어떻게
'what'	*upa*	*cua'*	*tare*	甚麼	무슨

Table 4.41: 'how' for THING.THINK

That nearly half of the sampled languages apply the interrogative meaning 'how' could be the reason for the emergence of the sub-cluster THING.THINK. The queried target of THING.THINK , i.e., an opinion, is not a concrete object that is the typical referent of the category THING. Rather, an opinion is abstract and invisible. For this kind of inquiry, a cross-cut usage of interrogatives meaning 'what' and 'how' is allowed. The exchange of these two forms scarcely makes an impact on the interpretation of the question. For instance, the understanding of the expression *what do you think* is not significantly different from *how do you think* in English. This might result from the questioned target being an action. This type of question can be comprehended as acquiring the 'kind' of action, or the 'manner' of action, i.e., how an action is conducted. Correspondingly, the interrogative form is often parallel to the one for MANNER. This reflects a semantic overlap between the category THING and MANNER.

4.5.4 THING.DO

Four contexts are assigned to this sub-cluster and they are given in (4.28). These four questions also display the similar content. The speakers are not asking for a concrete substance but a suggestion about the subsequent action they are expected to perform. In

English, the combination of the modal verb *should* and the general verb *do* functions as the predicate to denote the yet undefined action. In this regard, I name this sub-cluster **THING.DO**.

(4.28) *eng-x-bible-common*

42003010	*The crowds asked him, "**What** then should we do?"*
42003012	*Teacher, **what** should we do?*
42003014b	*"**What** should we do?" He answered, "Don't cheat or harass anyone, and be satisfied with your pay."*
44002037	*They said to Peter and the other apostles, "Brothers, **what** should we do?"*

Despite that the basic form meaning 'what' is attested in this sub-cluster in most sampled languages, a specialized verbal construction is found in three languages shown in Table 4.42 below.

	Gwich'in	**Khalkha**	**Burarra**
42003010	*deegwehee'yaa*	*yuu*	*nyiburr-yinmiya*
42003012	*deegwehee'yaa*	*yaah*	*nyiburr-yinmiya*
42003014b	*deegwehee'yaa*	*yaah*	*nyiburr-yinmiya*
44002037	*nats'ahts'ą'*	*yaah*	*nyiburr-yinmiya*

Table 4.42: Special forms for THING.DO

In Gwich'in, verbs in the dictionary are presented without internal analysis. In the first three questions of THING.DO, the attested form *deegwehee'yaa* refers to 'what are we going to do' (Alexander & Alexander 2011: 47). In 44002037, the interrogative *nats'ahts'ą'* 'how' is used. Khalkha has an interrogative verb *yaa-* 'to do what' (Janhunen 2012: 255). However, an explanation for the suffix *-h* is not found in the grammar. In Burarra, the interrogative verb *-yinmiya* 'do how' is employed (Glasgow & Glasgow 2011). This verb always requires a prefix. For these questions, *nyiburr-* indicating the excluded first or second person in the plural is attached to *-yinmiya*.

As discussed in §4.5.3 THING.THINK, it is possible that the interrogatives for 'what' and 'how' are both applicable when the inquired object refers to the abstract existence. Regarding the referent of THING.DO, i.e., an activity to be carried out, it seems that this pattern is also suited for this sub-cluster. For instance, the interrogative verb *-yinmiya* just mentioned in Burarra is translated as *do how* in English. Such a translation does not distort the purpose of the questioner. As an evidence of this usage, 14 sampled languages are found utilizing the interrogative meaning 'how' for THING.DO. The following Table 4.43 provides five of them as an example (Casad 1984: 198; Bohnhoff 2010: 312; Watters 1981: 335; Petzell 2008: 152; Wilkendorf 1998: 26).

	El Nayar Cora	**Dii**	**Ejagham**	**Kagulu**	**Nomaande**
42003010	*a'ini*	*nɛ́nná*	*nan*	*nhani*	*anyána*
42003012	*a'ini*	*nɛ́nná*	*nan*	*nhani*	*anyána*
42003014b	*a'ini*	*nɛ́nná*	*nan*	*nhani*	*anyána*
44002037	*a'ini*	*nɛ́nná*	*nan*	*nhani*	*anyána*
'what'	*ti'itaj*	*ɛ́ná*	*jen*	*choni*	*aátɛ́*

Table 4.43: 'how' for THING.DO

4.5.5 THING.SAY

This sub-cluster contains five contexts in total. The content is illustrated in (4.29) below.

(4.29) *eng-x-bible-common*

45009030 *So **what** are we going to say? Gentiles who weren't striving for righteousness achieved righteousness, the righteousness that comes from faith.*

45008031a *So **what** are we going to say about these things?*

43012027 *Now I am deeply troubled. **What** should I say?*

46011022 ***What** can I say to you?*

45006001 *So **what** are we going to say?*

The goal of these five questions is to ask for the content that someone is supposed to formulate verbally. The predicate is denoted by the verb *say* in English. Thus, this sub-cluster is dubbed **THING.SAY**.

Similar to the last two sub-clusters, i.e., THING.THINK and THING.DO, the referent of THING.SAY is not indicative of a concretely existing substance but an utterance. Accordingly, the use of the interrogative meaning 'how' can be expected, since the question targets at the action *say* and can be perceived as asking for the manner of this action. Among the samples languages, ten of them demonstrate such a usage. Table 4.44 below exhibits five examples (Ma 2012: 46; Woollams 1996: 225; Braine 1970: 204; Omar 1969: 201; Hanson 2010: 321).

	Parauk	**Batak Karo**	**Car Nicobarese**	**Iban**	**Yine**
45009030	*ka mawx*	*kuga*	*sitih*	*baka ni*	*gi*
45008031a	*ka mawx*	*kuga*	*sitih*	*baka ni*	*gi*
43012027	*ka mawx*	*kuga*	*sitih*	NA	*gi*
46011022	*ka mawx*	*kuga*	*sitih*	*baka ni*	*gi*
'what'	*patix*	*kai*	*asuh*	*nama*	*klu*

Table 4.44: 'how' for THING.SAY

Apart from the occurrence of the construction 'how', three languages mark THING.SAY with a unique interrogative verb or element, as given in the following Table 4.45.

	Gwich'in	**Northern Dagara**	**Kagulu**
45009030	*deegweheenjyaa*	*bo*	*chigambeki*
45008031a	*deegweheenjyaa*	*bvnv*	*chigambeki*
43012027	*deehihjyaa*	*bo*	*nigambeki*
46011022	*deenahaihjyaa*	*bvnv*	*nhani*

Table 4.45: Special forms for THING.SAY

For THING.SAY, a specialized verbal compound *deegweheenjyaa* is again found in Gwich'in. Exactly this form is not provided in the dictionary. But, two very similar entries are found. According to Alexander & Alexander (2011: 46), *deegeheenjyaa* and *deegiheenjyaa*

respectively refer to ‘what have they to say’ and ‘what are they going to say’. Given the morphological resemblance, it is extrapolated that *deegweheenjyaa* used in this sub-cluster is also related to the action *say*. The different part might be the marking of person or tense.

In Northern Dagara, *bo* and *bvnv* are a special pair of interrogatives. They are translated as ‘what’ in the grammar, but particularly indicate an utterance, i.e., questions like *what do I say?*, or an idea, i.e., questions like *what does this mean?* (Mwinlaaru 2017: 155-156). Finally, Kagulu combines the element *gamb-* ‘speak’ with the interrogative stem *-ki* ‘what’ to denote the meaning of ‘say what’ (Petzell 2008: 91). The prefix *chi-* appearing in the structure is an agreement marker indicating a non-human thing.

4.5.6 THING.HAPPEN

Two contexts consists this small subgroup and are given in (4.30).

(4.30) *eng-x-bible-common*

42018036	*When the man heard the crowd passing by, he asked **what** was happening.*
42015026	*He called one of the servants and asked **what** was going on.*

The speakers of these two questions are asking the addressee to describe the situation currently taking place. According to the verbs used in these contexts, this sub-cluster is assigned the label **THING.HAPPEN**. In terms of the interrogative coding, most languages do not exhibit any notable marking for this semantic domain. The general interrogative meaning ‘what’ is applied.

Nonetheless, we can still find a particular interrogative verb in two sampled languages, as shown in Table 4.46 below. In Tenharim-Parintintin-Diahoi, the construction *maraname* refers to ‘how is it’ (Betts 2012: 162). Gwich’in again provides a verbal interrogative expression *deegwii’in*. According to Alexander & Alexander (2011: 47), this form should be translated as ‘what goes on there’.

	Tenharim-Parintintin-Diahoi	Gwich'in
42018036	*maraname*	*deegwii'in*
42015026	*maraname*	*deegwii'in*

Table 4.46: Special forms for THING.HAPPEN

4.5.7 THING.SELECTION

As the result of the clustering with three subgroups, seven questions indicating the category SELECTION are classified together. In a more detailed clustering, they are again separated into two subgroups. According to the contextual information and the attested interrogative codings, it can be concluded that the first one with four contexts refers to a selection with multiple alternatives, which is given in (4.31) below. In the other three contexts, as shown in (4.32), two options have been explicitly provided. This tells that these three questions are about a choice made from two possibilities. Given these semantic features, these two sub-clusters are respectively labeled **THING.SELECTION.MULTIPLE** and **THING.SELECTION.TWO**.

(4.31) THING.SELECTION.MULTIPLE

40019018 ***Which ones**? Then Jesus said, "Don't commit murder. Don't commit adultery. Don't steal. Don't give false testimony."*

40022036 *Teacher, **what** is the greatest commandment in the Law?*

43010032 *For **which** of those works do you stone me?*

41012028 ***Which** commandment is the most important of all?*

(4.32) THING.SELECTION.TWO

41002009 ***Which** is easier — to say to a paralyzed person, 'Your sins are forgiven,' or to say, 'Get up, take up your bed, and walk'?*

40023019 ***Which** is greater, the gift or the altar that makes the gift holy?*

40023017 ***Which** is greater, the gold or the temple that makes the gold holy?*

In terms of the use of interrogatives, some sampled languages have different ways to encode these two subdomains. To distinguish a pair of options from a set of many possibilities, the languages in the following Table 4.47 build two separate forms.

		Japanese	North Alaskan Inupiatun	Icelandic	North Saami	Finnish
MULTIPLE	**40019018**	どの	*nalliŋich*	*hver*	*guđiid*	*mitä*
	40022036	どの	*nalliat*	*hvert*	*guhtemuš*	*mikä*
	43010032	どの	*nalliatigun*	*hvert*	*man*	*mikä*
	41012028	どれ	*nalliat*	*hvert*	*guhtemuš*	*mikä*
TWO	**41002009**	どちら	*nalliak*	*hvort*	*goabbá*	*kumpi*
	40023019	どちら	*nalliak*	*hvort*	*goabbá*	*kumpi*
	40023017	どちら	*nalliak*	*hvort*	*goabbá*	*kumpi*

Table 4.47: Codings for THING.SELECTION.MULTIPLE and THING.SELECTION.TWO

As presented in §4.4.5 PERSON.SELECTION, Japanese and North Alaskan Inupiatun have different interrogative forms for SELECTION.TWO, i.e., *dochira* どちら and *nalliak*, and SELECTION.MULTIPLE, i.e., *dore* どれ and *nalliat-*. The same application is also found in THING.SELECTION, as shown in Table 4.47. If the interrogative meaning 'which' functions like an adjective and is followed by a noun in Japanese, then the form *dono* どの should be used (Bunt 2003: 229), as can be seen in the first three contexts of SELECTION.MULTIPLE.

In Icelandic, *hvort* is the neuter form of *hvor* 'which' and is applied when there are only two options, whereas *hver* 'which/who' appears with more possibilities (Neijmann 2001: 285). In North Saami, *goabbá* is indicative of two options, whereas *guhtemuš* is translated as 'which of many' (Aikio & Ylikoski 2010: 104). The question word *guđiid* found in 40019018 is the declined form of *guhtemuš* in the accusative or genitive case, while *man* is inflected from *mii* 'what' for the genitive case (Aikio & Ylikoski 2010: 70). Finnish uses *kumpi* specifically to ask for a choice with two options, whereas *mikä* is generally indicative of 'what/which' (Karlsson 2008: 207). The form *mitä* appearing in 40019018 is inflected from *mikä* for the partitive case.

These examples demonstrate a formally marked semantic distinction between a selection made from two versus multiple choices. Yet, it is also possible that the speaker expects an answer in the plural or consisting of more than one option, as shown by the sampled languages in the following Table 4.48. In these languages, the interrogative context in 40019018 is noticeably marked with a form that is different from the one applied in other questions of THING.SELECTION. This form is neither identical to SELECTION.TWO nor SELECTION.MULTIPLE. According to the grammars of the exampled languages listed in Table 4.48 below, such a form signifies the plurality of the referents. That is, the questioner is already aware of the multiplicity of the inquired answer of 40019018. As can be seen in (4.31) above, the construction *which ones* in English also reveals the plural feature of the expected answer. In this sense, the context in 40019018 reflects a semantic subdomain subsumed under THING.SELECTION.MULTIPLE. Given this characteristic, this contexts is separated and named **THING.SELECTION.MULTIPLE.PL**. Correspondingly, the other three contexts of THING.SELECTION.MULTIPLE can be further interpreted as **THING.SELECTION.MULTIPLE.SG**.

		Northern Dagara	Welsh	Hungarian	Tetelcingo Nahuatl	Spanish	Turkish
MULTIPLE.PL	**40019018**	*abobe*	*pa rai*	*melyeket*	*cötlejuanu*	*cuáles*	*hangilerine*
	40022036	*buor*	*pa un*	*melyik*	*cötlaja*	*cuál*	*hangisidir*
MULTIPLE.SG	**43010032**	*buor*	*pa un*	*melyikért*	*cötlaja*	*cuál*	*hangisi*
	41012028	*buor*	*pa un*	*melyik*	*cötlaja*	*cuál*	*hangisidir*
TWO	**40023019**	*buor*	*pa un*	*melyik*	*cötlaja*	*cuál*	*hangisi*

Table 4.48: Special codings for THING.SELECTION.MULTIPLE.PL

Northern Dagara has the form *abobe* referring exclusively to the identification of non-human objects in the plural (Mwinlaaru 2017: 154-155). In contrast, a selection of a singular referent is asked with *bour*. In Welsh, the number distinction of the interrogative 'which' is marked with two formally related constructions. For referents in the plural, the form *pa rai* is used, whereas *pa un* is indicative of a singular target (King 2003: 100-101). The component *pa* serves as the general question word meaning 'which' and precedes a noun. Syntactically, both constructions cannot be followed by a noun. In Hungarian, *melyek* designates 'which' in the plural, whereas *melyik* is translated as 'which one' (Rounds 2001: 134). The element *cötl-*

in Tetelcingo Nahuatl generally indicates ‘which’. When it is combined with *-aja* ‘he’, then the whole structure refers to ‘which one’. Contrastingly, when *cötl-* appears with *-ejua* ‘they’, the meaning is extended to ‘which ones’ (Tuggy 1979: 29).[24] The question word *cuáles* in Spanish is the plural form of *cuál* ‘which’. Turkish adds the plural suffix *-leri(n)* to the interrogative *hangi-* ‘which’ to mark the plurality of the referents (van Schaaik 2020: 67).

In languages with a comprehensive system of noun classes, the plurality of the referents is commonly denoted with different class prefixes or agreement markers. The following Table 4.49 gives five examples (Robert 2016: 3; Watkins1937: 135; Cox 1975: 116; Soukka 1999: 136; Lindblom 1914: 22). The corresponding markers for the plurality is marked in bold.

		Wolof	Nyanja	Rundi	Noon	Tharaka
MULTIPLE.PL	**40019018**	***y****an*	***wo****tani*	***ibi****he*	***c****iida*	***ma****rîkû*
	40022036	*ban*	NA	*irihe*	*yiida*	*rîrîkû*
MULTIPLE.SG	**43010032**	*ban*	*yiti*	*ikihe*	*wiida*	*bûrîkû*
	41012028	*ban*	*liti*	*irihe*	*yiida*	*rîrîkûrîrîkû*
TWO	**40023019**	*lan*	*n’chiti*	*ikihe*	*ya*	*mbi*

Table 4.49: Prefixes to mark THING.SELECTION.MULTIPLE.PL

		Czech	Danish	German	Romanian	Croatian	Gagauz
MULTIPLE.PL	**40019018**	*která*	*hvilke*	*welche*	*care*	*koje*	*angılarını*
	40022036	*které*	*hvilket*	*welches*	*care*	*koja*	*angı*
MULTIPLE.SG	**43010032**	*který*	*hvilken*	*welches*	*care*	*koje*	*angısı*
	41012028	*které*	*hvilket*	*welches*	*care*	*koje*	*angı*
	41002009	*co*	*hvad*	*was*	*ce*	*što*	*ne*
TWO	**40023019**	*co*	*hvad*	*was*	*ce*	*što*	*ne*
	40023017	*co*	*hvad*	*was*	*ce*	*što*	*ne*

Table 4.50: ‘which’ for THING.SELECTION.MULTIPLE vs. ‘what’ for THING.SELECTION.TWO

24 The meaning of the element *-nu* is not given in the grammar.

Another formal possibility is that in some languages one sub-cluster of SELECTION is denoted with a coding borrowed from other categories, normally the form translated as 'what'. As displayed in Table 4.50 above, six sampled languages choose the interrogative generally indicating 'which' for SELECTION.MULTIPLE, while the form meaning 'what' is found in SELECTION.TWO.

Two languages displays an inverse situation, as shown in the next Table 4.51. This time, Eastern Bru, an Austro-Asiatic language spoken in Laos, and Gwich'in prefer to encode SELECTION.MULTIPLE with the construction meaning 'what', i.e., *ntrớu* and *jidii*, whereas the contexts of SELECTION.TWO are asked with the interrogative meaning 'which (one)'.

		Eastern Bru	**Gwich'in**
MULTIPLE.PL	**40019018**	*ntrớu*	*jidii*
	40022036	*ntrớu*	*jidii*
MULTIPLE.SG	**43010032**	*ntrớu*	*jidii shrit*
	41012028	*aléq*	*jidii*
	41002009	*aléq*	*jidii shrit*
TWO	**40023019**	*aléq*	*jidii shrit*
	40023017	*aléq*	*jidii shrit*

Table 4.51: 'what' for THING.SELECTION.MULTIPLE vs. 'which' for THING.SELECTION.TWO

		Dii	**Nomaande**
MULTIPLE.PL	**40019018**	*téé*	*hάányέ*
	40022036	*tέlá*	*hάányέ*
MULTIPLE.SG	**43010032**	*téé*	*hάányέ*
	41012028	*tέlá*	*hάányέ*
	41002009	*ęn*	*aátέ*
TWO	**40023019**	*ęn*	*aátέ*
	40023017	*ęn*	*aátέ*

Table 4.52: 'where' for THING.SELECTION.MULTIPLE vs. 'which' for THING.SELECTION.TWO

In Dii and Nomaande, as given in Table 4.52 above, the interrogative referring to 'where' is found in SELECTION.MULTIPLE, while the form meaning 'which' is used for SELECTION.TWO (Bohnhoff 2010: 312; Wilkendorf 1998: 26). However, due to the limited information about these languages, it cannot be ascertained yet whether the interrogative 'where' in these two languages also has the meaning of 'which' or this usage is a consequence of the special semantic facet of SELECTION.MULTIPLE.

4.5.8 THING.KIND

Eight contexts are assigned into this sub-cluster and they are presented in (4.33).

(4.33) *eng-x-bible-common*

43002018b	***What** miraculous sign will you show us?*
43006030a	***What** miraculous sign will you do, that we can see and believe you?*
42006032	*If you love those who love you, **why** should you be commended?*
40005046	*If you love only those who love you, **what** reward do you have?*
44019003	***What** baptism did you receive?*
41004030b	***What** parable can I use to explain it?*
41011028a	***What kind of** authority do you have for doing these things?*
44007049a	***What kind of** house will you build for me?*

Similar to the semantic feature of sub-cluster PERSON.KIND discussed in §4.4.7, the contexts in (4.33) reflect the intention to obtain the type or kind of a certain object. In this sense, these eight questions represent the subdomain named **THING.KIND**. Different from THING.SELECTION, the addressees of THING.KIND are not provided a set of options from which the expected answer should be generated. Instead, the purpose of the inquiry is to further describe or define the quality of the referent.

Syntactically, the interrogatives used in (4.33) are all followed by a noun and thus function like an adjective. In English, THING.KIND is asked with the question word *what* or the combination *what kind of*. It is unexpected that the question word *why* appears in 42006032. Yet, in many other languages the interrogative part of this context is replaced by the

expression translated as ‘what kind of blessing’, which results this context being classified into this sub-cluster.

In terms of the interrogative form applied for THING.KIND, some sampled languages have a unique structure, as given in Table 4.53 below.

	Burarra	Khalkha	Dogrib	Rundi	Japanese	North Saami	Karakalpak
43002018b	NA	*yamar*	*amìı*	*ki*	どんな	*makkár*	*qanday*
43006030a	*gun-guyinmiya*	*yamar*	NA	*ki*	どんな	*makkár*	*qanday*
42006032	NA	*yuun*	NA	*iki*	どれ	NA	*qanday*
40005046	*nyiburr-yinmiya*	*yamar*	NA	*ki*	なん	NA	*qanday*
44019003	*gun-guyinmiya*	*yamar*	*ayìı t'à*	*ki*	だれ	*makkár*	*qanday*
41004030b	*nguburr-yinmiya*	*yamar*	*dàhòt'ıı*	*iki*	どんな	*makkár*	*qanday*
41011028a	NA	*yamar*	*ayìı*	*ki*	何	*gean*	*qanday*
44007049a	*gun-guyinmiya*	*yamar*	*dàhòt'ıı*	*ki*	どんな	*makkár*	*qanday*

Table 4.53: Unique forms for THING.KIND

As discussed in §4.4.7 PERSON.KIND, Burarra and Khalkha present a morphologically undecomposable interrogative translated as ‘what kind of’, i.e., the stem *-guyinmiya* in Burarra and *yamar* in Khalkha. In most contexts of THING.KIND, these two forms are also applied. The form *dàhòt'ı(ı)* in Dogrib specifically refers to ‘what kind’ (Saxon & Siemens 1996: 12) and is found in two contexts of this sub-cluster. For other questions, the interrogatives *amìı* ‘what person’ and *ayìı* ‘what’ are used. The suffix *t'à* following *ayìı* in 44019003 serves to build an adverb (Saxon & Siemens 1996: 100).

Rundi has an invariable interrogative *ki* ‘what kind of’ that usually follows the modified noun (Cox 1975: 142). The contexts in 42006032 and 41004030b are asked with the form *iki* ‘what’. In Japanese, the combination *don'na* どんな indicates ‘which’ or ‘what kind of’ (Bunt 2003: 230). The other attested forms are *dore* どれ ‘which (from many)’, *nan* なん ‘what’, *nani* 何 ‘what’, and *dare* だれ ‘who’. In North Saami, most contexts of this sub-cluster are asked with *makkár* ‘what kind of’, while *gean* found in 41011028a is inflected from *gii* ‘who’ for the genitive case (Aikio & Ylikoski 2010: 70, 172). The last example is found in

Karakalpak, a Turkic language spoken in Uzbekistan. This language has the form *qanday* 'what kind of' that is applied in all eight contexts of this sub-cluster (Wurm 1951: 562).

In the sampled languages of the Turkic family, THING.KIND shares the same morphological form with MANNER, except for the unique coding of THING.KIND in Karakalpak mentioned above. That is, according to the context, this interrogative can be translated either as 'what kind of' or 'how'. Examples are given in Table 4.54 below.

	Turkish	**Turkmen**	**Uyghur**
43002018b	*nasıl*	*näme*	*qeni*
43006030a	*nasıl*	*nähili*	*qandaq*
42006032	NA	*näme*	*neri*
40005046	*neye*	*näme*	*qandaqmu*
44019003	*neye*	*nämä*	*neme*
41004030b	*nasıl*	*haýsy*	*qandaq*
41011028a	*hangi*	*haýsy*	*qaysi*
44007049a	*ne gibi*	*nähili*	NA

Table 4.54: Interrogatives for THING.KIND in Turkic family

According to Göksel & Kerslake (2005: 265), the form *nasıl*, which is normally translated as 'how', can also serve as a determiner and indicates 'what kind of' in Turkish. With this denotation, it is then replaceable by the construction *ne gibi* (lit. 'what like') found in 4407049a. The other forms found in the THING.KIND cluster are *neye* 'why' and *hangi* 'which'. In Turkmen, *nähili* can be used to ask for 'how' and 'what kind' (Hoey 2013: 28). This language also adopts *näme* 'what', *namä* inflected from *näme* for the dative case and *haýsy* 'which' for THING.KIND. Uyghur manifests a similar case to Turkmen. That is, the question word *qandaq* emphasizes the qualities of an object, while it is applicable for both meanings 'how' and 'what kind' (Engesæth et al. 2009: 25, 28).

Like the construction *what kind of* in English, some sampled languages pattern the expression for THING.KIND in a similar way. Normally, the general interrogative meaning 'what' is taken as the base. See six examples found in the following languages in Table 4.55.

	Romanian	Spanish	Irish	Cherokee	German	Dutch
43002018b	*ce*	*qué*	*cén*	*gado*	*welches*	NA
43006030a	*ce*	*qué*	*cén*	*gadono udsi*	*was für*	NA
42006032	*ce*	*de qué*	*cad*	*gado*	*welchem*	*wat voor*
40005046	*ce*	*qué*	*cad*	*gado*	*welchen*	NA
44019003	*ce*	*qué*	*cén*	*gado usdi*	*was*	*wat*
41004030b	*ce*	*qué*	*cén*	*gadoge*	*welchem*	*wat*
41011028a	*ce*	*qué*	*cén*	*gado*	*welcher*	NA
44007049a	*ce fel de*	*qué clase de*	*cén sórt*	*gado usdesdi*	*was für*	*wat voor*

Table 4.55: Interrogatives derived from 'what' for THING.KIND

As already appearing in §4.4.7 PERSON.KIND, the combination *ce fel de* in Romanian and *qué clase de* in Spanish resemble the English pattern 'what kind of'. In 44007049a, Irish also provides a variation *cén sórt*. In this structure, the interrogative *cén* 'which' is chosen as the base for THING.KIND (Stenson 2008b: 37). Another possibility is found in Cherokee. This language combines *gado* 'what' and *udsi* 'something' for the questions 'what is it that…' or 'what kind of' (Montgomery-Anderson 2008: 482).[25] In German and Dutch the interrogative for THING.KIND is also compositional. Yet, the components are changed to the question word meaning 'what' and the preposition meaning 'for', which constitute *was für* in German and *wat voor* in Dutch.

4.5.9 QUANTITY.MASS

In the grouping with eight sub-clusters, two contexts asked with *how much* in English are classified into one subgroup. They are given in (4.34) below. These two contexts are successive in the Bible and have alike content. In these two questions, the speakers inquire about the quantity of the object that the addressee owes. It can be informed from the contextual information that the target objects refer to oil and wheat, respectively. Normally, this kind of material cannot be separated into individuals and counted.

[25] In Montgomery-Anderson (2008), the form *usdesdi* attested in 44007049a is not found. However, considering the highly similar appearance with *udsi*, it is deduced that the meaning of *usdesdi* would not be hugely deviated from *usdi* 'something'.

(4.34) eng-x-bible-common

42016007 *Then the manager said to another, '**How much** do you owe?' He said, ' One thousand bushels of wheat.'*

42016005 *He said to the first, '**How much** do you owe my master?'*

In terms of the expression of uncountable objects, it is frequent that no grammatical marking for the number, i.e., singular or plural, can be added to the denoting noun phrases. Such a type is traditionally labeled UNCOUNTABLE or MASS nouns (Crystal 2008: 119). On the opposite, the separable substances are designated by the COUNTABLE nouns. For this sub-cluster, the name **QUANTITY.MASS** is given. The sub-cluster of QUANTITY.COUNT is assigned to the primary cluster MANNER/EXTENT by the algorithm. This will be discussed in §4.7.4 later. In terms of this semantic distinction, a formal division is often found in interrogative codings across languages. English, for instance, is known by having two constructions *how much* for mass referents, as shown in (4.34), and *how many* for countable objects.

In terms of the mass-count distinction, two issues emerge. The first one is about the different categorizations of these two QUANTITY-related sub-clusters. It is common across languages that the interrogative for QUANTITY is derived from the form for MANNER, such as *how* for *how many/much* in English. A description of this kind of composition will be given in §4.7.4 QUANTITY.COUNT. In this sense, an association between QUANTITY and MANNER is established. However, despite having a specific interrogative quantifier, some sampled languages still apply the form meaning 'what' for QUANTITY.MASS, as can be seen in the following Table 4.56 (Woollams 1996: 225; Petzell 2008: 92; Wilkendorf 1998: 26; Davies 2010: 444; Donaldson 2008: 103; Watkins 1937134; Cox 1975: 29). This kind of expression could be the cause of the classification of QUANTITY.MASS into the THING cluster. This might reflect the perception that an inseparable substance is commonly regarded as a whole, since it cannot be counted piece by piece. Thus, according to the context, the speaker inclines to directly ask 'what' this object is, instead of querying its amount. In this case, the question in 42016007 can be reformulated as *what do you owe?*. The alteration of the interrogative does not considerably impede the understanding of the inquiry intention.

	42016007	42016005	'what'
Batak Karo	*kai*	*kai*	*kai*
Kagula	*choni*	*choni*	*-ni*
Nomaande	*aátɛ́*	*aátɛ́*	*aátɛ́*
Madurese	*saponapa*	*saponapa*	*apa*
Dutch	*wat*	*wat*	*wat*
Nyanja	*wotani*	*chiyani*	*chiyani / -tani*
Rundi	*iki*	*iki*	*iki*

Table 4.56: 'what' for QUANTITY.MASS

The second issue is related to the interrogative forms used for QUANTITY.MASS and QUANTITY.COUNT. It should be noticed that the mass-count classification is deeply influenced by the culture and thus can vary from language to language. In other words, the use of interrogatives for QUANTITY.MASS and QUANTITY.COUNT is language-specific. Many languages employ the identical form for both mass and countable referents. An example is found in Mandarin in which the expression *duōshǎo* 多少 is used for all questions about QUANTITY. In contrast, a number of languages display a similar pattern in English, i.e., adopting different forms for these two sub-clusters. When a language marks the mass-count difference in interrogatives, the structure translated as 'how much' is normally attested in the context of QUANTITY.MASS, while the form for QUANTITY.COUNT is marked as 'how many'. However, it is not always the case. For instance, two sampled languages, Yine and Toro So Dogon, have two interrogatives respectively meaning 'how much' and 'how many', as seen Table 4.57 below. Yet, questions of both QUANTITY.MASS and QUANTITY.COUNT are all encoded with the form translated as 'how many'. The reason might be that speakers of these languages prefer to consider the referents in the questions in (4.34) to be countable. Moreover, it is also possible that although the interrogatives, i.e., *gi pejnu* in Yine and *aŋa* in Toro So Dogon, are translated into English as 'how many' (Hanson 2010: 126; Heath 2017: 284), their semantic denotation actually covers a larger scope. Yet, this speculation needs more support from the grammatical references.

	Yine	Toro So Dogon
'how much'	*gi pso*	*yagɔ baa*
'how many'	*gi pejnu*	*aŋa*
QUANTITY.MASS	*gi pejnu*	*aŋa*
QUANTITY.COUNT	*gi pejnu*	*aŋa*

Table 4.57: 'how many' for QUANTITY.MASS and QUANTITY.COUNT

The domain MASS and COUNT are both subsumed under the QUANTITY category, because it is recurrently attested that the corresponding interrogatives are related in form, just like the shared base *how* in English. Nevertheless, languages can also create a unique form separately to query the quantity of mass and countable referents. The following Table 4.58 gives examples from seven sampled languages.

	42016007	42016005	countable
Welsh	*faint*	*faint*	*sawl*
Finnish	*paljonko*	*paljonko*	*montako*
Hungarian	*mennyivel*	*mennyivel*	*hány*
Yucatec Maya	*buca'aj*	*buca'aj*	*jay p'éel*
Romanian	*cât*	*cât*	*câte / câți*
North Alaskan Inupiatun	*qanutulli*	*qanutun*	*qavsich*
Uyghur	*qanchilik*	*qanchilik*	*qanche*
Spanish	*cuánto*	*cuánto*	*cuántos / cuántas*

Table 4.58: Unanalyzable forms for QUANTITY.MASS

In Welsh, the form *faint* 'how much/many' is completely different from *sawl* 'how many' in the form. According to King (2003: 125-126), when *faint* appears before a singular noun, it refers to 'how much', while its meaning changes to 'how many' when this question word precedes a noun in the plural. In contrast, the form *sawl* is used specifically to denote 'how many'. In Finnish, the meaning 'much' is encoded in *paljon*. When this word is combined with the interrogative suffix *-ko*, it is then applicable for questions about 'how much'.

Conversely, the form *montako* is applied to ask for ‘how many’ (Karlsson 2008: 6, 140). Hungarian adopts the interrogative *mennyi* to enquire about the amount of mass referents and *hány* for countable objects (Rounds 2001: 135). In QUANTITY.MASS, the attested structure *mennyivel* is declined from *mennyi* in the singular instrumental case. In Yucate Maya, a Mayan language spoken in Mexico, *buca'aj* used in QUANTITY.MASS can be translated as ‘how much’, ‘how many’ or ‘what quantity’. However, for the coding specifically for ‘how many’ of an inanimate object, the interrogative particle *jay-* should be combined with the number classifier *p’éel* (Bolles & Bolles 2019: 41, 45).

In Romanian, the interrogative *cât* is attested in QUANTITY.MASS. This form is indicative of ‘how much’ when it appears in the singular. To denote the number of countable nouns, the plural form *câte* (for feminine nouns) or *câți* (for masculine and neuter nouns) ‘how many’ should be employed (Gönczöl-Davies 2008: 55). In North Alaskan Inupiatun, the interrogative *qanutun* ‘how much/how long’ has similar initials with *qavsich* ‘how many (are they)’ (Seiler 2012: 164, 168). Yet, a further morphological decomposition is not given in the grammar. A resembling case is also found in Uyghur. The form *qanchilik* for QUANTITY.MASS means both ‘how much’ and ‘how many’. This question word and the form *qanche* exclusively meaning ‘how many’ share the same beginning whose origin is not explained in the reference (Engesæth et al. 2009: 265). The last example is provided in Spanish. In QUANTITY.MASS, the question word *cuánto* is utilized. This form should be followed by a masculine noun in the singular or an uncountable noun. In this sense, it usually means ‘how much’. In contrast, the plural form *cuántos* (for masculine nouns) and *cuántas* (for feminine nouns) refer to ‘how many’.

Similar to the pattern *how much* in English, the forms attested in QUANTITY.MASS are also frequently compositional in the sampled languages. Such a construction usually takes the interrogative meaning ‘how’ or ‘what’ as the base. In some cases, interrogatives meaning ‘where’ and ‘which’ can occur in the structure as well. Table 4.59 below provides seven examples.

In Danish, *hvor* ‘where/how’ appears in both constructions for mass and countable objects. The mass-count distinction is marked by the second component *meget* ‘much’ and *mange* ‘many’. German combines *wie* ‘how’ separately with *viel* ‘much’ and *viele* ‘many’ for these two semantic subdomains. In Icelandic, the form *hve* ‘how’ is used to build *hve mikið* ‘how

much' and *hve margar* 'how many'. In 42016007, only the interrogative *hvað* 'what/how' appears. In North Saami, *man* 'how' occur with *ollu* 'much' to ask for the quantity of mass objects, whereas the morphologically independent form *galle* is indicative of 'how many' (Kahn & Valijärvi 2017: 232). Noon displays a structural difference, i.e., *na* 'how' is placed after *hín* 'amount to/be equal' to ask for QUANTITY.MASS. For countable nouns, the interrogative stem *-era* should appear with a suitable agreement marker (Soukka 1999: 136, 183).

	42016007	**42016005**	**countable**
Danish	*hvor meget*	*hvor meget*	*hvor mange*
German	*wieviel*	*wieviel*	*wie viele*
Icelandic	*hvað*	*hve mikið*	*hve margar*
North Saami	*man ollu*	*man ollu*	*galle*
Noon	*hín na*	*hín na*	*-era*
Parecís	*xoanere*	*xoanere*	*xoanama*
Turkish	*ne kadar*	*ne kadar*	*kaç*

Table 4.59: Analyzable forms for QUANTITY.MASS

The last two sampled languages in Table 4.59 choose the interrogative for 'what' as the base for QUANTITY.MASS. In Parecís, when the question word *xoana* 'what' is linked with the nominalizer *-re*, it is used as an interrogative quantifier for mass nouns. When *xoana* is followed by the suffix *-ma* which probably means 'quantity', the entire combination means 'how many' (Brandão 2014: 336). Turkish combines *ne* 'what' with *kadar* 'until/as…as' to indicate 'how much', whereas *kaç* is used for countable nouns (Göksel & Kerslake 2005: 60, 125).

4.5.10 Summary

This chapter presented the primary cluster of THING and its internal classification. With a relatively large amount of contexts and sophisticated semantic differentiation, more sub-cluster are identified within this group, like the last cluster PERSON. In the first sub-cluster

THING.PATIENT discussed in §4.5.2, the questions target at the semantic role patient, i.e., a referent influenced by the action performed by someone else.

In §4.5.3, the THING.THINK sub-cluster shows a mixed application of the interrogative codings of THING and MANNER, while no dedicated interrogative form is found in the sampled languages for this subgroup. This demonstrates a semantic overlap between these two categories. Nevertheless, such a situation arouses the suspicion whether this sub-cluster can be seen as a cross-linguistically applicable comparative concept. A discussion about this issue will be given in Chapter 6.

Three sub-clusters presenting interrogative verbs indicating the motion *do*, *say* and *happen* are elaborated in §4.5.4 to §4.5.6. Although the majority of sampled languages do not specially mark the interrogative in these sub-clusters, there are also a few special verbal form attested in the corresponding contexts.

The sub-cluster THING.SELECTION is presented in §4.5.7. Addressees of this kind of question are requested to give the answer within a set of options. Based on the number of options, a finer classification is manually performed, which leads to the further establishment of two subsets, i.e., THING.SELECTION.TWO and THING.SELECTION.MULTIPLE. The former indicates two possibilities for the answer, whereas the latter is composed of questions with a set of more than two choices. On the basis of the specialized marking in interrogatives, THING.SELECTION.MULTIPLE can be again divided into THING.SELECTION.MULTIPLE.SG and THING.SELECTION.MULTIPLE.PL. The queried goal of the THING.SELECTION.MULTIPLE.SG is expected in the singular, whereas questioners of THING.SELECTION.MULTIPLE.PL seek an answer in the plural.

In §4.5.8, interrogative contexts of the THING.KIND sub-cluster aim at the quality of the referent object. The answer is not supposed to be selected from a limited set of choices, unlike in THING.SELECTION. The last sub-cluster of the primary group THING is QUANTITY.MASS presented in §4.5.9. In the contexts of this sub-cluster, the amount of a mass object is inquired about. In terms of the uncountability of the referent item, some languages choose the coding meaning ‘what’ to mark the questions, which might account for the grouping of this sub-cluster into the cluster THING.

4.6 Cluster of INTENTION

The cluster of the INTENTION category will be presented in this chapter. An overview and the internal structure of this cluster will be first provided in §4.6.1. In §4.6.2, I will give some contexts suggested by the algorithm as representatives of the INTENTION cluster. In §4.6.3, I will discuss the existence of a possible sub-cluster INTENTION.PURPOSE.

4.6.1 Overview

This cluster comprises 89 interrogative contexts. According to Table 4.1 in §4.1, the average silhouette width of this cluster lies at 0.51, which is the highest among all six primary clusters. Table 4.60 provides a selection of verses of this cluster. As can be seen, the question words used in English, German, and Mandarin for these contexts are all related to causality. Among contexts in the Bible, the questioners usually want to query the intention of someone's behavior or performance. Questions asking for the reason for natural phenomena, for instance, *why is the sky blue*, are rarely found in biblical texts. In order to distinguish these two types, this cluster is labeled **INTENTION**.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	42006002	0.62635	*why*	*warum*	為什麼
2	42018019	0.61730	*why*	*warum*	為什麼
3	51002020	0.61660	*why*	*warum*	為什麼
4	40020006	0.61491	*why*	*warum*	為什麼
5	42024005	0.61207	*why*	*warum*	為什麼
6	43008046b	0.59369	*why*	*wie*	為什麼
7	42019031	0.54092	*why*	*weshalbe*	為什麼
8	62003012	0.46844	*why*	*weswegen*	為什麼
9	40009011	0.46219	*why*	*warum*	為何
10	44005009	0.37198	*how*	*warum*	怎麼
11	44005004	0.36075	*what*	*warum*	怎麼
12	41014004	0.29203	*why*	*wozu*	為什麼

Table 4.60: Verse selection of cluster INTENTION

Figure 4.11 below shows the internal distribution of INTENTION with symbols indicating interrogative codings used in the English translation *eng-x-bible-common*. As displayed in this graph, most of the data points are spread on the left side. There is no clearly identifiable grouping within them. Only a few dots are located away from the majority. They are encoded with the question word *how* and *what* in English, respectively.

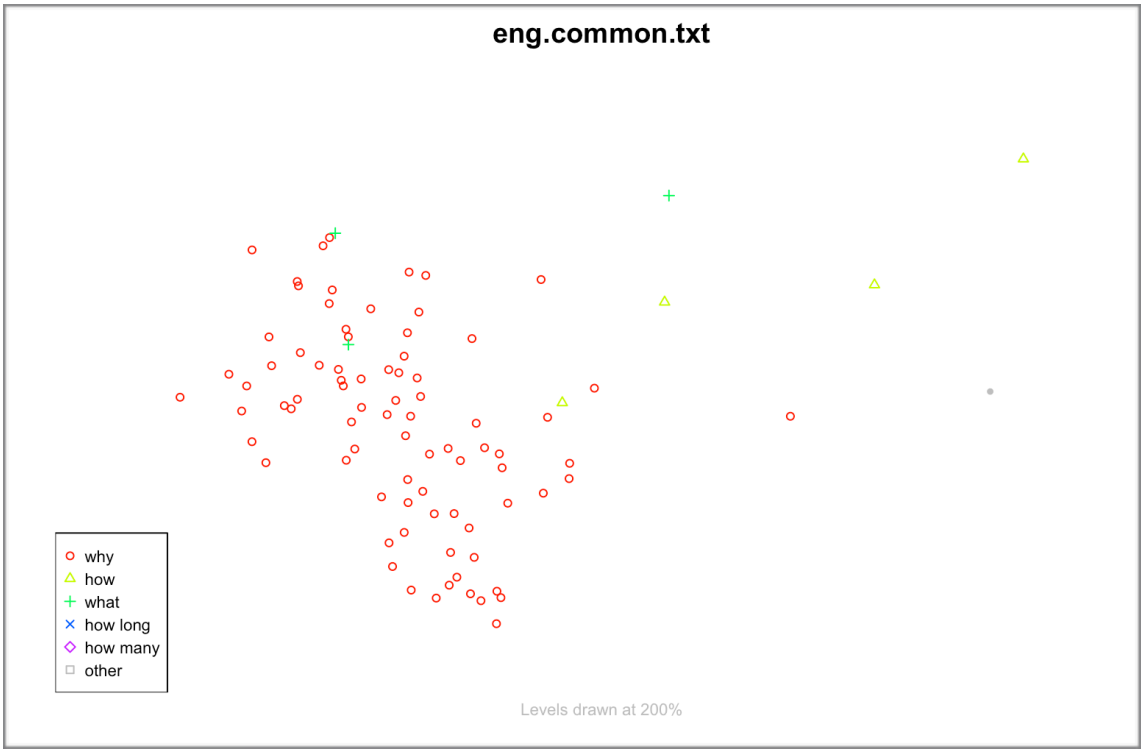

Figure 4.11: MDS plot of cluster INTENTION

Figure 4.12 below shows the result of the second level of clustering. According to this Figure, 89 contexts are optimally classified into 14 groups. Only one group with three contexts displays a possible relationship to the concept of PURPOSE (see §4.6.3). In terms of the other 13 groups, no special semantic meaning can be inferred through the usage of interrogatives or textual content. The clustering with two groups also appears to be a good outcome. These two groups contain 56 and 33 contexts, respectively. Yet, it is still difficult to identify their semantic meaning. In this sense, I choose the first ten contexts with the highest silhouette width as the representative of the whole cluster INTENTION. They will be presented in the following §4.6.2.

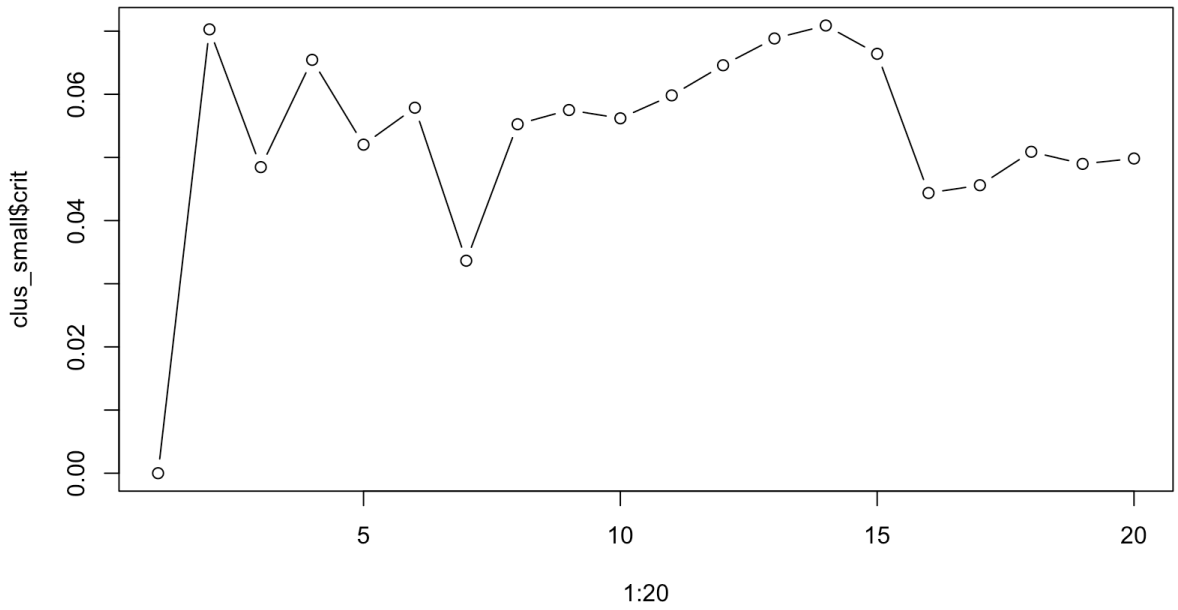

Figure 4.12: Suggested sub-clusters of INTENTION

The reason for the numerous grouping with 14 groups might be related to the fact that some languages have more than one form for questions about INTENTION or REASON. However, there is very minimal semantic distinction between these forms, which leads to the difficulty to explain the grouping from a semantic perspective. A typical example demonstrating this issue is German. This language has several interrogatives equating to the question word *why* in English, e.g., *warum*, *weshalb*, *weswegen*, *wieso*, *aus welchem Grund*. Although these forms have different etymological roots and morphological structures, they are normally exchangeable without bringing any alteration to the expression. Another example from the sample is Balantak. According to van der Berg & Busenitz (2012: 205), the question word *kadai* refers to 'why'. This form has two semantic equivalents, i.e., *nokadai* and *nongko'upa*. The latter is derived from *upa* 'what'. In the contexts of INTENTION, these three interrogatives are all found without exhibiting any clear semantic differentiation.

In another situation, the distinction between different codings is irrelevant to the semantic meaning but is decided by the formality of the utterance. For instance, two interrogatives, *naze* なぜ and *dōshite* どうして, are attested in the Japanese translation of this cluster. According to Bunt (2003: 227-228), the former serves as the formal structure to enquire about the reason, whereas the latter is considered an equivalent applied in a more casual

conversation. However, such a kind of different usage does not conduce to the identification of semantic meanings.

4.6.2 Representative contexts of INTENTION

In (4.35), a selection of five contexts representing INTENTION is shown.[26]

(4.35) *eng-x-bible-common*

42006002	***Why** are you breaking the Sabbath law?*
42018019	***Why** do you call me good?*
51002020	*If you died with Christ to the way the world thinks and acts, **why** do you submit to rules and regulations as though you were living in the world?*
40020006	***Why** are you just standing around here doing nothing all day long?*
42024005	***Why** do you look for the living among the dead?*

The target of these questions is to obtain the cause motiving someone to conduct a certain action. In terms of the morphological structure of interrogatives, it is recurrently attested that the form is derived from other categories. Among the sampled languages, two main types are found. The first type is based on the interrogative 'what'. Most sampled languages adopt this pattern to build the structure for INTENTION. Table 4.61 below provides four examples.

	Wolof	**Garifuna**	**Tabasco Chontal**	**Northern Dagara**
42006002	*lu tax*	*cásan uágu*	*cua' uc'a*	*bṍv n'v so*
42018019	*lu tax*	*ca uágu*	*cua' uc'a*	*bṍv n'v so*
51002020	*lu tax*	*cáti uágu*	*cua' uc'a*	*bṍv yə̃w*
'what'	*lu*	*ca*	*cua'*	*bṍv*

Table 4.61: 'what'-pattern for INTENTION

[26] The other five representative contexts are 44009004, 40009004, 40022018, 42019023, and 45009020c.

In Wolof, a Niger-Congo language spoken in Gambia and Senegal, the construction for INTENTION is composed of *lu* 'what' and the verb *tax* 'to cause' (Robert 2016: 12). In Garifuna, when the preposition *uágu* 'on/about' follows the question word *ca* 'what', they jointly form the construction meaning 'why' (Haurholm-Larsen 2016: 176). The element *-san* and *-ti* serve as interrogative particles (Suazo 2002: 222; González 2012: 75). Tabasco Chontal, a Mayan language spoken in Mexico, combines *cua'* 'what' with the preposition *uc'a* 'because' to denote 'why' (Keller & Luciano 1997: 62, 268). In Northern Dagara, two forms are used for INTENTION. The first one found in 42006002 and 42018019 comprises *bṽv* 'what', the identifying pronoun *n'v*, and *so* 'own'. The other construction appearing in 51002020 incorporates *bṽv* 'what' and the postposition *yãw* 'for the sake of' (Mwinlaaru 2017: 156, 218).

The second common basis for INTENTION is the form indicating 'how'. The following Table 4.62 provides examples from three sampled languages.

	El Nayar Cora	**Kagulu**	**Eastern Bru**
42006002	*a'ini*	*nhani*	*nǒ́q*
42018019	*a'ini een cɨn*	*nhani*	*nǒ́q*
51002020	*a'ini een cɨn*	*nhani*	*nǒ́q*
'how'	*a'ini*	*nhani*	*nǒ́q*

Table 4.62: 'how'-pattern for INTENTION

In El Nayar Cora, the question word *a'ini* 'how' is used together with *een* 'be' and *cɨn* 'with' to inquire about the cause (Casad 1984: 200). In Kagulu, the interrogative *nhani* can be indicative of 'how' as well as 'why' (Petzell 2008: 176-177). The same case also occurs in Eastern Bru. According to Miller (2017: 10), the interrogative *nǒ́q* marks both 'why' and 'how'.

4.6.3 A possible sub-cluster for INTENTION.PURPOSE

Three contexts are classified together as a subgroup that appears to be related to the concept of PURPOSE. They are illustrated in (4.36).

(4.36) *eng-x-bible-common*

42019033	*As they were untying the colt, its owners said to them, "**Why** are you untying the colt?"*
42019031	*If someone asks, "**Why** are you untying it?" just say, "Its master needs it."*
41011003	***Why** are you doing this?*

According to the grammatical description, the semantic differentiation between various interrogatives for REASON can be drawn between the concept of CAUSE and PURPOSE in some languages. The former specifically refers to the motive for conducting a certain action. In contrast, PURPOSE emphasizes the aim or goal that is supposed to be achieved through the involved performance. The interrogatives found in the languages presented in Table 4.63 below suggest the possible existence of the sub-cluster INTENTION.PURPOSE. In these languages, the interrogative indicating a purpose or the meaning 'for what' is applied in the three contexts in (4.36).

	Huallaga Huánuco Quechua	**Ayacucho Quechua**	**Parecís**	**Acehnese**	**Paraguayan Guaraní**
42019033	*imapätaj*	*imapaqtaq*	*xoana hoka*	*keupeue*	*maerã piko*
42019031	*imapätaj*	*imapaqtaq*	*xoare maheta*	*keupeue*	*maerãpa*
41011003	*imapätaj*	*imapaqtaq*	*xoare maheta*	*peusabab*	*maerãpa*
CAUSE	*imanirtaj*	*imanasqataq*	*xoana hoka*	*pakon*	*mba'ére piko*

Table 4.63: Examples of INTENTION.PURPOSE

In Huallaga Huánuco Quechua, the purposive suffix *-pä*/*-paq* and the content question marker *-taj* are attached to the interrogative root *ima* 'what' (Weber 1996: 433). On the contrary, *imanirtaj* for INTENTION.CAUSE is composed of *ima* 'what' and *ni-* 'say' (Weber 1989: 25). According to Zariquiey & Córdova (2008: 101, 280), a similar pattern for PURPOSE is found in *imapaqtaq* in the related language Ayacucho Quechua as well. Parecís combines the question word *xoare* 'what' with the purposive particle *maheta* for INTENTION.PURPOSE,

whereas the interrogative construction for the other contexts of INTENTION is composed of *xoare* or *xoana*, an interrogative particle, and the connective particle *hoka* (Brandão 2014: 80, 336, 339).

In Achenese, the form *keupeue* appears in two contexts of INTENTION.PURPOSE. According to Durie (1985: 151, 174-175), *keu* is a preposition meaning a goal or a purpose. For INTENTION.PURPOSE, it is combined with the question word *peue* 'what'. The form *peusabab* is not found in the grammar. However, it might comprise *peue* and *seubap* 'reason' (Durie 1985: 213). When acquiring a general reason, Achenese attaches *kon* 'reason' to *pa*, a variant of *peue*, to build the interrogative *pakon*.

In Paraguayan Guaraní, a Tupian language spoken in Paraguay, the interrogative *maerã* refers to 'for what', whereas *mba'ére* is a general form meaning 'why'. Both constructions are derived from *mba'e* 'what/thing'. To build *maerã* 'what for', *mba'e* is combined with the so-called nominal destinative aspect marker *-rã*. Conversely, the general interrogative *mba'ére* is composed of *mba'e* and the enclitic *=re* 'at' (Estigarribia 2020: 111). The elements co-occurring with these two forms are the interrogative enclitic *=piko* and *=pa.*

However, as can be seen in (4.36), the distinction between CAUSE and PURPOSE is not formally marked in English. In terms of German, although this language has two forms *wozu* (lit. 'to what') and *wofür* (lit. 'for what') that usually imply a purpose, the general question word *warum* and *weshalb* 'why' are used for this sub-cluster. The same situation also occurs in the Spanish translation. Spanish applies *para qué* for PURPOSE and *por qué* for CAUSE or a general reason. The latter is attested in all three contexts of this sub-cluster. Seemingly, even if a language differentiates the semantic domain PURPOSE and CAUSE, it is commonly grammatically correct that the form for PURPOSE is substituted by the one for CAUSE or a general reason. Yet, the other way around is seldom seen. In this sense, it lacks solid evidence for the conclusion that these three contexts denote the semantic domain PURPOSE.

4.6.4 Summary

This chapter presented the cluster of INTENTION. Different from the other four primary groups discussed before, it is difficult to interpret the internal structure of this cluster, although the algorithm has suggested several subdivisions. In this sense, ten contexts with the best silhouette width have been chosen in §4.6.2 as representatives of this cluster. This situation

might stem from the fact that many languages have various forms to ask for a reason or intention, while these codings are exchangeable without the interrogative meaning being changed. In other words, the formal differentiation of interrogative codings does not signify semantic distinction within this cluster.

Based on the special construction, it is suspected in §4.6.3 that there is a sub-cluster querying the purpose of an action. However, even if some languages, e.g., Spanish, have such a dedicated form to ask for the purpose, they do not use it in the contexts of this sub-cluster. Thus, the existence of the sub-cluster INTENTION.PURPOSE is in doubt.

4.7 Cluster of MANNER/EXTENT

In this chapter, the cluster representing the semantic domain MANNER/EXTENT will be illustrated. In §4.7.1, the basic information about this cluster will be given. Within the relevant contexts, the following four sub-clusters are identified and will be discussed in the corresponding sections:

- §4.7.2 — MANNER
- §4.7.3 — MANNER.STATEMENT
- §4.7.4 — QUANTITY.COUNT
- §4.7.5 — QUANTITY.FREQUENCY

4.7.1 Overview

The last primary cluster is composed of 56 contexts. The average silhouette width of this cluster is 0.33. Most contexts of this cluster are encoded with the question word *how* in English. Given that *how* in English is normally used for the concept of MANNER or EXTENT, the label **MANNER/EXTENT** is assigned to this cluster. Table 4.64 below provides information on twelfth selected contexts for a brief view. Figure 4.13 shows the internal distribution of cluster MANNER/EXTENT.

Nr.	Verse ID	Silhouette width	English	German	Mandarin
1	45010014c	0.54015	*how*	*wie*	怎
2	54003005	0.53906	*how*	*wie*	怎
3	45010014a	0.53629	*how*	*wie*	怎
4	43003012	0.51131	*how*	*wie*	怎麼
5	46015035a	0.48071	*how*	*wie*	怎樣
6	43009010	0.47590	*how*	*wie*	怎麼
7	43007015	0.39484	*how*	*wieso*	怎麼
8	40022012	0.39147	*how*	*wie*	怎麼
9	43001048	0.14151	*how*	*wie*	怎麼
10	42010026b	0.10328	*how*	*wie*	怎麼
11	41008020	0.00983	*how many*	*wie viele*	多少
12	40018021	0.00544	*how many times*	*wievielmal*	多少

Table 4.64: Verse selection of cluster MANNER/EXTENT

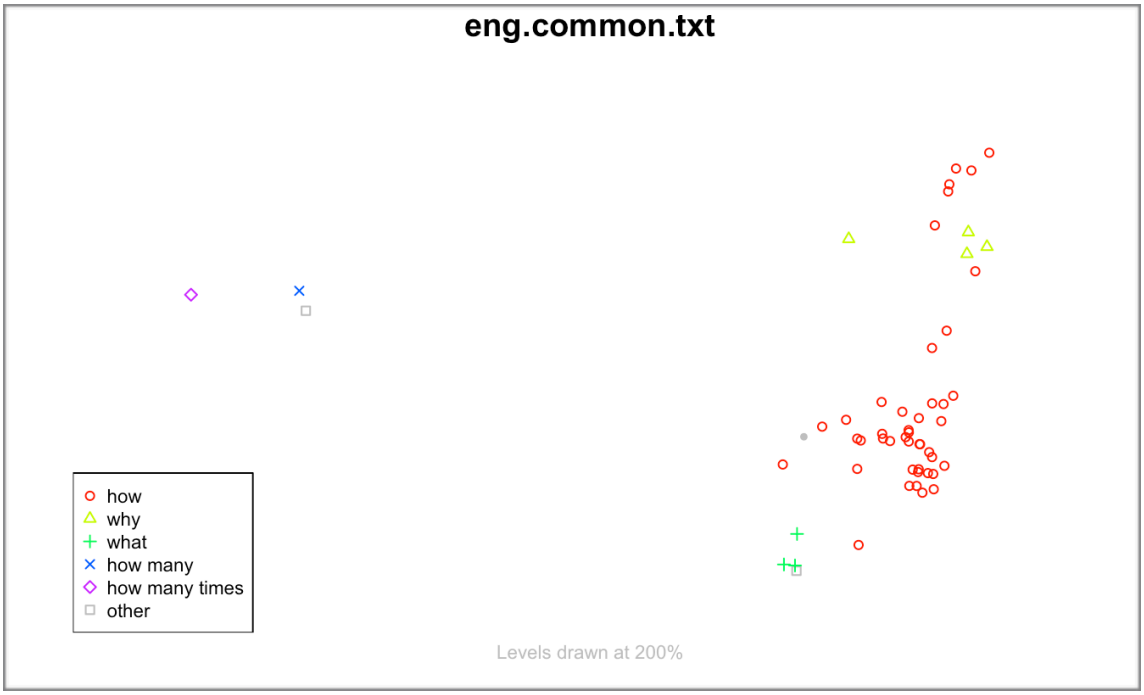

Figure 4.13: MDS plot of cluster MANNER/EXTENT

As can be seen in Figure 4.13, the majority of dots gather on the right side and spread vertically. Most of the contexts are encoded with the question word *how* in English. At the top of this group, there are four green triangles representing contexts asked with *why* in English.

Moreover, contexts with the question word *what* in English are located at the bottom, as the green crosses signalize. Far away from this big group, three dots are prominently found on the left in the graph. As indicated by the legends, they are marked with the construction *how many* and *how many times* in English, respectively, which signifies the association with the category QUANTITY.

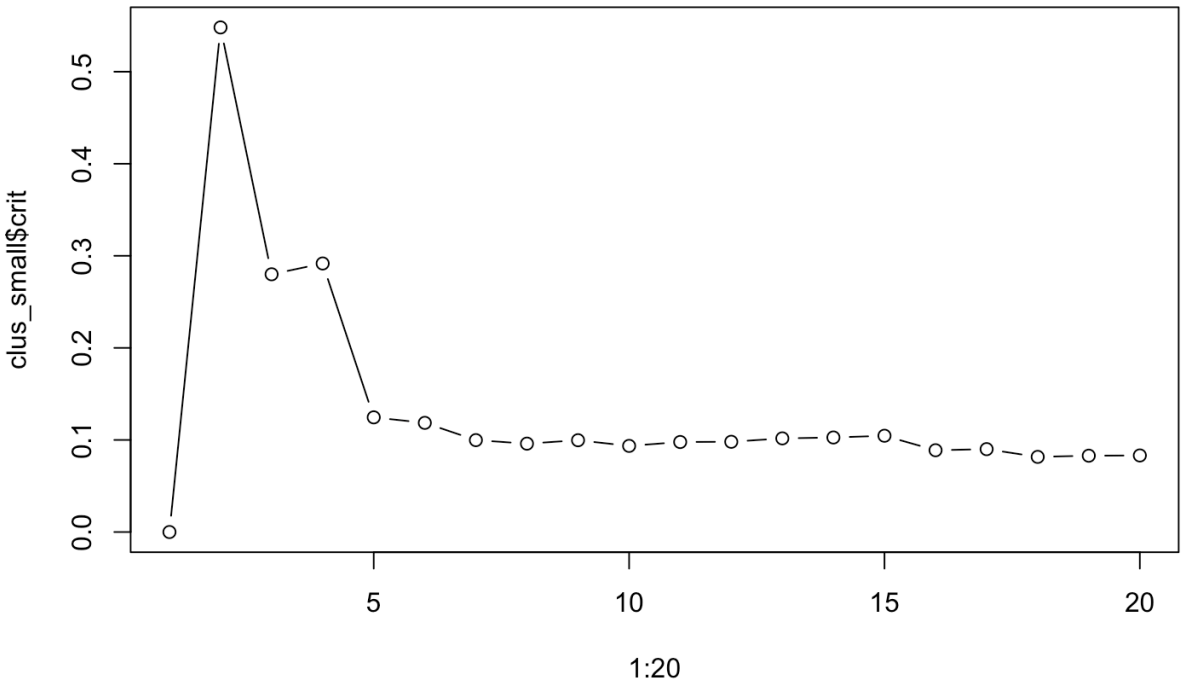

Figure 4.14: Suggested sub-clusters of MANNER/EXTENT

Figure 4.14 shows the results of the second level of clustering. The optimal grouping is found with two sub-clusters. The first one is composed of those three contexts related to the category QUANTITY, as just noted. The other 53 contexts comprise the second subgroup, which needs further internal classification. Thus, the suboptimal result with four subgroups is also taken into account. This time, 53 contexts are assigned to three subgroups. They contain respectively thirty-nine, eleven, and three contexts. The interpretation on clusters with 39 and 11 contexts will be given in §4.7.2 and §4.7.3, respectively. The last set with three contexts is semantically unclear.[27]

[27] The verse IDs of these three contexts are 46012017b, 46012017a, and 41009050.

4.7.2 MANNER

In the subgroup with 39 contexts, most contexts are with a high silhouette width. However, it is still difficult to allocate a suitable label for this sub-cluster. The reason is that the interrogative form for the MANNER category can normally express numerous meanings. This characteristic is found across languages. Take English as an example. According to dictionaries, some common situations in which the question word *how* can appear are given in (4.37) below.

(4.37) a. ***How*** *did you make the soup?*
(in what manner or way)

b. ***How*** *would you know whether it's rainy?*
(for what reason; capability; 'why')

c. ***How*** *are you?*
(in what state or condition)

d. ***How*** *big is the stadium?*
(to what extent)

As can be seen above, only the usage in (4.37a) reflects the inquiry about how an action or an activity is performed, i.e., the manner. The other three questions in (4.37b) to (4.37d) display quite different domains from the MANNER category. And for these usages, there is no extra morphological markedness added to the form *how* in English. Such a case is also attested in many sampled languages. This leads to the clustering being unable to classify these different situations according to the formal marking. In this case, a manual inspection of the interrogative denotation is needed.

According to the contextual information, five contexts are selected manually from the subgroup with 39 contexts as the representatives of **MANNER**, as given in (4.38) below. The other 34 contexts ask rather for a reason, i.e., the usage in (4.37b). In these five questions, the

function of *how* in English is to query the way of an action being conducted or the emergence of a state.

(4.38) *eng-x-bible-common*

41004013 *Then **how** will you understand all the parables?*
46015035a *But someone will say, "**How** are the dead raised?"*
43009010 *So they asked him, "**How** are you now able to see?"*
40021020 *"**How** did the fig tree dry up so fast?" they asked.*
43009026b ***How** did he heal your eyes?*

In terms of the morphological structure, most languages have a monomorphemic and independent coding for MANNER. That is, the form for MANNER is normally a basic lexeme and cannot be further divided into any meaningful element. Yet, two Austronesian languages in the sample show exceptions. These two languages have a construction composed of two lexemes — Iban with *baka ni*, lit. 'like which' (Omar 1969: 201) and Balantak with *koi upa*, lit. 'like what' (van den Berg & Busenitz 2012: 202).

In these two languages, the form for MANNER is derived from SELECTION and THING, respectively. These two categories are also the main source for the derivational constitution of MANNER. Considering the derivation from SELECTION, Chamorro provides a similar pattern *taimanu*, lit. *tai-manu* 'like this-which' (Topping & Dungca 1973: 160). The following Table 4.65 lists four languages demonstrating the derivation from THING (Schachter & Otanes 1983: 506; Weber 1989: 39, 327; Weber 1996: 432; Estigarribia 2020: 111-112; Durie1985: 151).

	Tagalog	**Huallaga Huánuco Quechua**	**Paraguayan Guaraní**	**Acehnese**
41004013	*paano*	*imanöpataj*	*(piko) mba'éicha*	*pakriban*
THING	*ano*	*ima*	*mba'e*	*pa=peue=pue*

Table 4.65: Derivation from THING for MANNER

4.7.3 MANNER.STATEMENT

This sub-cluster incorporates eleven contexts. Four of them with the highest silhouette width are given (4.39) as an illustration.

(4.39) *eng-x-bible-common*

43006042	***How** can he now say, 'I have come down from heaven?'*
43004009	***Why** do you, a Jewish man, ask for something to drink from me, a Samaritan woman?*
43012034a	***How** can you say that the Human One must be lifted up?*
42020041	***Why** do they say that the Christ is David's son?*

In these contexts, the function of *how* in English is in correspondence with the usage in (4.37b) above. That is, the purpose of the speaker is to obtain the factor that drives someone to make a statement or utterance. In terms of the question word attested in the English translation, it shows a mixture of *how* and *why*. In this sense, this sub-cluster can be seen as semantically connected to cluster INTENTION discussed in §4.6.2 to a certain degree. Considering that the content of these eleven contexts relates to someone's statement, this sub-cluster is labeled **MANNER.STATEMENT**.

	Wolof	**Parecís**	**Western Huasteca Nahuatl**	**Korean**	**Welsh**
43006042	*lu tax*	*xoana hoka*	*quenicatza*	어떻게	*sut*
43004009	*nan*	*xoana hoka*	*para tlen*	어떻게	*sut*
41012035	*lu tax*	*aliyakereya*	*para tlen*	왜	*pam*
'how'	*nan*	*aliyakere*	*quenicatza*	어떻게	*sut*
'why'	*lu tax*	*xoana hoka*	*para tlen*	왜	*pam*

Table 4.66: Examples for MANNER.STATEMENT

The interrogative *how* used in this subgroup usually co-occurs with the modal verb *can* in English. Pragmatically, this combination indicates that the speaker wants to express surprise

or disapproval of what someone said or did. In this case, this kind of question is normally considered rhetorical. Nevertheless, since for some questions it is explicitly marked that the speaker asks a question and the answer is provided in the subsequent text, it would be reckless to exclude all these eleven contexts from the class of content question. Thus, they are retained as the evidence of multiple meanings of the interrogative form 'how'.

Similar to the situation in English, many sampled languages apply both interrogatives meaning 'how' and 'why' in this sub-cluster. Table 4.66 above provides five examples (Robert 2016: 5, 16; Brandão 2014: 334,336; Beller & Beller 1977: 222-223; King 2003: 70).

4.7.4 QUANTITY.COUNT

As described at the start of this chapter, three contexts related to QUANTITY are assigned to the same subgroup. According to the content and the formal marking, the context in 40018021 is further separated and it will be discussed in the next section. The other two contexts representing **QUANTITY.COUNT** will be elaborated in this section and they are illustrated in (4.40) below.

(4.40) *eng-x-bible-common*

40015034	*Jesus said, "**How much** bread do you have?". They responded, "Seven loaves and a few fish."*
41008020	*And when I broke seven loaves of bread for those four thousand people, **how many** baskets full of leftovers did you gather?*

Same as the counterpart of QUANTITY.MASS discussed in §4.5.9, two questions of QUANTITY.COUNT also target at the amount of an object. The referents of the interrogative here are bread and basket, respectively. In English, the bread in the first question is asked with the interrogative construction *how much* that usually encodes uncountable items. Yet, in the German and Danish translations, for instance, the bread is reckoned to be countable, and, correspondingly, the interrogative *wie viele* and *hvor mange* 'how many' are employed. This demonstrates the second issue discussed in §4.5.9 QUANTITY.MASS. That is, the differentiation

between countable and uncountable objects varies hugely across languages. One cannot simply equate lexical items as belonging to COUNT or MASS across languages.

The fact that the interrogative for QUANTITY is frequently derived from the form meaning 'how' leads to the categorization of QUANTITY.COUNT into MANNER/EXTENT. To some degree, the concept of QUANTITY is in correspondence with EXTENT in the sense that they both refer to the range or amount of existence. Examples displaying this morphological pattern in Danish, German, and Icelandic have already been provided in Table 4.59 in §4.5.9 QUANTITY.MASS. Besides those, Table 4.67 below presents some more languages of this type.

	40015034	**41008020**	**'how'**
Maasina Fulfulde	*hono foti*	*hono foti*	*hono*
Finnish	*montako*	*kuinka monta*	*kuinka*
Nama	*matikō*	*matikō*	*mati*
Dutch	*hoeveel*	*hoeveel*	*hoe*
Tenharim-Parintintin-Diahoi	*maramomi*	*maramomi*	*marã*
Yine	*gi pejnu*	*gi pejnu*	*gi*

Table 4.67: Derivation from MANNER for QUANTITY.COUNT

In Maasina Fulfulda, *hono* 'how' is combined with the element *foti* to yield the construction for QUANTITY (Breedveld 1995: 41). Finnish has two ways to ask for the number of a countable item. One can either attach the interrogative suffix *-ko* to the word *monta* 'many' or combines *kuinka* 'how' with *monta* (Karlsson 2008: 6, 114). Both forms are presented in 40015034 and 41008020, respectively. Nama, a Khoe-Kwadi language spoken in Namibia, adds the number-driving suffix *-kō* to *mati* 'how' for the interrogative quantifier *matikō* (Hagman 1977: 51). The form *hoeveel* in Dutch is composed of *hoe* 'how' and *veel* 'many/much'. It is generally applicable for countable and uncountable objects (Donaldson 2008: 159). In Tenharim-Parintintin-Diahoi, there exists a connection between *marã* 'how' and *marmomi* 'how many/much' (Betts 2012: 161). However, a detailed description of the derivation is not provided in the grammar. Finally, Yine builds the interrogative quantifier *gi pejnu* 'how many' with *gi* 'how' and *pejnu* 'every' (Hanson 2010: 326).

4.7.5 QUANTITY.FREQUENCY

The context 40018021, as illustrated in (4.41) below, was originally classified to the sub-cluster QUANTITY.COUNT. Yet, after scrutinizing the interrogatives employed in this context, the content in this question is considered as representing the semantic subdomain **QUANTITY.FREQUENCY**. Thus, this context is manually separated as an independent sub-cluster.

(4.41) *eng-x-bible-common*

40018021 *Lord, **how many times** should I forgive my brother or sister who sins against me?*

The questioner of 40018021 intends to inquire about the number or the times of an occurrence. Semantically, this kind of information can also be counted and thus pertains to the category of QUANTITY, which is corroborated by the frequently attested morphological resemblance between the interrogatives for QUANTITY.FREQUENCY and QUANTITY.COUNT. It should be noticed that the interrogative form for QUANTITY.FREQUENCY is often absent in references of many lesser-described languages. In this case, only the recognizable part, which is normally the form for QUANTITY.COUNT, is marked during the data collection. In some other languages, a monolexemic form appearing in this question shows a high similarity with QUANTITY.COUNT, but it is not found in the grammar. For this situation, the whole structure is extracted and taken as the interrogative for QUANTITY.FREQUENCY.

Among the sampled languages with available references, two morphological patterns for QUANTITY.FREQUENCY are recurrently attested. The first one is like *how many times* in English, i.e., it is composed of the interrogative quantifier translated as ‘how many’ and another element. Due to the deficiency of grammatical information, the meaning of the other element is sometimes unknown. Some examples are given in Table 4.68 below (Strehlow et al. 2018: 294; Woollams 1996: 48, 225; Heath 2017: 284; Hagman 1977: 51; Naughton 2005: 102; Tuggy 1979: 29; Emenanjo 2015: 390; King 2003: 124; Weber 1989: 327).

The second pattern is similar to the construction of *how often* in English. In this structure, the interrogative equivalent to *how* in English is combined with the adverb meaning ‘often’.

This usage of the form ‘how’ appears to correspond with the function of querying EXTENT, as noted in (4.37d) above. Besides English, this composition is also found in the other three languages, as shown in the following Table 4.69. In these languages, the form ‘how’ is also the derivational source for the interrogative quantifier. These two patterns, i.e., ’how many times’ and ‘how often’, can coexist in a language. For instance, in different translations in English and German, either forms can be found.

	40018021	**‘how many’**
Western Arrarnta	*nthakintjarangama*	*nthakintja*
Batak Karo	*piga kali*	*piga*
Toro So Dogon	*aŋa baa*	*aŋa*
Nama	*matikō !nāde*	*matikō*
Czech	*kolikrát*	*kolik*
Tetelcingo Nahuatl	*quiejquechpa*	*quiejquech*
Igbo	*ùḫò ole*	*ole*
Welsh	*sawl gwaith*	*sawl*
Huallaga Huánuco Quechua	*ayca cutitaj*	*ayca*

Table 4.68: Examples of the pattern ‘how many times’

	40018021	**‘how’**
German	*wie oft ‘how often’/wievielmal ‘how many times’*	*wie*
Icelandic	*hve oft*	*hve*
Dutch	*hoe vaak*	*hoe*

Table 4.69: Examples of the pattern ‘how often’

4.7.6 Summary

The chapter discussed the cluster of MANNER/EXTENt and its internal distribution. Showing the reverse situation of INTENTION presented in §4.6, it is commonly seen across languages that the interrogative form translated as ‘how’ in English bears multiple semantic features and denotes various interrogative categories, e.g., MANNER, EXTENT and REASON. That is, there

exists a one-to-more relation between interrogative form and meaning. In this case, the uniformity of the formal marking obstructs the identification of the semantic differentiation and the internal clustering conducted by the algorithm. Thus, the interrogative contexts encoding MANNER, i.e., the way to perform a certain action, can only be manually examined and selected, as given in §4.7.2.

In §4.7.3, the sub-cluster MANNER.STATEMENT emerged through the automatic clustering. In order to ask the reason for making a statement, the sampled languages present a hybrid usage of 'how' and 'why' in this subgroup. Different from other sub-clusters, there is no unique form found in any sampled language for this kind of question. Therefore, it is uncertain whether the interrogative contexts of MANNER.STATEMENT can be regarded as a comparative concept, just as the sub-cluster PLACE.FROM.SOURCE.

The interrogative coding corresponding to 'how' in English also serves frequently as an element of other interrogative constructions, as for QUANTITY.COUNT elaborated in §4.7.4 and QUANTITY.FREQUENCY in §4.7.5. The former contains two contexts acquiring the quantity of a countable object, which appears as the counterpart of QUANTITY.MASS found in the primary group THING. The goal of QUANTITY.FREQUENCY is to query the times of an occurrence. According to the clustering results, the corresponding context was originally classified into QUANTITY.COUNT. Yet, in the light of the salient semantic facet and dedicated interrogative codings, the sub-cluster QUANTITY.FREQUENCY is manually established. Given that the concept of QUANTITY is similar to the extent of existence to some degree, the interrogative codings in QUANTITY.COUNT and QUANTITY.FREQUENCY appear to present the meaning of EXTENT.

5 Derivations between interrogative contexts

This chapter will illustrate the derivations between contexts within each primary cluster as well as across categories. In §5.1, some general remarks about the identification of derivational relations will be given. In §5.2 to §5.6, the internal derivations found within the primary cluster TIME, PLACE, PERSON, THING and MANNER/EXTENT will be discussed, respectively. Finally, the derivational links across different categories will be sketched in §5.7.

5.1 General notes

During the interpretation of clusters and sub-clusters in Chapter 4, it occurs recurrently that interrogatives are formally derived from the others. Even for the dedicated form applied for a certain context, it is often related to another basic interrogative. To draw a general map of the derivational relations between interrogative contexts, the forms used in the most representative context of each identified sub-cluster, i.e., the context with the highest silhouette width, are employed for the observation in this chapter. The following Table 5.1 lists the verse IDs of these 38 contexts.

According to the morphological structure of these interrogatives, the derivational directions can be inferred. To be noticed, the current analysis only takes the transparent composition of the structures into account. That is, the etymological developments, the synchronically unanalyzable constructions and the inflected forms are not considered. In the following content, the layout of the derivational maps is based on the internal structure of each primary cluster which has been displayed in the corresponding section in Chapter 4, e.g., Figure 4.2 in §4.2.1 plotting the distribution within TIME. In this regard, the distances between contexts in the derivational maps are approximately equal to their dissimilarities across the sampled languages. However, the position of some labels has been slightly adapted to a more comprehensible visual arrangement. To emphasize the tight semantic relation and make the maps more lucid, contexts with similar meanings are enclosed into a circle. Two derivationally connected contexts or circles are linked with a line with an arrow indicating the developmental direction. The thickness of the lines stands for the intensity of the derivation. That is, the more sampled languages display a certain derivational pattern between a couple of contexts or circles, the thicker the linking line is painted between them.

Primary cluster	Sub-cluster	Verse ID
TIME	TIME.GENERAL.PAST	40025039
	TIME.GENERAL.FUTURE	42021007a
	TIME.SELECTION	43006025
	DURATION.FUTURE	66006010a
	DURATION.PAST	41009021
PLACE	PLACE.EVENT	43001038b
	PLACE.OBJECT.SG	43009012
	PLACE.OBJECT.PL	43008010
	PLACE.GOAL	43013036
	PLACE.FROM.ORIGIN	66007013b
	PLACE.FROM.SOURCE	40013054a
PERSON	PERSON.ROLE.AGENT	41005030
	PERSON.ROLE.PATIENT	43018004
	PERSON.ROLE.GOAL	43006068
	PERSON.ASCRIPTION	40016013
	PERSON.ROLE.RECIPIENT	58003018
	PERSON.SELECTION.MULTIPLE	42010036b
	PERSON.SELECTION.TWO	42022027
	PERSON.POSSESSOR	42020024a
	PERSON.IDENTITY.2SG	43021012
	PERSON.IDENTITY.3SG	40021010
	PERSON.IDENTITY.PL	40012048b
	PERSON.KIND	46003005b
THING	THING.PATIENT	41006024
	THING.THINK	40022042a
	THING.DO	42003010
	THING.SAY	45009030
	THING.HAPPEN	42018036
	THING.SELECTION.MULTIPLE.PL	40019018
	THING.SELECTION.MULTIPLE.SG	40022036
	THING.SELECTION.TWO	41002009
	THING.KIND	43002018b
	QUANTITY.MASS	42016007
INTENTION	INTENTION	42006002
MANNER/EXTENT	MANNER	41004013
	MANNER.STATEMENT	43006042
	QUANTITY.COUNT	40015034
	QUANTITY.FREQUENCY	40018021

Table 5.1: Representative contexts of 38 sub-clusters

It has to be noted that the derivational direction between two contexts is not necessarily one-way. Instead, both directions are possible. When one context is considerably more frequently derived from the other one across the sampled languages, this direction is correspondingly regarded as dominant and exhibited in the derivational map. Since this chapter intends to broadly sketch the derivational relations between identified contexts, only the frequent connections with an obvious direction are drawn in the maps below. This means, even though a pair of contexts are not linked in the graph, they could still be derivationally

related, which is just too weak to be presented here or there is no obvious directionality between them.

Also, the derivations deduced in this study are based on the interrogatives applied in 38 representative contexts. The use of these interrogatives is purely the decision of the translators of the Bible, which may be divergent from the description in grammars. Besides, considering that the interrogatives are directly extracted from authentic discourse, elements carrying other grammatical information are very likely involved in the derivational analysis. Compared to the strictly grammatical examples in references, the derivations found here more closely mirror the interrogatives in practical use.

5.2 Derivations within TIME

The following Figure 5.1 shows the derivations between the contexts of the cluster TIME.

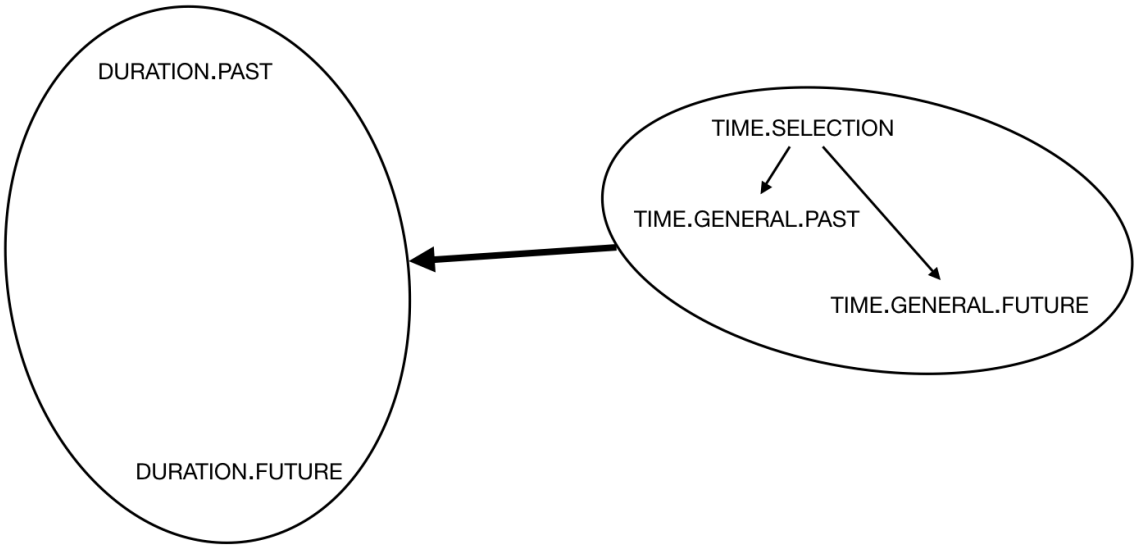

Figure 5.1: Derivations within TIME

As can be seen on the circle on the right, TIME.SELECTION appears to be the source of TIME.GENERAL.PAST and TIME.GENERAL.FUTURE. In contrast, no clear derivational direction can be recognized between TIME.GENERAL.PAST and TIME.GENERAL.FUTURE in the sampled languages. Therefore, they are not visually linked in the map. As mentioned in the last section, such a case does not indicate that the interrogatives attested in these two contexts are completely independent. Instead, the number of the sampled languages in which

TIME.GENERAL.FUTURE is derived from TIME.GENERAL.PAST is very close to the reverse case. In this sense, we can only tell that these two contexts are related to each other. The same situation is also found between DURATION.PAST and DURATION.FUTURE within the circle on the left.

It is conspicuous that the three contexts on the right in Figure 5.1, i.e., TIME.SELECTION, TIME.GENERAL.FUTURE and TIME.GENERAL.PAST, all serve as the main source to build the structures of DURATION.PAST and DURATION.FUTURE. The derivation from the three source contexts to DURATION.FUTURE is stronger than DURATION.PAST. Table 5.2 below provides examples demonstrating the derivations described above.

Derivation	**Language**	**Example**
SELECTION → GENERAL.PAST	Baoulé	*ônin → tyen ônin*
SELECTION → GENERAL.FUTURE	Baoulé	*ônin → ble ônin*
GENERAL&SELECTION → DURATION.PAST	Maasina Fulfulde	*mande → gila mande*
GENERAL&SELECTION → DURATION.FUTURE	Japanese	いつ → いつまで

Table 5.2: Examples of derivations within TIME

5.3 Derivations within PLACE

The derivational links within the cluster PLACE are illustrated in Figure 5.2 below. Six PLACE-related contexts can be classified into three groups. At the bottom left, PLACE.EVENT, PLACE.OBJECT.PL and PLACE.OBJECT.SG all query the stative condition of the questioned target, as discussed in §4.3 before. Among them, PLACE.OBJECT.PL seems to more frequently act as the source of the other two contexts. The contexts of PLACE.EVENT and PLACE.OBJECT.SG are also related with each other in form, whereas there lacks enough evidence to suggest any derivational preference.

Two contexts involving the direction 'from' are found at the bottom right. Similar to PLACE.EVENT and PLACE.OBJECT.SG, it is difficult to tell which of this pair serves as the derivational source of the other. However, it is certain that PLACE.FROM.ORIGIN and PLACE.FROM.SOURCE are often derived from the three contexts of the stative location. The same derivation is also found from PLACE.OBJECT & PLACE.EVENT to PLACE.GOAL whose

label is located at the top in Figure 5.2. Table 5.3 below exemplifies the derivations within PLACE.

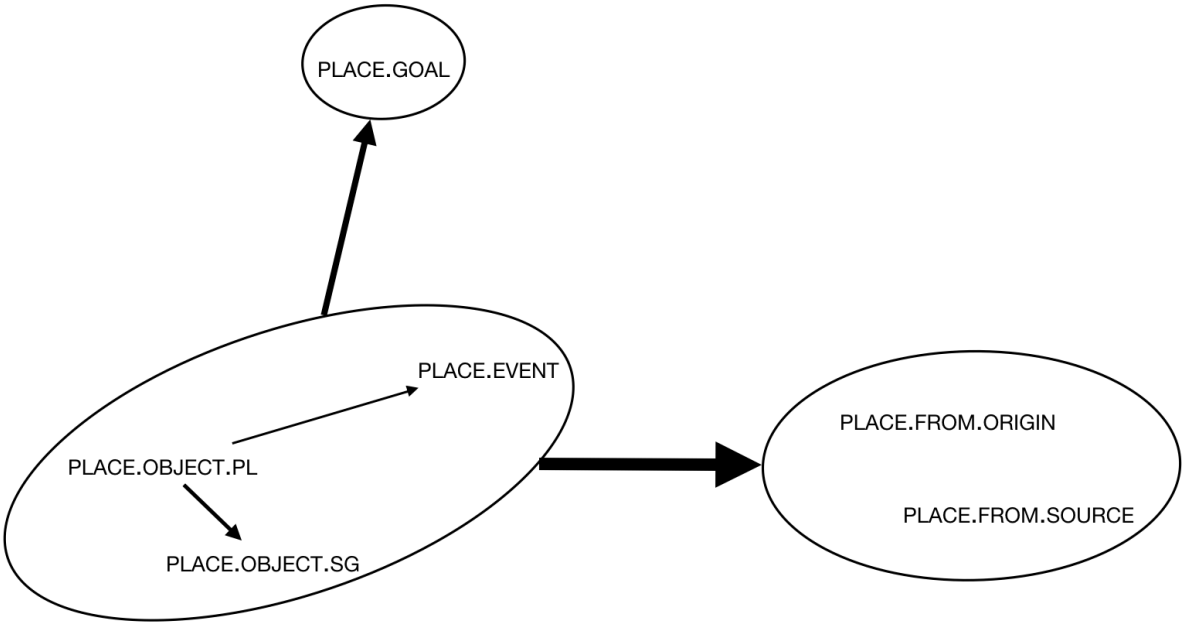

Figure 5.2: Derivations within PLACE

Derivation	Language	Example
OBJECT.PL → OBJECT.SG	Masaaba	*ena → waheena*
OBJECT.PL → EVENT	Makarsa	*kemae → kemaeki*
OBJECT&EVENT → GOAL	Spanish	*dónde → adónde*
OBJECT&EVENT → FROM	Romanian	*unde → de unde*

Table 5.3: Examples of derivations within PLACE

5.4 Derivations within PERSON

The cluster PERSON has the highest number of identified sub-clusters of which the attested interrogatives are closely related. Correspondingly, the derivational relations within PERSON are remarkably complicated, as displayed in the following Figure 5.3.

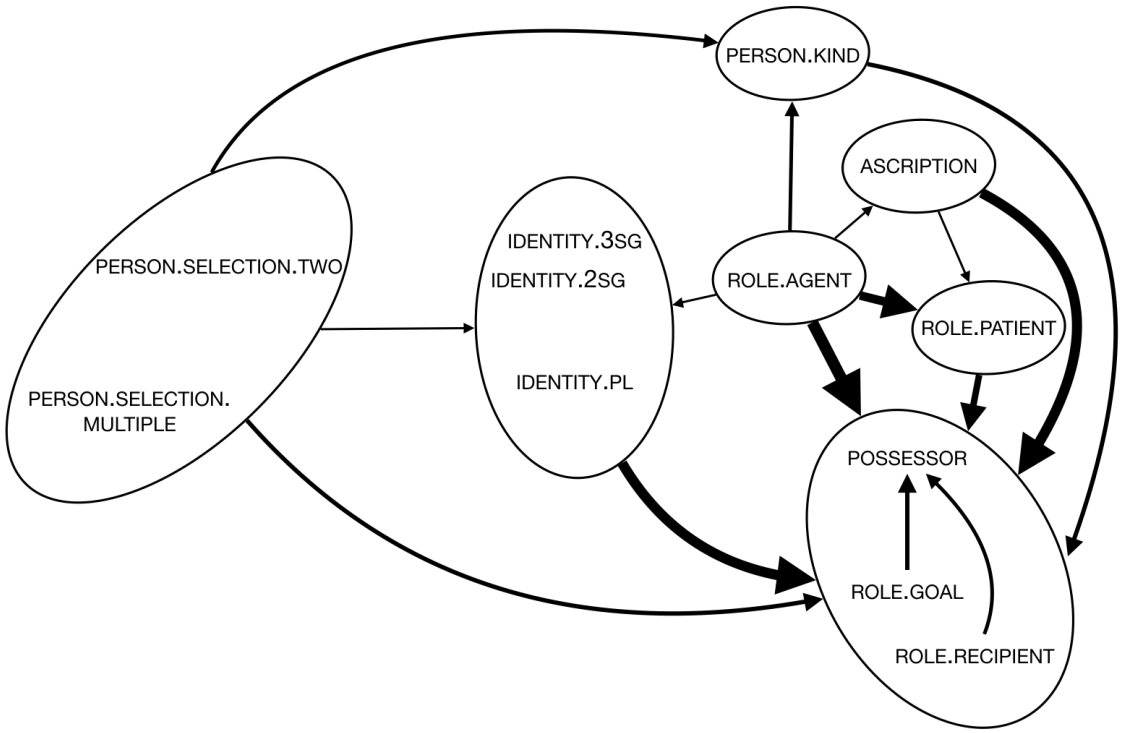

Figure 5.3: Derivations within PERSON

Two contexts of PERSON.SELECTION are located within the circle on the left. They are morphologically related, whereas there does not show an ascertained derivational direction between them. The PERSON.SELECTION-group serves commonly as the source of the other PERSON-related contexts. One of the notable connections from PERSON.SELECTION heads towards PERSON.KIND whose label is located at the top in Figure 5.3. Similar to PERSON.SELECTION, PERSON.KIND can also take part in the derivation as the start point. A relatively strong link from PERSON.KIND is pointing to the group composed of PERSON.POSSESSOR, PERSON.ROLE.GOAL and PERSON.ROLE.RECIPIENT that will be described subsequently.

The context PERSON.IDENTITY.3SG, PERSON.IDENTITY.2SG and PERSON.IDENTITY.PL gather at the center. Similar to PERSON.SELECTION, there exist internal derivations within this group but without obvious derivational direction. These three PERSON.IDENTITY-related contexts are found frequently as one of the important origins of the POSSESSOR-GOAL-RECIPIENT group as well.

On the right side of the map, the derivational relations appear to be intricate. Four circles assemble here, i.e., PERSON.ROLE.AGENT, PERSON.ASCRIPTION, PERSON.ROLE.PATIENT and the group of PERSON.POSSESSOR, PERSON.ROLE.GOAL and PERSON.ROLE.RECIPIENT. The context of

PERSON.ROLE.AGENT commonly initiates a derivation to the other members. The goals of two intense relations from PERSON.ROLE.AGENT are PERSON.ROLE.PATIENT and the POSSESSOR-GOAL-RECIPIENT group, respectively. Considering that PERSON.ASCRIPTION is not specially marked in the majority of sampled languages and is asked with the general form of PERSON, it is understandable that this context also plays as the source of PERSON.ROLE.PATIENT and POSSESSOR-GOAL-RECIPIENT group. Except for being the endpoint, PERSON.ROLE.PATIENT can also act as the source of POSSESSOR-GOAL-RECIPIENT group.

Finally, POSSESSOR-GOAL-RECIPIENT group is always found at the end of a derivational development from all other PERSON-related contexts. Within this group, POSSESSOR is consistently more derived from GOAL and RECIPIENT. Table 5.4 below provides examples of the derivational links within PERSON.

Derivation	**Language**	**Example**
SELECTION → POSSESSOR-GOAL-RECIPIENT	Gagauz	*kim → kimin/kimä*
SELECTION → IDENTITY	Tetelcingo Nahuatl	*öque → öquenu/öquemeju*
SELECTION → KIND	Western Arrarnta	*ngunha → ngunhama*
KIND → POSSESSOR-GOAL-RECIPIENT	Hopi	*hak → hakiy/hakimuy*
IDENTITY → POSSESSOR-GOAL-RECIPIENT	Khalkha	*hen → hend/henii*
AGENT → IDENTITY	Makasar	*inai → inaiki'/inaika*
AGENT → KIND	Wolof	*ku → kuy*
AGENT → PATIENT	Turkish	*kim → kimi*
AGENT → POSSESSOR-GOAL-RECIPIENT	Uyghur	*kim → kimge/kimlerni/kimnig*
AGENT → ASCRIPTION	Korean	누구 → 누구에게
ASCRIPTION → PATIENT	Parauk	*mawx → pui mawx*
ASCRIPTION → POSSESSOR-GOAL-RECIPIENT	Nyanja	*yani → ayani/nchayani*
PATIENT→ POSSESSOR-GOAL-RECIPIENT	Ejagham	*énê → bhaghé énê*
GOAL → POSSESSOR	Noon	*ɓa → cuuɓa*
RECIPIENT → POSSESSOR	Vietnamese	*ai → của ai*

Table 5.4: Examples of derivations within PERSON

5.5 Derivations within THING

The internal derivations of the cluster THING is presented in Figure 5.4 below. Although the context QUANTITY.MASS has been assigned into this primary cluster, this concept is semantically more associated to the other QUANTITY contexts of the primary cluster MANNER/EXTENT.[28] Thus, the derivation relations involving QUANTITY.MASS will be presented in the map of MANNER/EXTENT in §5.6.

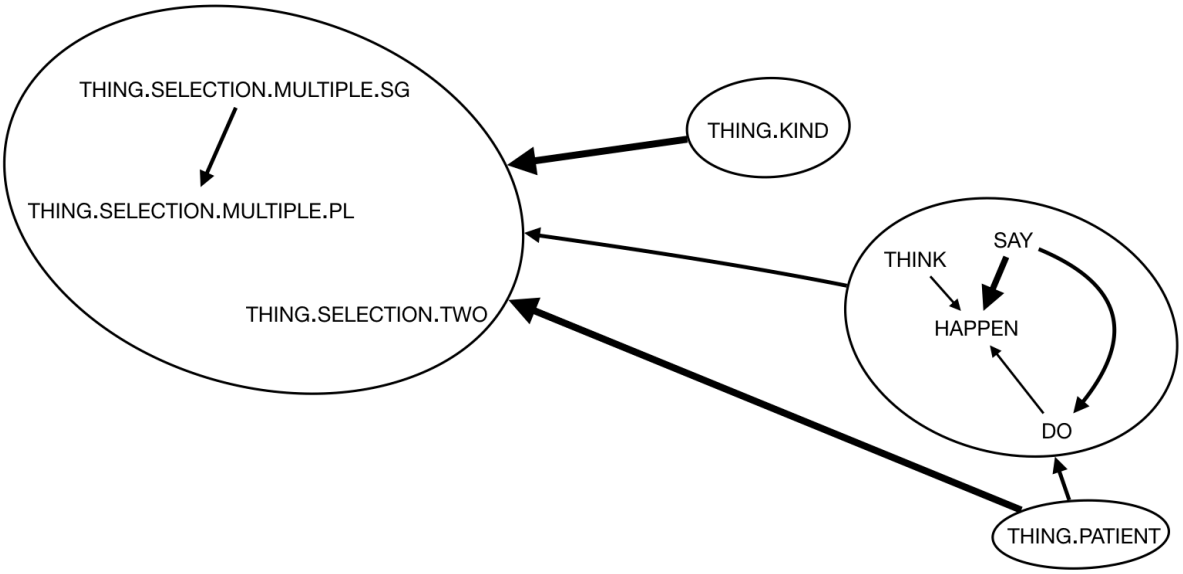

Figure 5.4: Derivations within THING

Three contexts related to THING.SELECTION are found within the circle on the left in Figure 5.4. Among these three contexts, a clear derivation starts from THING.SELECTION.MULTIPLE.SG to THING.SELECTION.MULTIPLE.PL. The context THING.KIND is located at the center of Figure 5.4. It serves as one of the sources of THING.SELECTION.

Four contexts assemble within the circle on the right in Figure 5.4. They all contain questions asking for a concrete action. In this sense, this group is given the label THING.VERB. Among THING.VERB, THING.HAPPEN is always found as the derivational endpoint. There is another development commencing from THING.SAY to THING.DO. This group acts as a source of THING.SELECTION as well. At the bottom right, THING.PATIENT appears to be the origin of THING.VERB and THING.SELECTION. Table 5.5 below illustrates the derivations within THING with selected examples.

[28] The explanation of this clustering result has been given in §4.5.9.

Derivation THING	Language	Example
SELECTION.MULTIPLE.SG → SELECTION.MULTIPLE.PL	Catalan	*quin → quins*
KIND → SELECTION	Balantak	*upa → koi upa*
PATIENT → SELECTION	Cherokee	*gado → gado usdi*
VERB → SELECTION	Parecís	*xoare → xoarenai/xoarehare*
PATIENT → THING.VERB	Balinese	*napi → punapike*
SAY → DO	Dogrib	*ayìı → ayìı dàts'ıı*
SAY → HAPPEN	Tenharim-Parintintin-Diahoi	*marã → maraname*
DO → HAPPEN	Turkmen	*näme → nämedigini*
THINK → HAPPEN	Hopi	*hin → hintaqamuy*

Table 5.5: Examples of derivations within THING

5.6 Derivations within MANNER/EXTENT

Figure 5.5 below displays the internal derivations of the cluster MANNER/EXTENT. Considering that QUANTITY.MASS is noticeably connected to the four sub-clusters of MANNER/EXTENT, this context is included in the current category.

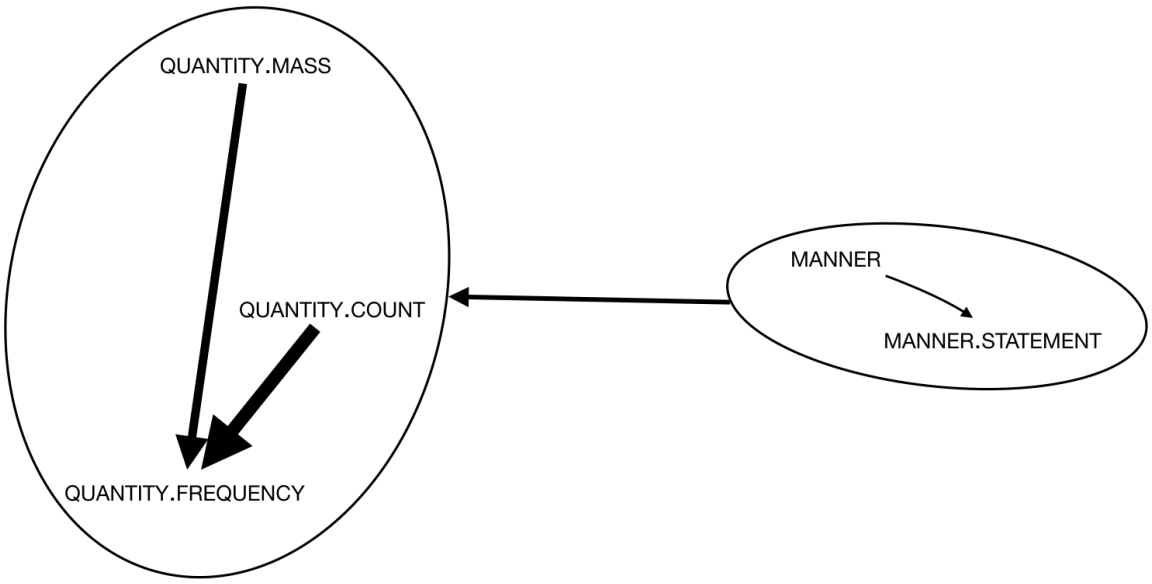

Figure 5.5: Derivations within MANNER/EXTENT

As shown in the map, five contexts are distributed into two groups in accordance with the similarities between them. In the group on the right side, MANNER.STATEMENT is more derived from MANNER. The other group at the left comprises three contexts related to QUANTITY, i.e., QUANTITY.COUNT, QUANTITY.MASS and QUANTITY.FREQUENCY. A significant strong connection begins from QUANTITY.COUNT to QUANTITY.FREQUENCY. Furthermore, QUANTITY.FREQUENCY is also commonly derived from QUANTITY.MASS. Between the two groups in Figure 5.5, there exists an apparent derivation from MANNER & MANNER.STATEMENT to QUANTITY. Examples displaying these links are given in the following Table 5.6.

Derivation	**Language**	**Example**
MANNER → MANNER.STATEMENT	Dogrib	*dànì → dànìghǫ*
COUNT → FREQUENCY	Welsh	*sawl → sawl gwaith*
MASS → FREQUENCY	German	*wieviel → wievielmal*
MANNER&MANNER.STATEMENT → QUANTITY	English	*how → how many*

Table 5.6: Examples of derivations within MANNER/EXTENT

5.7 Derivation across categories

Apart from the internal derivations within each primary cluster, it is common that the interrogative of a certain context is derived from another semantic category. Cysouw (2005b: 2) portrays the connections between nine major classes. On this basis, Hölzl (2018: 83) adds some links involving two finer categories, i.e., KIND and ACTIVITY. Based on the attested interrogatives of this study, the following Figure 5.6 summarizes the derivations across six primary categories.

As can be seen in Figure 5.6, the interrogatives used for THING are recurrently found as the source of the other classes whereby the derivation from THING to INTENTION is considerably strong. Apart from having THING as the origin, the derivations to TIME often start from MANNER/EXTENT and PLACE as well. Compared to THING and PLACE, INTENTION, MANNER/EXTENT and TIME are more often derived from the other categories.

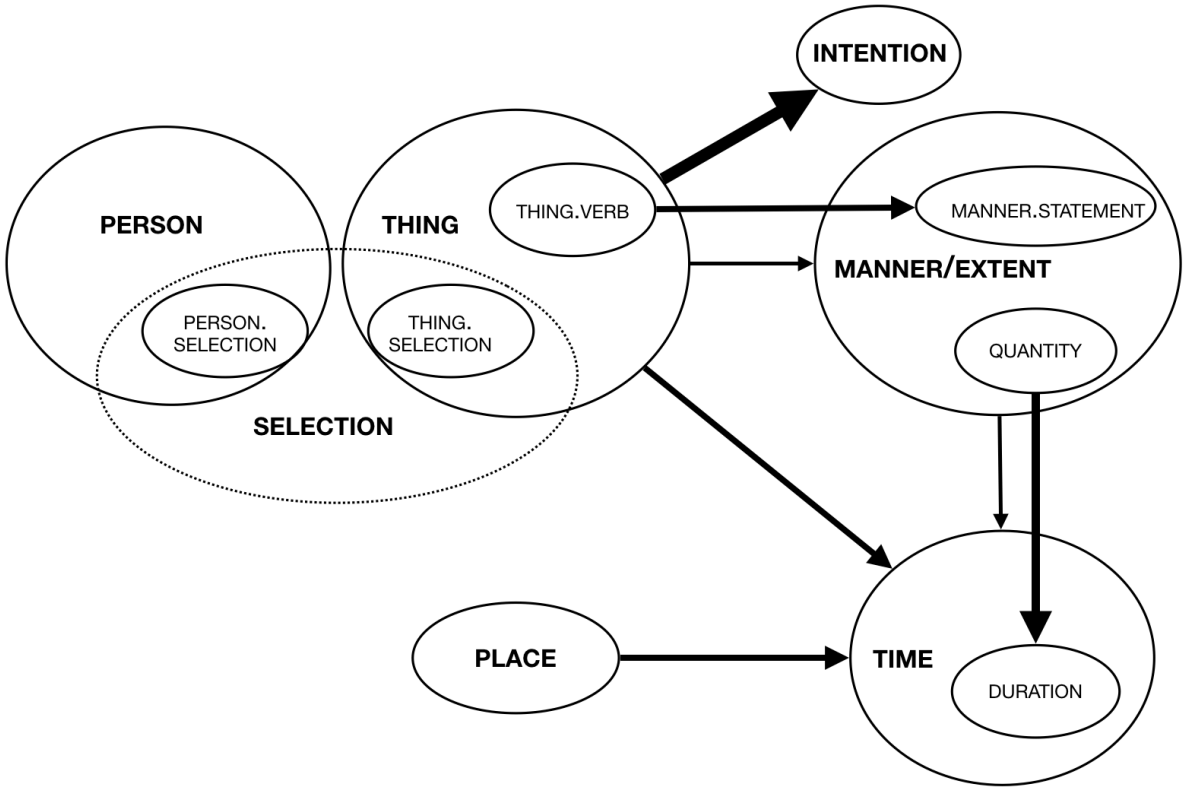

Figure 5.6: Derivations across categories

At the top left in Figure 5.6, PERSON seems to rarely show an explicit derivational relation with any other class. Nonetheless, a substantially high similarity occurs between PERSON.SELECTION and THING.SELECTION. This is not in accordance with the quantitative analysis in Chapter 4 which did not suggest a separate cluster for SELECTION. The main factor in this quantitative result, as has been discussed, is that the general forms for PERSON and THING are also applicable to SELECTION-related questions in many sampled languages. Besides, some languages even do not have a specialized structure for SELECTION.

However, a hypothesis of a separate cluster for SELECTION can be proposed through the manual inspection of the data, as indicated by the dotted circle in Figure 5.6. The reason for this hypothesis is that a number of sampled languages do not differentiate the interrogative for these two questions and often mark them with the identical form particularly denoting a choice made from a set of options, e.g., *which* in English and *welch-* in German. This resemblance is also in line with the fact that SELECTION, regardless of whether the referent is human or non-human, is usually taken as an independent primary category in many previous research and descriptive grammars.

Besides the parallelism between PERSON.SELECTION and THING.SELECTION, two other connections in Figure 5.6 are also identified between a pair of sub-clusters of different categories. The contexts of MANNER/EXTENT can be further divided into two groups, i.e., MANNER & MANNER.STATEMENT and QUANTITY, as noted in §5.6. The contexts of QUANTITY manifest particularly notable derivations to DURATION subsumed under TIME. The sub-cluster MANNER.STATEMENT appears to be often derived from the four questions of THING.VERB that are encircled at the bottom center in Figure 5.4 above. Despite being classified into two primary categories, this five contexts are all involved with concrete actions for which the interrogative for MANNER, e.g., *how* in English, is employable in many sampled languages. Table 5.7 below presents these pivotal derivations across categories with examples.

Derivation	**Language**	**Example**
THING → INTENTION	Irish	*cad* → *cad chuige*
THING → TIME	Mandarin	什麼 → 什麼時候
THING → MANNER/EXTENT	Tagalog	*ano* → *paano*
THING.SELECTION = PERSON.SELECTION	Batak Karo	*apai* = *apai*
MANNER/EXTENT → TIME	Garifuna	*ída* → *ída bugagí*
THING.VERB → MANNER&STATEMENT	Japanese	どう → どうして
QUANTITY → DURATION	Toro So Dogon	*yagɔ baa* → *waaru yagɔ baa*

Table 5.7: Examples of derivation across categories

6 Conclusion

This chapter will summarize the clustering results of interrogative contexts presented in this study and outline some prospects for future research about content interrogatives. In §6.1, I will first summarize the clustering results with discussion in more detail. Considering the remarkable semantic similarities and the derivational connections between certain categories and sub-clusters, as shown in Chapter 4 and Chapter 5, the categorization of some interrogative sub-clusters is slightly adjusted, which will also be explained in this section. Then, several interesting aspects that have not yet been addressed in this research will be proposed in §6.2.

6.1 Summary of the clustering results

This study has explored the semantic diversity of content interrogatives and attempted to identify the representative context of each interrogative category. The algorithm *Partitioning Around Medoids* (PAM) successfully classified the contexts into various groups according to the similarities between the interrogative codings attested in 90 biblical translations. In total, six primary clusters and thirty-eight sub-clusters[29] are identified. In correspondence with semantic properties, each grouping is assigned a label for the interpretation.

6.1.1 Primary clusters

As the result of the first level of clustering, 413 contexts are grouped into six primary clusters, i.e., TIME, PLACE, PERSON, THING, INTENTION and MANNER/EXTENT. The identification of primary clusters is congruous with the interrogative inventory provided in the descriptive grammars of most languages. The grouping is decided by the maximal formal differentiations and is supposed to imply the underlying semantic distinctions between these interrogative categories across languages.

Notwithstanding, the automatic decision involving contexts of SELECTION is conspicuously different from most grammatical descriptions. In the references of many languages, the domain of SELECTION is commonly deemed to be a major interrogative category for which

[29] Given that the existence of sub-cluster INTENTION.PURPOSE is uncertain, as argued in §4.6.3, this possible sub-cluster is not included in the final results.

specialized codings can be found, e.g., *which* in English and *welch-* in German. However, according to the first level of clustering in this study, the interrogative contexts relevant to this domain do not independently constitute a primary class. Instead, they are assigned separately to the cluster PERSON and THING. Via the internal categorization, they are then divided into two sub-clusters, i.e., PERSON.SELECTION (see §4.4.5) and THING.SELECTION (see §4.5.7). According to the unique interrogative forms attested in some languages, a more subtle number distinction between two and multiple options is further manually drawn.

Even though the relevant contexts are separately assigned to the category PERSON and THING, the occurrence of the dedicated form for SELECTION can always be found in PERSON.SELECTION and THING.SELECTION, as shown in Figure 6.1 below. Through the specific coding, the semantic meanings of PERSON.SELECTION and THING.SELECTION are easily interpreted and distinguished from other contexts of PERSON and THING. Besides, if SELECTION is taken as an independent category, as done in many grammars, the emergence of these two sub-categories discloses a pair of more sophisticated domains of SELECTION that are semantically related to the human vs. non-human distinction.

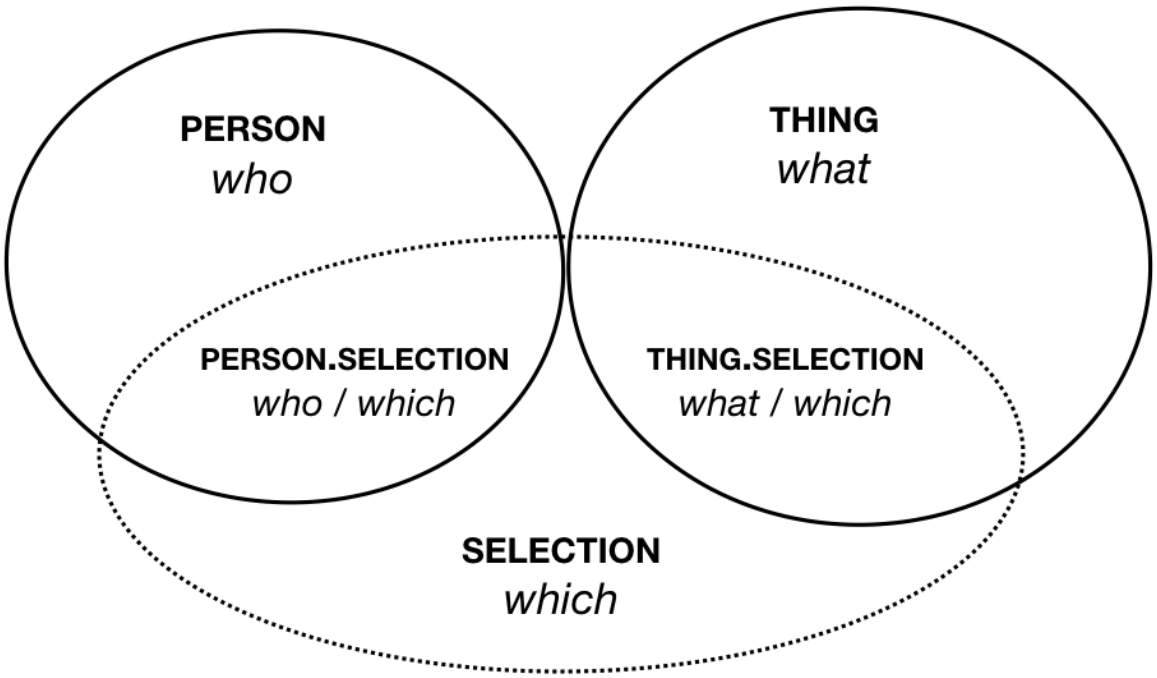

Figure 6.1: Domain of SELECTION

The partition of contexts belonging to SELECTION results from the hybrid usage of the dedicated form of SELECTION and the interrogatives for PERSON and THING. This outcome aligns with the view of Idiatov (2009: 7-8) that the distinctive line between selective and non-selective contexts is sometimes difficult to draw. Furthermore, he argues how the questioner conceptualizes the real referent plays a more important role than the essence of the referent. In this sense, it can be expected that the way to inquire about SELECTION differs across languages as well as practical contexts.

6.1.2 Sub-clusters

Thirty-eight sub-clusters with a total of 177 contexts are generated from the second level of clustering within six primary categories.[30] With regard to the other 236 interrogative contexts, the second level of clustering failed to assigned them to an interpretable sub-cluster. Given that each primary cluster has a different number of contexts, the results of the internal clustering vary. The following Table 6.1 provides a summary of sub-clusters.

To be noticed, in order to emphasize the close connection and the common practice of being viewed as an independent category, as just discussed, the five sub-clusters of PERSON.SELECTION and THING.SELECTION are rearranged into the major class SELECTION in Table 6.1. The same case is found with the sub-cluster QUANTITY.MASS. As has been remarked in §5.5 and §5.6 before, despite being assigned to the category THING, QUANTITY.MASS is semantically more connected to the other two QUANTITY-related sub-clusters that are subsumed under the category MANNER/EXTENT. Given this noteworthy link, QUANTITY.MASS is moved to the category MANNER/EXTENT in Table 6.1.

As can be seen in Table 6.1, the number of sub-clusters within PERSON and THING is considerably larger than the other primary clusters. This might reflect that, within these two categories, there are more subtle semantic differentiations for which languages apply unique constructions. Yet, it should be noticed that the small amount of the identifiable sub-clusters within the other primary categories does not necessarily indicate the lesser internal semantic division. Rather, the lack of available relevant contexts in the Bible could be the chief factor, e.g., questions asking for information about TIME and PLACE.

[30] The content of these 177 contexts can be found in Appendix B.

During the identification and interpretation of the results of the internal grouping, some difficulties arise in terms of INTENTION and MANNER/EXTENt. Although these two clusters also contain a reasonable number of contexts, only a few sub-domains for MANNER/EXTENT and even no convinced one for INTENTION can be ascertained through the second level of clustering. Yet, the cause of this situation differs between these two primary clusters.

Primary cluster	Sub-cluster	Primary cluster	Sub-cluster
TIME	TIME.GENERAL.PAST TIME.GENERAL.FUTURE TIME.SELECTION DURATION.FUTURE DURATION.PAST	**PLACE**	PLACE.EVENT PLACE.OBJECT.SG PLACE.OBJECT.PL PLACE.GOAL PLACE.FROM.ORIGIN PLACE.FROM.SOURCE
INTENTION	INTENTION	**MANNER/EXTENT**	MANNER MANNER.STATEMENT QUANTITY.MASS QUANTITY.COUNT QUANTITY.FREQUENCY
PERSON	PERSON.ROLE.AGENT PERSON.ROLE.PATIENT PERSON.ROLE.RECIPIENT PERSON.ROLE.GOAL PERSON.ASCRIPTION PERSON.IDENTITY.2SG PERSON.IDENTITY.3SG PERSON.IDENTITY.PL PERSON.POSSESSOR PERSON.KIND	**THING**	THING.PATIENT THING.THINK THING.DO THING.SAY THING.HAPPEN THING.KIND
SELECTION	PERSON.SELECTION.TWO PERSON.SELECTION.MULTIPLE		THING.SELECTION.TWO THING.SELECTION.MULTIPLE.PL THING.SELECTION.MULTIPLE.SG

Table 6.1: Summary of the clustering results

In respect of MANNER/EXTENT, it is not rare across languages that a coding, e.g., *how* in English, can carry multiple meanings. The identical form leads to the difficulty of categorization of contexts and recognition of semantic differentiations. Therefore, the outcome of the internal clustering of this primary cluster is scant. The typical contexts of MANNER have to be manually examined and selected.

An opposite case occurs with the cluster INTENTION. It is common that a language has various constructions to query the intention of an action, while the functional distinction between these forms is seldom specified in grammars. One possible difference is, for example, a few languages tell apart GOAL 'for what' from CAUSE 'because of what'.

According to the attested forms, the second level of clustering suggested that the contexts subsumed under INTENTION should be further divided into 14 sub-groups. Nevertheless, it is bewildering to interpret the specific meaning of each based on the applied interrogative codings. Regardless of the potentially existing semantic distinction, miscellaneous interrogative forms referring to INTENTION may be purely alternatives for one another in a language, which does not involve differences in meaning. It is also possible that the functional nuances are too subtle to be formally marked in sampled languages.

For most sub-clusters in Table 6.1, their existence is justified by the dedicated interrogative forms found in one or some certain languages, as has been manifested in Chapter 4. These forms are frequently related to a general interrogative coding. They can be derived from the basic coding of the corresponding primary category, e.g., *wohin* 'where to' developed from *wo* 'where' and denoting PLACE.GOAL in German. In addition, they can also be inflected or carry a marker explicitly indicating a grammatical attribute, e.g., *whose* for PERSON.POSSESSOR in English and *cuáles* 'which.PL' for THING.SELECTION.MULTIPLE.PL in Spanish. Apart from the formal connection, languages can also build a unique structure specifically for some minor categories, especially those involving verbal concepts or copula expressions, e.g., THING.DO and PLACE.OBJECT. Given that the specialized form can sometimes only be found in a few or even one language, some sub-clusters cannot be detected by the automatic grouping. In this situation, they are identified through manual examination. Sub-clusters of this type are, for instance, TIME.PAST, TIME.FUTURE, TIME.SELECTION and QUANTITY.FREQUENCY.

6.1.3 Subsidiary concept

Nevertheless, there are also cases that the internal clustering suggests the existence of a sub-cluster for which no specific coding is attested in any sampled languages. Three sub-clusters pertain to this kind, i.e., THING.THINK, MANNER.STATEMENT and PLACE.FROM.SOURCE. In respect of the attested forms adopted for these sub-clusters, it shows a mixture of various categories, i.e., THING 'what' and MANNER 'how' for THING.THINK, INTENTION 'why' and MANNER 'how' for MANNER.STATEMENT, and PLACE.EVENT/FROM.ORIGIN 'where/where from' and MANNER 'how' for PLACE.FROM.SOURCE. Since no special coding can assist the interpretation, the meaning of these sub-clusters can only be extrapolated based on the textual content.

The fact that there is no separate lexicalization of these three sub-clusters makes them distinct from all other sub-categories for which a specialized form can be found in at least one sampled language. The lack of formal evidence causes the difficulty in solidly delineating their semantic domains. This suggests that these sub-clusters might not be 'real' sub-categories but only subsidiary in between certain primary categories. For this case, I propose the name **subsidiary concept** for this kind of sub-cluster. Three subsidiary concepts occurring in this study are visualized in Figure 6.2 below.

However, subsidiary concepts are not identical to their both neighboring concepts either. The forms of both categories are exchangeable for these subsidiary concepts, which marks them different from other sub-categories of each side. In this sense, it seems not far-fetched that there could exist separate lexicalization for these subsidiary concepts in some languages that are not studied here.

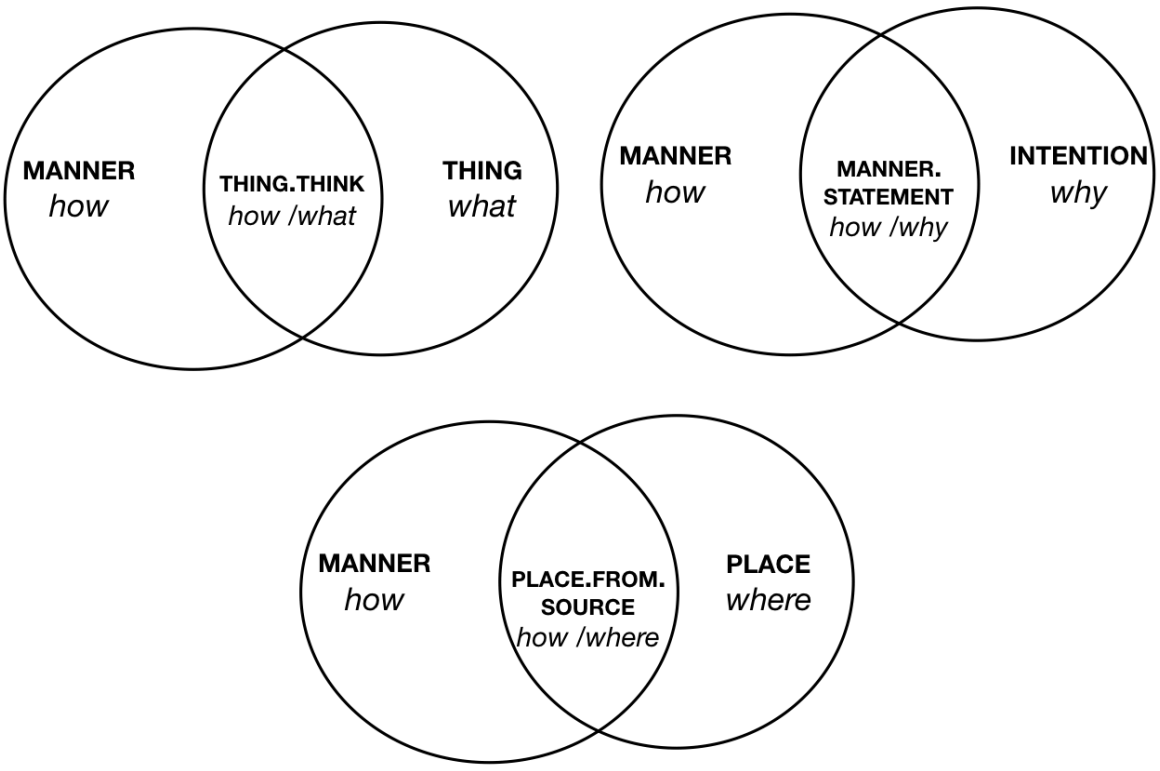

Figure 6.2: Domain of subsidiary concepts

6.2 Prospects

This study has empirically investigated the subtle divisions within interrogative categories. Many topics still remain undiscussed. For example, due to the limited time, only a small amount of languages have been taken from each family as the sample. The results presented in the preceding chapters only demonstrate the distribution of interrogative contexts across sampled languages. It is difficult to sketch any family-particular or areal pattern for codings of the interrogative clusters and sub-clusters. In future research with a larger and more balanced language sample, it is expected to draw the characteristics of each language family. Moreover, a language could be singled out to represent the typical pattern of each family.

Also, when there are enough grammatical references, the diachronic development of content interrogatives could be more sufficiently considered, especially during the morphological analysis (see §3.4.3). With the evidence of historically shared elements, the semantic relations between interrogative contexts can be better revealed. Besides, this would help to further identify derivational directions between interrogative categories and sub-categories. It will also be interesting to see whether a certain derivation solely occurs in a few language families or is recurrently attested.

Finally, it is expected to utilize the identified interrogative categories and sub-clusters to supplement the interrogative paradigm of languages with lesser grammatical information or those with a compact interrogative system. Besides, according to descriptive grammars, there are languages with an 'economical' interrogative system or a small-size inventory of content interrogatives.[31] By examining the elements present in corresponding contexts in the biblical translation, we might be able to find out whether there exists any grammatical device to further tell apart the sophisticated interrogative meanings.

[31] Take Ewe as an example. As argued by Aikhenvald (2015: 236), this language has "one of the most economical systems in the world" in which there are only two question markers. One is the particle *ka* which is added after a noun phrase, e.g., *afi-ka* 'where' (lit. *place-ka*), while the other question word is *néne* 'how many' (Ameka 1991: 53-54).

Appendix A: Sampled languages and used Bible translations

The Table below presents information about the sampled languages of this study. Language abbreviations in ISO 639-3 standard are given in the second column **Code**. To be noticed, there are two translational versions for English and German, respectively. Although the ISO 639-3 code of Mandarin Chinese is **cmn** and of Yue Chinese/Cantonese **yue**, the name of the Bible translations in these two languages are both coded with **zho**.

Language	Code	Family	Bible translation
Acehnese	ace	Austronesian	*ace-x-bible*
Ayacucho Quechua	quy	Quechuan	*quy-x-bible*
Balantak	blz	Austronesian	*blz-x-bible*
Balinese	ban	Austronesian	*ban-x-bible*
Baoulé	bci	Niger-Congo	*bci-x-bible*
Batak Karo	btx	Austronesian	*btx-x-bible*
Burarra	bvr	Maningrida	*bvr-x-bible*
Car Nicobarese	caq	Austro-Asiatic	*caq-x-bible*
Catalan	cat	Indo-European	*cat-x-bible-evangelica*
Central Yupik	esu	Eskimo-Aleut	*esu-x-bible*
Chamorro	cha	Austronesian	*cha-x-bible-2003*
Cherokee	chr	Iroquoian	*chr-x-bible*
Chuj	cac	Mayan	*cac-x-bible-sansebastian*
Coatlán Mixe	mco	Mixe-Zoque	*mco-x-bible*
Croatian	hrv	Indo-European	*hrv-x-bible-2000*
Czech	ces	Indo-European	*ces-x-bible-living*
Danish	dan	Indo-European	*dan-x-bible-hverdagsdansk*
Dii	dur	Niger-Congo	*dur-x-bible*
Dogrib	dgr	Eyak-Athabaskan	*dgr-x-bible*
Dutch	nld	Indo-European	*nld-x-bible-2007*
Eastern Bru	bru	Austro-Asiatic	*bru-x-bible*
Ejagham	etu	Niger-Congo	*etu-x-bible*

Language	Code	Family	Bible translation
El Nayar Cora	crn	Uto-Aztecan	*crn-x-bible*
English	eng	Indo-European	*eng-x-bible-common* *eng-x-bible-goodnews*
Finnish	fin	Uralic	*fin-x-bible-1992*
Francisco León Zoque	zos	Mixe-Zoque	*zos-x-bible*
Gagauz	gag	Altaic	*gag-x-bible-latin*
Garifuna	cab	Arawakan	*cab-x-bible*
German	deu	Indo-European	*deu-x-bible-schlachter* *deu-x-bible-newworld*
Gwich'in	gwi	Eyak-Athabaskan	*gwi-x-bible*
Halh Mongolian	khk	Altaic	*khk-x-bible*
Highland Popoluca	poi	Mixe-Zoque	*poi-x-bible*
Hopi	hop	Uto-Aztecan	*hop-x-bible*
Huallaga Huánuco Quechua	qub	Quechuan	*qub-x-bible*
Hungarian	hun	Uralic	*hun-x-bible-2003*
Iban	iba	Austronesian	*iba-x-bible*
Icelandic	isl	Indo-European	*isl-x-bible*
Igbo	ibo	Niger-Congo	*ibo-x-bible*
Indonesian	ind	Austronesian	*ind-x-bible-firman*
Inga	inb	Quechuan	*inb-x-bible*
Irish	gle	Indo-European	*gle-x-bible*
Japanese	jpn	Isolate	*jpn-x-bible-colloquial*
Jarai	jra	Austronesian	*jra-x-bible*
Kagulu	kki	Niger-Congo	*kki-x-bible*
Karakalpak	kaa	Altaic	*kaa-x-bible-latin*
Korean	kor	Isolate	*kor-x-bible-1985*
Kuku-Yalanji	gvn	Pama-Nyungan	*gvn-x-bible*
Lowland Tarahumara	tac	Uto-Aztecan	*tac-x-bible*
Ma'anyan	mhy	Austronesian	*mhy-x-bible*

Language	Code	Family	Bible translation
Maasina Fulfulde	ffm	Niger-Congo	*ffm-x-bible*
Machiguenga	mcb	Arawakan	*mcb-x-bible*
Madurese	mad	Austronesian	*mad-x-bible*
Makasar	mak	Austronesian	*mak-x-bible*
Mandarin Chinese	cmn	Sino-Tibetan	*zho-x-bible-contemp*
Masaaba	myx	Niger-Congo	*myx-x-bible*
Nama	naq	Khoe-Kwadi	*naq-x-bible*
Nomaande	lem	Niger-Congo	*lem-x-bible*
Noon	snf	Niger-Congo	*snf-x-bible*
North Alaskan Inupiatun	esi	Eskimo-Aleut	*esi-x-bible*
North Saami	sme	Uralic	*sme-x-bible*
Northern Dagara	dgi	Niger-Congo	*dgi-x-bible*
Northern Kissi	kqs	Niger-Congo	*kqs-x-bible*
Nyanja	nya	Niger-Congo	*nya-x-bible*
Paraguayan Guaraní	gug	Tupian	*gug-x-bible*
Parauk	prk	Austro-Asiatic	*prk-x-bible*
Parecís	pab	Arawakan	*pab-x-bible*
Romanian	ron	Indo-European	*ron-x-bible-2006*
Rundi	run	Niger-Congo	*run-x-bible*
Sirionó	srq	Tupian	*srq-x-bible*
Spanish	spa	Indo-European	*spa-x-bible-newworld*
Tabasco Chontal	chf	Mayan	*chf-x-bible*
Tagalog	tgl	Austronesian	*tgl-x-bible-1996*
Tenharim-Parintintin-Diahoi	pah	Tupian	*pah-x-bible*
Tetelcingo Nahuatl	nhg	Uto-Aztecan	*nhg-x-bible*
Tharaka	thk	Niger-Congo	*thk-x-bible*
Toro So Dogon	dts	Niger-Congo	*dts-x-bible*
Totontepec Mixe	mto	Mixe-Zoque	*mto-x-bible*
Turkish	tur	Altaic	*tur-x-bible-newworld*

Language	Code	Family	Bible translation
Turkmen	tuk	Altaic	*tuk-x-bible-mukaddes*
Uyguhr	uig	Altaic	*uig-x-bible-latin*
Vietnamese	vie	Austro-Asiatic	*vie-x-bible-bandich*
Welsh	cym	Indo-European	*cym-x-bible-colloquial2013*
Western Arrarnta	are	Pama-Nyungan	*are-x-bible*
Western Huasteca Nahuatl	nhw	Uto-Aztecan	*nhw-x-bible*
Wolof	wol	Niger-Congo	*wol-x-bible*
Yine	pib	Arawakan	*pib-x-biblw*
Yucatec Maya	yua	Mayan	*yua-x-bible*
Yue Chinese/Cantonese	yue	Sino-Tibetan	*zho-x-bible-new*

Appendix B: Contexts of 38 sub-clusters

The following Table lists 177 interrogative contexts (*eng-x-bible-common*) that are respectively assigned to the 38 identifiable sub-clusters discussed in Chapter 4. The representative context of each sub-cluster (see Table 5.1 in Chapter 5) is marked in grey. They are used to present the derivational connections between interrogative categories and sub-clusters in Chapter 5.

Verse	Label	Cluster	Context
66006010a	DURATION.FUTURE	1	They cried out with a loud voice , " Holy and true Master , how long will you wait before you pass judgment ?
41009019b	DURATION.FUTURE	1	How long will I put up with you ? Bring him to me . "
42009041a	DURATION.FUTURE	1	Jesus answered , " You faithless and crooked generation , how long will I be with you
43010024	DURATION.FUTURE	1	The Jewish opposition circled around him and asked , " How long will you test our patience ? If you are the Christ , tell us plainly . "
41009021	DURATION.PAST	1	Jesus asked his father , " How long has this been going on ? " He said , " Since he was a child .
42021007a	TIME.GENERAL.FUTURE	1	They asked him , " Teacher , when will these things happen ?
42017020	TIME.GENERAL.FUTURE	1	Pharisees asked Jesus when God’s kingdom was coming . He replied , " God’s kingdom isn’t coming with signs that are easily noticed .
40025039	TIME.GENERAL.PAST	1	When did we see you sick or in prison and visit you ? ’
40025038	TIME.GENERAL.PAST	1	When did we see you as a stranger and welcome you , or naked and give you clothes to wear ?
40025037	TIME.GENERAL.PAST	1	“ Then those who are righteous will reply to him , ‘ Lord , when did we see you hungry and feed you , or thirsty and give you a drink ?
43006025	TIME.SELECTION	1	When they found him on the other side of the lake , they asked him , " Rabbi , when did you get here ? "
43001038b	PLACE.EVENT	2	" Rabbi (which is translated Teacher) , where are you staying ? "
43011034	PLACE.EVENT	2	He asked , " Where have you laid him ? " They replied , " Lord , come and see . "
42022009	PLACE.EVENT	2	They said to him , " Where do you want us to prepare it ? "

Verse	Label	Cluster	Context
40002004	PLACE.EVENT	2	He gathered all the chief priests and the legal experts and asked them where the Christ was to be born .
43006005	PLACE.EVENT	2	Jesus looked up and saw the large crowd coming toward him . He asked Philip , " Where will we buy food to feed these people ? "
42017037	PLACE.EVENT	2	The disciples asked , " Where , Lord ? " Jesus said , " The vultures gather wherever there's a dead body . "
66007013b	PLACE.FROM.ORIGIN	2	and where did they come from ? "
43019009	PLACE.FROM.ORIGIN	2	He went back into the residence and spoke to Jesus , " Where are you from ? " Jesus didn't answer .
40013054a	PLACE.FROM.SOURCE	2	When he came to his hometown , he taught the people in their synagogue . They were surprised and said , " Where did he get this wisdom ?
41006002a	PLACE.FROM.SOURCE	2	On the Sabbath , he began to teach in the synagogue . Many who heard him were surprised . " Where did this man get all this ?
40013056	PLACE.FROM.SOURCE	2	And his sisters , aren't they here with us ? Where did this man get all this ? "
43013036	PLACE.GOAL	2	Simon Peter said to Jesus , " Lord , where are you going ? " Jesus answered , " Where I am going , you can't follow me now , but you will follow later . "
43016005	PLACE.GOAL	2	But now I go away to the one who sent me . None of you ask me , ' Where are you going ? '
43007035	PLACE.GOAL	2	The Jewish opposition asked each other , " Where does he intend to go that we can't find him ? Surely he doesn't intend to go where our people have been scattered and are living among the Greeks ! He isn't going to teach the Greeks , is he ?
43008010	PLACE.OBJECT.PL	2	Jesus stood up and said to her , " Woman , where are they ? Is there no one to condemn you ? "
42017017	PLACE.OBJECT.PL	2	Jesus replied , " Weren't ten cleansed ? Where are the other nine ?
43009012	PLACE.OBJECT.SG	2	They asked , " Where is this man ? " He replied , " I don't know . "
40002002	PLACE.OBJECT.SG	2	They asked , " Where is the newborn king of the Jews ? We've seen his star in the east , and we've come to honor him . "
43007011	PLACE.OBJECT.SG	2	The Jewish leaders were looking for Jesus at the festival . They kept asking , " Where is he ? "

Verse	Label	Cluster	Context
43008019	PLACE.OBJECT.SG	2	They asked him , " Where is your Father ? " Jesus answered , " You don’t know me and you don’t know my Father . If you knew me , you would also know my Father . "
40016013	PERSON.ASCRIPTION	3	Now when Jesus came to the area of Caesarea Philippi , he asked his disciples , " Who do people say the Human One is ? "
41008029b	PERSON.ASCRIPTION	3	Who do you say that I am ? " Peter answered , " You are the Christ . "
42009018	PERSON.ASCRIPTION	3	Once when Jesus was praying by himself , the disciples joined him , and he asked them , " Who do the crowds say that I am ? "
43021012	PERSON.IDENTITY.2SG	3	Jesus said to them , " Come and have breakfast . " None of the disciples could bring themselves to ask him , " Who are you ? " They knew it was the Lord .
43001022a	PERSON.IDENTITY.2SG	3	They asked , " Who are you ? We need to give an answer to those who sent us .
43008025a	PERSON.IDENTITY.2SG	3	" Who are you ? " they asked .
44022008	PERSON.IDENTITY.2SG	3	I answered , ‘ Who are you , Lord ? ’ ‘ I am Jesus the Nazarene , whom you are harassing , ’ he replied .
59004012	PERSON.IDENTITY.2SG	3	There is only one lawgiver and judge , and he is able to save and to destroy . But you who judge your neighbor , who are you ?
43008053	PERSON.IDENTITY.2SG	3	Are you greater than our father Abraham ? He died and the prophets died , so who do you make yourself out to be ? "
43001021	PERSON.IDENTITY.2SG	3	They asked him , " Then who are you ? Are you Elijah ? " John said , " I’m not . " " Are you the prophet ? " John answered , " No . "
40021010	PERSON.IDENTITY.3SG	3	And when Jesus entered Jerusalem , the whole city was stirred up . " Who is this ? " they asked .
42009009a	PERSON.IDENTITY.3SG	3	Herod said , " I beheaded John , so now who am I hearing about ? " Herod wanted to see him .
42007049	PERSON.IDENTITY.3SG	3	The other table guests began to say among themselves , " Who is this person that even forgives sins ? "
42005021a	PERSON.IDENTITY.3SG	3	The legal experts and Pharisees began to mutter among themselves , " Who is this who insults God ?
40012048b	PERSON.IDENTITY.PL	3	Who are my brothers ? "
66007013a	PERSON.IDENTITY.PL	3	Then one of the elders said to me , " Who are these people wearing white robes ,
44019015	PERSON.IDENTITY.PL	3	The evil spirit replied , " I know Jesus and I’m familiar with Paul , but who are you ? "

Verse	Label	Cluster	Context
46003005b	PERSON.KIND	3	What is Paul ? They are servants who helped you to believe . Each one had a role given to them by the Lord :
46003005a	PERSON.KIND	3	After all , what is Apollos ?
40008027	PERSON.KIND	3	The people were amazed and said , " What kind of person is this ? Even the winds and the lake obey him ! "
42020024a	PERSON.POSSESSOR	3	" Show me a coin . Whose image
40022020a	PERSON.POSSESSOR	3	" Whose image
40022042b	PERSON.POSSESSOR	3	Whose son is he ? " " David’s son , " they replied .
42011019	PERSON.POSSESSOR	3	If I throw out demons by the authority of Beelzebul , then by whose authority do your followers throw them out ? Therefore , they will be your judges .
42020033	PERSON.POSSESSOR	3	In the resurrection , whose wife will she be ? All seven were married to her . "
44004007c	PERSON.POSSESSOR	3	or in what name did you do this ? "
41005030	PERSON.ROLE.AGENT	3	At that very moment , Jesus recognized that power had gone out from him . He turned around in the crowd and said , " Who touched my clothes ? "
42008045	PERSON.ROLE.AGENT	3	" Who touched me ? " Jesus asked . When everyone denied it , Peter said , " Master , the crowds are surrounding you and pressing in on you ! "
43007020	PERSON.ROLE.AGENT	3	The crowd answered , " You have a demon . Who wants to kill you ? "
42022064	PERSON.ROLE.AGENT	3	They blindfolded him and asked him repeatedly , " Prophesy ! Who hit you ? "
42012014	PERSON.ROLE.AGENT	3	Jesus said to him , " Man , who appointed me as judge or referee between you and your brother ? "
43012038a	PERSON.ROLE.AGENT	3	This was to fulfill the word of the prophet Isaiah : Lord , who has believed through our message ?
44007027	PERSON.ROLE.AGENT	3	The one who started the fight against his neighbor pushed Moses aside and said , ‘ Who appointed you as our leader and judge ?
45007024	PERSON.ROLE.AGENT	3	I’m a miserable human being . Who will deliver me from this dead corpse ?
42003007	PERSON.ROLE.AGENT	3	Then John said to the crowds who came to be baptized by him , " You children of snakes ! Who warned you to escape from the angry judgment that is coming soon ?
45008031b	PERSON.ROLE.AGENT	3	If God is for us , who is against us ?

Verse	Label	Cluster	Context
41011028b	PERSON.ROLE.AGENT	3	Who gave you this authority to do them ? "
40003007	PERSON.ROLE.AGENT	3	Many Pharisees and Sadducees came to be baptized by John . He said to them , " You children of snakes ! Who warned you to escape from the angry judgment that is coming soon ?
43021020	PERSON.ROLE.AGENT	3	Peter turned around and saw the disciple whom Jesus loved following them . This was the one who had leaned against Jesus at the meal and asked him , " Lord , who is going to betray you ? "
43006060	PERSON.ROLE.AGENT	3	Many of his disciples who heard this said , " This message is harsh . Who can hear it ? "
48003001	PERSON.ROLE.AGENT	3	You irrational Galatians ! Who put a spell on you ? Jesus Christ was put on display as crucified before your eyes !
43005012	PERSON.ROLE.AGENT	3	They inquired , " Who is this man who said to you , ‘ Pick it up and walk ’ ? "
41016003	PERSON.ROLE.AGENT	3	They were saying to each other , " Who’s going to roll the stone away from the entrance for us ? "
41005031	PERSON.ROLE.AGENT	3	His disciples said to him , " Don’t you see the crowd pressing against you ? Yet you ask , ‘ Who touched me ? ’ "
47002002	PERSON.ROLE.AGENT	3	If I make you sad , who will be there to make me glad when you are sad because of me ?
48005007a	PERSON.ROLE.AGENT	3	You were running well — who stopped you from obeying the truth ?
42016011	PERSON.ROLE.AGENT	3	If you haven’t been faithful with worldly wealth , who will trust you with true riches ?
46004007a	PERSON.ROLE.AGENT	3	Who says that you are better than anyone else ?
43006068	PERSON.ROLE.GOAL	3	Simon Peter answered , " Lord , where would we go ? You have the words of eternal life .
43012038b	PERSON.ROLE.GOAL	3	To whom is the arm of the Lord fully revealed ?
43018004	PERSON.ROLE.PATIENT	3	Jesus knew everything that was to happen to him , so he went out and asked , " Who are you looking for ? "
43020015b	PERSON.ROLE.PATIENT	3	Who are you looking for ? " Thinking he was the gardener , she replied , " Sir , if you have carried him away , tell me where you have put him and I will get him . "
58003018	PERSON.ROLE.RECIPIENT	3	And against whom did he swear that they would never enter his rest , if not against the ones who were disobedient ?
58003017	PERSON.ROLE.RECIPIENT	3	And with whom was God angry for forty years ? Wasn’t it with the ones who sinned , whose bodies fell in the desert ?

Verse	Label	Cluster	Context
42010036b	PERSON.SELECTION.MULTIPLE	3	Which one of these three was a neighbor to the man who encountered thieves ? "
42022027	PERSON.SELECTION.TWO	3	So which one is greater , the one who is seated at the table or the one who serves at the table ? Isn't it the one who is seated at the table ? But I am among you as one who serves .
42007042	PERSON.SELECTION.TWO	3	When they couldn't pay , the lender forgave the debts of them both . Which of them will love him more ? "
40027021	PERSON.SELECTION.TWO	3	The governor said , " Which of the two do you want me to release to you ? " " Barabbas , " they replied .
40021031	PERSON.SELECTION.TWO	3	" Which one of these two did his father's will ? " They said , " The first one . " Jesus said to them , " I assure you that tax collectors and prostitutes are entering God's kingdom ahead of you .
42016007	QUANTITY.MASS	4	Then the manager said to another , ' How much do you owe ? ' He said , ' One thousand bushels of wheat . ' He said , ' Take your contract and write eight hundred . '
42016005	QUANTITY.MASS	4	" One by one , the manager sent for each person who owed his master money . He said to the first , ' How much do you owe my master ? '
42003010	THING.DO	4	The crowds asked him , " What then should we do ? "
42003012	THING.DO	4	Even tax collectors came to be baptized . They said to him , " Teacher , what should we do ? "
42003014b	THING.DO	4	What should we do ? " He answered , " Don't cheat or harass anyone , and be satisfied with your pay . "
44002037	THING.DO	4	When the crowd heard this , they were deeply troubled . They said to Peter and the other apostles , " Brothers , what should we do ? "
42018036	THING.HAPPEN	4	When the man heard the crowd passing by , he asked what was happening .
42015026	THING.HAPPEN	4	He called one of the servants and asked what was going on .
43002018b	THING.KIND	4	What miraculous sign will you show us ? "
43006030a	THING.KIND	4	They asked , " What miraculous sign will you do , that we can see and believe you ?
42006032	THING.KIND	4	" If you love those who love you , why should you be commended ? Even sinners love those who love them .
40005046	THING.KIND	4	If you love only those who love you , what reward do you have ? Don't even the tax collectors do the same ?

Verse	Label	Cluster	Context
44019003	THING.KIND	4	Then he said , " What baptism did you receive , then ? " They answered , " John’s baptism . "
41004030b	THING.KIND	4	What parable can I use to explain it ?
41011028a	THING.KIND	4	They asked , " What kind of authority do you have for doing these things ?
44007049a	THING.KIND	4	Heaven is my throne , and the earth is my footstool . ‘ What kind of house will you build for me , ’ says the Lord ,
41006024	THING.PATIENT	4	She left the banquet hall and said to her mother , " What should I ask for ? " " John the Baptist’s head , " Herodias replied .
40020021	THING.PATIENT	4	" What do you want ? " he asked . She responded , " Say that these two sons of mine will sit , one on your right hand and one on your left , in your kingdom . "
43001038a	THING.PATIENT	4	When Jesus turned and saw them following , he asked , " What are you looking for ? " They said ,
41010036	THING.PATIENT	4	" What do you want me to do for you ? " he asked .
40006031b	THING.PATIENT	4	‘ What are we going to drink ? ’ or
43004027a	THING.PATIENT	4	Just then , Jesus’ disciples arrived and were shocked that he was talking with a woman . But no one asked , " What do you want ? " or
43018038	THING.PATIENT	4	" What is truth ? " Pilate asked . After Pilate said this , he returned to the Jewish leaders and said , " I find no grounds for any charge against him .
40019020	THING.PATIENT	4	The young man replied , " I’ve kept all these . What am I still missing ? "
40006031a	THING.PATIENT	4	Therefore , don’t worry and say , ‘ What are we going to eat ? ’ or
40019027	THING.PATIENT	4	Then Peter replied , " Look , we’ve left everything and followed you . What will we have ? "
42007024	THING.PATIENT	4	After John’s messengers were gone , Jesus spoke to the crowds about John . " What did you go out into the wilderness to see ? A stalk blowing in the wind ?
41009010	THING.PATIENT	4	So they kept it to themselves , wondering , " What’s this ‘ rising from the dead ’ ? "
43006030b	THING.PATIENT	4	What will you do ?
44010004	THING.PATIENT	4	Startled , he stared at the angel and replied , " What is it , Lord ? " The angel said , " Your prayers and your compassionate acts are like a memorial offering to God .

Verse	Label	Cluster	Context
41009033	THING.PATIENT	4	They entered Capernaum . When they had come into a house , he asked them , " What were you arguing about during the journey ? "
40006031c	THING.PATIENT	4	‘ What are we going to wear ? ’
41008037	THING.PATIENT	4	What will people give in exchange for their lives ?
42024019	THING.PATIENT	4	He said to them , " What things ? " They said to him , " The things about Jesus of Nazareth . Because of his powerful deeds and words , he was recognized by God and all the people as a prophet .
44002012	THING.PATIENT	4	They were all surprised and bewildered . Some asked each other , " What does this mean ? "
43007036	THING.PATIENT	4	What does he mean when he says , ‘ You will look for me , but you won’t find me , and where I am you can’t come ’ ? "
40026015	THING.PATIENT	4	and said , " What will you give me if I turn Jesus over to you ? " They paid him thirty pieces of silver .
41001027	THING.PATIENT	4	Everyone was shaken and questioned among themselves , " What’s this ? A new teaching with authority ! He even commands unclean spirits and they obey him ! "
44022016	THING.PATIENT	4	What are you waiting for ? Get up , be baptized , and wash away your sins as you call on his name . ’
41011005	THING.PATIENT	4	Some people standing around said to them , " What are you doing , untying the colt ? "
45009030	THING.SAY	4	So what are we going to say ? Gentiles who weren’t striving for righteousness achieved righteousness , the righteousness that comes from faith .
45008031a	THING.SAY	4	So what are we going to say about these things ?
43012027	THING.SAY	4	" Now I am deeply troubled . What should I say ? ‘ Father , save me from this time ’ ? No , for this is the reason I have come to this time .
46011022	THING.SAY	4	Don’t you have houses to eat and drink in ? Or do you look down on God’s churches and humiliate those who have nothing ? What can I say to you ? Will I praise you ? No , I don’t praise you in this .
45006001	THING.SAY	4	So what are we going to say ? Should we continue sinning so grace will multiply ?
40019018	THING.SELECTION.MULTIPLE.PL	4	The man said , " Which ones ? " Then Jesus said , " Don’t commit murder . Don’t commit adultery . Don’t steal . Don’t give false testimony .
40022036	THING.SELECTION.MULTIPLE.SG	4	" Teacher , what is the greatest commandment in the Law ? "

Verse	Label	Cluster	Context
43010032	THING.SELECTION.MULTIPLE.SG	4	Jesus responded , " I have shown you many good works from the Father . For which of those works do you stone me ? "
41012028	THING.SELECTION.MULTIPLE.SG	4	One of the legal experts heard their dispute and saw how well Jesus answered them . He came over and asked him , " Which commandment is the most important of all ? "
41002009	THING.SELECTION.TWO	4	Which is easier — to say to a paralyzed person , ‘ Your sins are forgiven , ’ or to say , ‘ Get up , take up your bed , and walk ’ ?
40023019	THING.SELECTION.TWO	4	You blind people ! Which is greater , the gift or the altar that makes the gift holy ?
40023017	THING.SELECTION.TWO	4	You foolish and blind people ! Which is greater , the gold or the temple that makes the gold holy ?
40022042a	THING.THINK	4	" What do you think about the Christ ?
40026066	THING.THINK	4	What do you think ? " And they answered , " He deserves to die ! "
40021028	THING.THINK	4	“ What do you think ? A man had two sons . Now he came to the first and said , ‘ Son , go and work in the vineyard today . ’
43011056	THING.THINK	4	They were looking for Jesus . As they spoke to each other in the temple , they said , " What do you think ? He won’t come to the festival , will he ? "
40017025a	THING.THINK	4	" Yes , " he said . But when they came into the house , Jesus spoke to Peter first . " What do you think , Simon ?
40018012	THING.THINK	4	What do you think ? If someone had one hundred sheep and one of them wandered off , wouldn’t he leave the ninety-nine on the hillsides and go in search for the one that wandered off ?
42006002	INTENTION	5	Some Pharisees said , " Why are you breaking the Sabbath law ? "
42018019	INTENTION	5	Jesus replied , " Why do you call me good ? No one is good except the one God .
51002020	INTENTION	5	If you died with Christ to the way the world thinks and acts , why do you submit to rules and regulations as though you were living in the world ?
40020006	INTENTION	5	Around five in the afternoon he went and found others standing around , and he said to them , ‘ Why are you just standing around here doing nothing all day long ? ’
42024005	INTENTION	5	The women were frightened and bowed their faces toward the ground , but the men said to them , " Why do you look for the living among the dead ?

Verse	Label	Cluster	Context
44009004	INTENTION	5	He fell to the ground and heard a voice asking him , " Saul , Saul , why are you harassing me ? "
40009004	INTENTION	5	But Jesus knew what they were thinking and said , " Why do you fill your minds with evil things ?
40022018	INTENTION	5	Knowing their evil motives , Jesus replied , " Why do you test me , you hypocrites ?
42019023	INTENTION	5	Why then didn’t you put my money in the bank ? Then when I arrived , at least I could have gotten it back with interest . ’
45009020c	INTENTION	5	" Why did you make me like this ? "
41004013	MANNER	6	" Don’t you understand this parable ? Then how will you understand all the parables ?
46015035a	MANNER	6	But someone will say , " How are the dead raised ?
43009010	MANNER	6	So they asked him , " How are you now able to see ? "
40021020	MANNER	6	When the disciples saw it , they were amazed . " How did the fig tree dry up so fast ? " they asked .
43009026b	MANNER	6	How did he heal your eyes ? "
43006042	MANNER.STATEMENT	6	They asked , " Isn’t this Jesus , Joseph’s son , whose mother and father we know ? How can he now say , ‘ I have come down from heaven ’ ? "
43004009	MANNER.STATEMENT	6	The Samaritan woman asked , " Why do you , a Jewish man , ask for something to drink from me , a Samaritan woman ? " (Jews and Samaritans didn’t associate with each other .)
43012034a	MANNER.STATEMENT	6	The crowd responded , " We have heard from the Law that the Christ remains forever . How can you say that the Human One must be lifted up ?
42020041	MANNER.STATEMENT	6	Jesus said to them , " Why do they say that the Christ is David’s son ?
41012035	MANNER.STATEMENT	6	While Jesus was teaching in the temple , he said , " Why do the legal experts say that the Christ is David’s son ?
43008033	MANNER.STATEMENT	6	They responded , " We are Abraham’s children ; we’ve never been anyone’s slaves . How can you say that we will be set free ? "
46015012	MANNER.STATEMENT	6	So if the message that is preached says that Christ has been raised from the dead , then how can some of you say , " There’s no resurrection of the dead " ?
43014009	MANNER.STATEMENT	6	Jesus replied , “ Don’t you know me , Philip , even after I have been with you all this time ? Whoever has seen me has seen the Father . How can you say , ‘ Show us the Father ’ ?

Verse	Label	Cluster	Context
40022043	MANNER.STATEMENT	6	He said , " Then how is it that David , inspired by the Holy Spirit , called him Lord when he said ,
48002014	MANNER.STATEMENT	6	But when I saw that they weren’t acting consistently with the truth of the gospel , I said to Cephas in front of everyone , " If you , though you’re a Jew , live like a Gentile and not like a Jew , how can you require the Gentiles to live like Jews ? "
41009012	MANNER.STATEMENT	6	He answered , " Elijah does come first to restore all things . Why was it written that the Human One would suffer many things and be rejected ?
40015034	QUANTITY.COUNT	6	Jesus said , " How much bread do you have ? " They responded , " Seven loaves and a few fish . "
41008020	QUANTITY.COUNT	6	" And when I broke seven loaves of bread for those four thousand people , how many baskets full of leftovers did you gather ? " They answered , " Seven . "
40018021	QUANTITY.FREQUENCY	6	Then Peter said to Jesus , " Lord , how many times should I forgive my brother or sister who sins against me ? Should I forgive as many as seven times ? "

References

Aijmer, Karin. 2008. Parallel and comparable corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: an international handbook*, vol. 1, 275–291. Berlin, New York: De Gruyter Mouton. (doi:10.1515/booksetHSK29)

Aikhenvald, Alexandra Y. 2015. *The art of grammar: a practical guide* (Oxford Linguistics). Oxford: Oxford University Press.

Aikio, Ante & Jussi Ylikoski. 2010. The Structure of North Saami. Salt Lake City: Department of Linguistics, The University of Utah. (Course Handout).

Alexander, Ronelle & Ellen Elias-Bursać. 2006. *Bosnian, Croatian, Serbian, a grammar: with sociolinguistic commentary*. Madison: The University of Wisconsin.

Alexander, Virginia & Clarence Alexander. 2011. Gwich'in to English Dictionary. (Unpublished manuscript).

Ameka, Felix K. 1991. *Ewe: Its Grammatical Constructions and Illocutionary Devices*. Canberra: Australian National University. (PhD Thesis.)

Andrade, Manuel J. 1955. *A grammar of Modern Yucatec* (Microfilm Collection of Manuscripts on Middle American Cultural Anthropology 41.2). Chicago: University of Chicago Library.

Arkadiev, Peter & Ivano Caponigro. 2021. Conveying content questions without wh-words: evidence from Abaza. *Proceedings of Sinn und Bedeutung* 25. 73–94. (doi:10.18148/SUB/2021.V25I0.925)

Aronoff, Mark. 1976. *Word formation in generative grammar* (Linguistic Inquiry Monographs 1). Cambridge, Mass: MIT Press.

Asyik, Abdul Gani. 1987. *A Contextual Grammar of Acehnese Sentences (Complementation)*. Ann Arbor: University of Michigan. (PhD Thesis.)

Bakker, Dik. 2011. Language Sampling. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology* (Oxford Handbooks in Linguistics), 100–130. Oxford: Oxford University Press. (doi:10.1093/oxfordhb/9780199281251.013.0007)

Bartholomew, Doris A. & Ralph Engel. 1987. Gramática Zoque. In Ralph Engel & Mary Engel (eds.), *Diccionario zoque de Francisco León* (Vocabularios (y Diccionarios) Indígenas "Mariano Silva y Aceves" 30), 329–416. Mexico: Instituto Lingüístico de Verano.

Beck, David. 2016. Some language-particular terms are comparative concepts. *Linguistic Typology* 20(2). 395–402. (doi:10.1515/lingty-2016-0013)

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg & Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of human language*, vol. 1: Method and Theory, 123–156. Stanford, Calif: Stanford University Press.

Beller, Richard & Patricia Beller. 1977. Huasteca Nahuatl. In Ronald W. Langacker (ed.), *Studies in Uto-Aztecan Grammar 2: Modern Aztec Grammatical Sketches* (Summer Institute of Linguistics Publications in Linguistics 56), 199–306. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Betts, LaVera. 1981. *Dicionário parintintín-português português-parintintín*. Brasília, DF: Summer Institute of Linguistics.

Betts, LaVera. 2012. *Kagwahiva Dictionary*. Anápolis, GO: Associação Internacional de Linguística SIL-Brasil.

Bhat, D. N. S. 2000. The Indefinite-Interrogative Puzzle. *Linguistic Typology* 4(3). 365–400.

Bhat, D.N. Shankara. 2004. *Pronouns* (Oxford Studies in Typology and Linguistic Theory). Oxford: Oxford University Press.

Biber, D. & S. Conrad & R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge Approaches to Linguistics). Cambridge: Cambridge University Press.

Birk, D. B. W. 1976. *The Malakmalak Language, Daly River (Western Arnhem Land)* (Pacific Linguistics Series B, 45). Canberra: Australian National University.

Blake, Frank Ringgold. 1925. *A grammar of the Tagálog language, the chief native idiom of the Philippine Islands* (American Oriental Series 1). New Haven, Conn: American Oriental Society.

Bohnhoff, Lee Edward. 2010. *A description of Dii: Phonology, grammar, and discourse*. Ngaoundéré, Cameroun: Dii Literature Team.

Bolles, David & Alejandra Bolles. 2019. *A grammar and anthology of the Yucatecan Mayan language*. Milford, CT, Cancun, Q.R.: Ms.

Booij, Geert E. 2015. Morphological Analysis. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 449–472. Oxford: Oxford University Press. (doi:10.1093/oxfordhb/9780199677078.013.0020)

Braine, Jean C. 1970. *Nicobarese Grammar (Car Dialect)*. Berkeley: University of California at Berkeley. (PhD Thesis.)

Brandão, Ana Paula Barros. 2014. *A reference grammar of Paresi-Haliti (Arawak)*. Austin: The University of Texas at Austin. (PhD Thesis.)

Breedveld, J. O. 1995. *Form and Meaning in Fulfulde: A Morphophonological Study of Maasinankoore* (CNWS Publications 32). Leiden: Research School CNWS.

Britton, A. Scott. 2005. *Guaraní: Guaraní-English , English-Guaraní concise dictionary* (Hippocrene Concise Dictionaries). New York: Hippocrene Books.

Brown, Cecil H. 2013. Hand and Arm. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/129. (Accessed 2023-08-13.)

Brown, Lea. 2001. *A Grammar of Nias Selatan*. Sydney: University of Sydney. (PhD Thesis.)

Browne, Wayles & Theresa Alt. 2004. *A Handbook of Bosnian, Croatian and Serbian*. Durham NC: Slavic and East European Language Research Center (SEELRC), Duke University.

Bunt, Jonathan. 2003. *The Oxford Japanese grammar and verbs*. 1st edn. Oxford: Oxford University Press.

Burgess, Donald H. 1984. Western Tarahumara. In Ronald W. Langacker (ed.), *Studies in Uto-Aztecan grammar 4: Southern Uto-Aztecan grammatical sketches* (Summer Institute of Linguistics Publications in Linguistics 56), 1–149. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Bybee, Joan L. & Clay Beckner. 2015. Usage-Based Theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 953–980. Oxford: Oxford University Press. (doi:10.1093/oxfordhb/9780199677078.013.0032)

Casad, Eugene H. 1984. Cora. In Ronald W. Langacker (ed.), *Studies in Uto-Aztecan grammar 4: Southern Uto-Aztecan grammatical sketches* (Summer Institute of Linguistics Publications in Linguistics 56), 153–459. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Cheng, Lisa Lai Shen. 1991. *On the Typology of Wh-Questions*. Cambridge, Mass: Massachusetts Institute of Technology. (PhD Thesis.)

Childs, G. Tucker. 1995. *A Grammar of Kisi, a Southern Atlantic Language*. Berlin: De Gruyter Mouton.

Chisholm, William & Milic, Louis Tonko & Greppin, John A. C. (eds.). 1984. *Interrogativity: a colloquium on the grammar, typology, and pragmatics of questions in seven diverse languages, Cleveland, Ohio, October 5th, 1981-May 3rd, 1982* (Typological Studies in Language 4). Amsterdam, Philadelphia: J. Benjamins Pub. Co.

Christodouloupoulos, Christos & Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation* 49. 375–395.

Chung, S. 2020. *Chamorro Grammar*. Santa Cruz: University of California.

Clark, Larry. 1998. *Turkmen reference grammar* (Turcologica 34). Wiesbaden: Harrassowitz.

Comrie, Bernard. 1976. *Aspect: an introduction to the study of verbal aspect and related problems* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Cox, Elizabeth Ellen. 1975. *Kirundi grammar book*. Bujumbura: Methodist Missionary Soc.

Croft, William. 1995. Modern Syntactic Typology. In Masayoshi Shibatani & Theodora Bynon (eds.), *Approaches to Language Typology*, 85–145. Oxford: Oxford University Press.

Croft, William. 2002. *Typology and Universals*. 2nd edn. Cambridge: Cambridge University Press. (doi:10.1017/CBO9780511840579)

Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393. (doi:10.1515/lingty-2016-0012)

Croft, William & Keith Poole. 2008. Multidimensional scaling and other techniques for uncovering universals. *Theoretical Linguistics* 34(1). 75–84. (doi:10.1515/THLI.2008.007)

Crystal, David. 2008. *A dictionary of linguistics and phonetics* (The Language Library). 6th edn. Malden, MA: Blackwell Publishing.

Cysouw, Michael. 2004. *Interrogative words: an exercise in lexical typology. Presentation presented at the Bantu grammar: description and theory workshop*. ZAS Berlin.

Cysouw, Michael. 2005a. Quantitative methods in typology (Quantitative Methoden in der Typologie). In Reinhard Köhler & Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik / Quantitative Linguistics - Ein internationales Handbuch / An International Handbook*, 554–557. Berlin: De Gruyter Mouton.

Cysouw, Michael. 2005b. The typology of content interrogatives. Paper presented at Association for Linguistic Typology. Padang.

Cysouw, Michael. 2007. Content Interrogatives in Pichis Ashéninca: Corpus Study and Typological Comparison. *International Journal of American Linguistics* 73(2). 133–163. (doi:10.1086/519056)

Cysouw, Michael. 2008. Generalizing Language Comparison. *Theoretical Linguistics* 34(1). 47–51. (doi:10.1515/THLI.2008.003)

Cysouw, Michael. 2014. Inducing semantic roles. In Silvia Luraghi & Heiko Narrog (eds.), *Perspectives on Semantic Roles* (Typological Studies in Language 106), 23–68. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/tsl.106.02cys)

Cysouw, Michael & Chris Biemann & Matthias Ongyerth. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *Language Typology and Universals* 60(2). 158–171. (doi:doi:10.1524/stuf.2007.60.2.158)

Cysouw, Michael & Jeff Good. 2013. Languoid, Doculect and Glossonym: Formalizing the Notion "Language." *Language Documentation & Conservation* 7. 331–359.

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

da Milano, Federica. 2007. Demonstratives in parallel texts: a case study. *Language Typology and Universals* 60(2). 135–147. (doi:doi:10.1524/stuf.2007.60.2.135)

Dahl, Östen. 2001. Principles of areal typology. In Martin Haspelmath & Ekkehard König & Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook*, vol. 2, 1456–1470. Berlin, Boston: De Gruyter Mouton.

Dahl, Östen. 2007. From questionnaires to parallel corpora in typology. *Language Typology and Universals* 60(2). 172–181. (doi:10.1524/stuf.2007.60.2.172)

Dahl, Östen. 2016. Thoughts on language-specific and crosslinguistic entities. *Linguistic Typology* 20(2). 427–437. (doi:10.1515/lingty-2016-0016)

Dahl, Östen & Bernhard Wälchli. 2016. Perfects and iamitives: two gram types in one grammatical space. *Letras de Hoje* 51(3). 325–348. (doi:10.15448/1984-7726.2016.3.25454)

Davies, William D. 2010. *A grammar of Madurese*. Berlin: De Gruyter Mouton.

de Jong Boudreault, Lynda J. 2009. *A grammar of Sierra Popoluca (Soteapanec, a Mixe-Zoquean language)*. Austin: University of Texas at Austin. (PhD Thesis.)

de Swart, Henriëtte & Jos Tellings & Bernhard Wälchli. 2022. Not…Until across European Languages: A Parallel Corpus Study. *Languages* 7(1). 56. (doi:10.3390/languages7010056)

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

den Dikken, Marcel. 2003. On the morphosyntax of *wh*-movement. In Cedric Boeckx & Kleanthes K. Grohmann (eds.), *Multiple Wh-Fronting* (Linguistik Aktuell/Linguistics Today 64), 77–98. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/la.64.07dik)

Diessel, Holger. 2003. The relationship between demonstratives and interrogatives. *Studies in Language* 27(3). 635–655. (doi:10.1075/sl.27.3.06die)

Dindelegan, Gabriela Pană & Maiden, Martin (eds.). 2013. *The grammar of Romanian*. 1st edn. Oxford: Oxford University Press.

Divjak, Dagmar & Nick Fieller. 2014. Cluster analysis: Finding structure in linguistic data. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy* (Human Cognitive Processing 43), 405–441. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/hcp.43.16div)

Dixon, R. M. W. 2012. *Basic linguistic theory*. Vol. 3: Further grammatical topics. Oxford: Oxford University Press.

Donaldson, Bruce. 2008. *Dutch: a comprehensive grammar*. 2nd edn. London: Routledge.

Dryer, Matthew S. 1989. Large Linguistic Areas and Language Sampling. *Studies in Language* 13(2). 257–292. (doi:10.1075/sl.13.2.03dry)

Dryer, Matthew S. 1997. Are Grammatical Relations Universal? In Joan L. Bybee & John Haiman & Sandra A. Thompson (eds.), *Essays on Language Function and Language Type*, 115–143. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/z.82.09dry)

Dryer, Matthew S. 2013. Position of Interrogative Phrases in Content Questions. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/93. (Accessed 2023-08-1)

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/. (Accessed 2023-08-13.)

Durie, Mark. 1985. *A Grammar of Acehnese on the Basis of a Dialect of North Aceh* (Verhandelingen van Het Koninklijk Instituut Voor Taal-, Land- En Volkenkunde 112). Dordrecht: Foris Publications.

Dyer, Chris & Victor Chahuneau & Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648. Atlanta, Georgia: Association for Computational Linguistics.

Einarsson, Stefán. 1961. *Icelandic: grammar, texts, glossary*. 4th edn. Baltimore: Johns Hopkins Press.

Elson, Benjamin F. & Donaciano Gutiérrez G. 1999. *Diccionario Popoluca de la Sierra Veracruz*. Coyoacán, D.F.: Instituto Lingüístico de Verano.

Emenanjo, Emmanuel Nolue. 2015. *A grammar of contemporary Igbo: Constituents, features and processes*. Port Harcourt: M & J Grand Orbit Communications Ltd.

Engel, Ralph & Mary Allhiser de Engel. 1987. *Diccionario Zoque de Francisco Leon* (Serie de Vocabularios y Diccionarios Indígenas Mariano Silva y Aceves 30). México, D.F.: Instituto Lingüístico de Verano.

Engesæth, Tarjei & Mahire Yakup & Arienne Dwyer. 2009. *Teklimakandin salam: hazirqi zaman Uyghur tili qollanmisi = Greetings from the Teklimakan: a handbook of modern Uyghur*. Vol. 1. Lawrence: University of Kansas Scholarworks.

Előry, Vilma (ed.). 2007. *Értelmező szótár +: értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelvhasználati tanácsok és fogalomköri csoportok* (Magyar Nyelv Kézikönyvei 13–14). 1st edn. Budapest: Tinta könyviadó.

Érsek, Iván. 1977. *Langenscheidts praktisches Lehrbuch Ungarisch: ein Standardwerk für Anfänger*. Berlin: Langenscheidt.

Estigarribia, Bruno. 2020. *A grammar of Paraguayan Guarani*. London: UCL press.

Evans, Nicholas. 2011. Semantic Typology. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology* (Oxford Handbooks in Linguistics), 504–533. Oxford: Oxford University Press. (doi:10.1093/oxfordhb/9780199281251.013.0024)

Evans, Nicholas. 2020. Introduction: Why the comparability problem is central in typology. *Linguistic Typology* 24(3). 417–425. (doi:10.1515/lingty-2020-2055)

Everitt, Brian S. & Sabine Landau & Morven Leese & Daniel Stahl. 2011. *Cluster analysis*. 5th edn. Chichester, West Sussex, U.K: John Wiley & Sons.

Frajzyngier, Zygmunt & Erin Shay. 2002. *A grammar of Hdi* (Mouton Grammar Library 21). Berlin, New York: De Gruyter Mouton.

Gili, Joan. 1967. *Introductory Catalan Grammar: with a Brief Outline of the Language and Literature: A Selection from Catalan Writers, and a Catalan-English and English-Catalan Vocabulary*. Oxford: Dolphin Book Co.

Glasgow, Kathleen. 1994. *Burarra-Gun-Nartpa Dictionary*. Berriman, Australia: Summer Institute of Linguistics.

Glasgow, Kathleen Glasgow & David Glasgow. 2011. *Burarra-English Interactive Dictionary* (AuSIL Interactive Dictionary Series A-1). The Australian Society for Indigenous Languages (AuSIL). (Ed. Lecompte, Maarten.)

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A comprehensive grammar*. London, New York: Routledge.

Gönczöl-Davies, Ramona. 2008. *Romanian: an essential grammar*. London, New York: Routledge.

González, Felipe David Sánchez. 2012. *Diccionario Garífuna-Español guatemalteco, para uso de las escuelas bilingües interculturales de los municipios de Puerto Barrios y*

Livingston del departamento de Izabal. Guatemala: Universidad de San Carlos de Guatemala. (Master's Thesis.)

Gray, David. 2015. *A short descriptive grammar of the Turkmen language*. Cheltenham: SIL-NEG.

Green, Rebecca. 1987. *A Sketch Grammar of Burarra*. Canberra: Australian National University. (B.A. Hons Thesis.)

Gudai, Darmansyah. 1988. *A Grammar of Ma'anyan, A Language of Central Kalimantan*. Canberra: Australian National University. (PhD Thesis.)

Gwich 'in Social and Cultural Institute. 2009. *Gwich'in Topical Dictionary: Gwichyah Gwich 'in & Teet/ 'it Gwich 'in Dialects*. 6th edn. Fort McPherson: Gwich'in Social and Cultural Institute.

Haacke, Wilfrid Heinrich Gerhard. 1976. *A Nama grammar: the noun-phrase*. Cape Town: University of Cape Town. (Master's Thesis.)

Haan, Judith. 2002. *Speaking of questions: an exploration of Dutch question intonation* (LOT 52). Utrecht: Netherlands Graduate School of Linguitics.

Hagège, Claude. 2008. Towards a typology of interrogative verbs. *Linguistic Typology* 12(1). 1–44. (doi:10.1515/LITY.2008.031)

Hagman, Roy Stephen. 1977. *Nama Hottentot grammar*. Bloomington: Indiana University.

Hanson, Rebecca. 2010. *A grammar of Yine (Piro)*. Victoria: La Trobe University. (PhD Thesis.)

Harris, Brian. 1988. Bi-text, a new concept in translation theory. *Language Monthly* 54(March). 8–10.

Haspelmath, Martin. 1997. *Indefinite pronouns* (Oxford Studies in Typology and Linguistic Theory). Oxford: Oxford University Press.

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132. (doi:10.1515/LINGTY.2007.011)

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Olmen & Tanja Mortelmans & Frank Brisard (eds.), *Aspects of Linguistic Variation*, 83–114. Berlin, Boston: De Gruyter Mouton. (doi:doi: 10.1515/9783110607963-004)

Haspelmath, Martin & Andrea D. Sims. 2010. *Understanding morphology* (Understanding Language Series). 2nd edn. London: Hodder Education.

Haurholm-Larsen, Steffen. 2016. *A Grammar of Garifuna*. Bern: University of Bern. (PhD Thesis.)

Hawkesworth, Celia. 1998. *Colloquial Croatian and Serbian: the complete course for beginners*. London, New York: Routledge.

Heath, Jeffrey. 2017. *A Grammar of Yorno So (Toro So subgroup of Dogon, Mali)*. Language Description Heritage Library (MPI).

Heine, Bernd & Ulrike Claudi & Friederike Hünnemeyer. 1991. *Grammaticalization: a conceptual framework*. Chicago: University of Chicago Press.

Hengeveld, Kees & Maria Luiza Braga & Elisiene De Melo Barbosa & Jaqueline Silveira Coriolano & Juliana Jezuino Da Costa & Mariana De Souza Martins & Diego Leite De Oliveira et al. 2012. Semantic categories in the indigenous languages of Brazil. *Functions of Language* 19(1). 38–57. (doi:10.1075/fol.19.1.02hen)

Hershberger, Henry D. & Ruth Hershberger. 1982. *Kuku-Yalanji dictionary* (Work Papers of SIL- AAB. Series B, 7). Darwin: Summer Institute of Linguistics.

Hoey, Elliott Michael. 2013. *Grammatical Sketch of Turkmen*. Santa Barbara: University of California at Santa Barbara. (Master's Thesis.)

Holmes, Ruth Bradley & Betty Sharp Smith. 1977. *Beginning Cherokee*. 2nd edn. Norman: University of Oklahoma Press.

Hölzl, Andreas. 2018. *A typology of questions in Northeast Asia and beyond: an ecological perspective* (Studies in Diversity Linguistics 20). Berlin: Languange Science Press.

Hoogshagen Noordsy, Searle & Hilda Halloran de Hoogshagen. 1993. *Diccionario Mixe de Coatlán, Oaxaca* (Serie de Vocabularios y Diccionarios Indígenas Mariano Silva y Aceves 32). 1st edn. México, D.F.: Instituto Lingüístico de Verano.

Hopkins, Nicholas. 2012. *A dictionary of the Chuj Mayan language: As spoken in San Mateo Ixtatán, Huehuetenango, Guatemala*. Florida: Jaguar Tours.

Hoppmann, Dorothea. 2011. *Einführung in die koreanische Sprache: auf der Grundlage des gleichnamigen von Bruno Lewin und Tschong Dae Kim verfaßten Lehrbuchs*. 2., unveränd. Aufl. Hamburg: Buske.

Hoymann, Gertie. 2010. Questions and responses in ǂĀkhoe Hai||om. *Journal of Pragmatics* 42(10). 2726–2740. (doi:10.1016/j.pragma.2010.04.008)

Huddleston, Rodney. 1994. The Contrast between Interrogatives and Questions. *Journal of Linguistics*. Cambridge University Press 30(2). 411–439.

Hundt, Marianne. 2020. Corpus-Based Approaches to World Englishes. In Daniel Schreier & Marianne Hundt & Edgar W. Schneider (eds.), *The Cambridge Handbook of World Englishes*, 506–533. 1st edn. Cambridge: Cambridge University Press. (doi: 10.1017/9781108349406.022)

Idiatov, Dmitry. 2007. *A typology of non-selective interrogative pronominals*. Antwerp: University of Antwerp. (PhD Thesis.)

Idiatov, Dmitry & Johan van der Auwera. 2004. On interrogative pro-verbs. *Proceedings of the Workshop on the Syntax, Semantics, and Pragmatics of Quesitons*, 17–23.

Imani, Ayyoob & Masoud Jalili Sabet & Philipp Dufter & Michael Cysouw & Hinrich Schütze. 2021. ParCourE: A Parallel Corpus Explorer for a Massively Multilingual Corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 63–72. Online: Association for Computational Linguistics. (doi:10.18653/v1/2021.acl-demo.8)

Janhunen, Juha (ed.). 2003. *The Mongolic Languages* (Routledge Language Family Series 5). *The Mongolic Languages* (Routledge Language Family Series 5). London: Routledge.

Janhunen, Juha. 2012. *Mongolian* (London Oriental and African Language Library 19). Amsterdam, Philadelphia: John Benjamins. (doi:10.1075/loall.19)

Jensen, Joshua. 2014. *Syntactic Structures in an Austronesian Language*. Berlin, Boston: De Gruyter Mouton. (doi:doi:10.1515/9781614516804)

Johanson, Lars & Csató, Éva Á. (eds.). 1998. *The Turkic languages* (Routledge language family descriptions). 1. publ. London: Routledge.

Jones, Walton Glyn & Kirsten Gade. 1981. *Danish: A grammar*. Copenhagen: Nordisk Forlag.

Jukes, Anthony Robert. 2006. *Makassarese (basa Mangkasara'): A description of an Austronesian language of South Sulawesi*. Melbourne: University of Melbourne. (PhD Thesis.)

Kahn, Lily & Riitta-Liisa Valijärvi. 2017. *North Sámi: An Essential Grammar*. New York: Routledge.

Kalectaca, Milo & Ronald W. Langacker. 1978. *Lessons in Hopi*. Tucson: University of Arizona Press.

Kanungo, Tapas & Philip Resnik & Song Mao & Doe-Wan Kim & Qigong Zheng. 2005. The Bible and multilingual optical character recognition. *Communications of the ACM* 48(6). 124–130. (doi:10.1145/1064830.1064837)

Karlsson, Fred. 2008. *Finnish: an essential grammar* (Routledge Essential Grammars). 2nd edn. London, New York: Routledge.

Kaufman, Leonard & Peter J. Rousseeuw. 2005. *Finding groups in data: an introduction to cluster analysis* (Wiley Series in Probability and Mathematical Statistics). Hoboken, New Jersey: John Wiley & Sons.

Keller, Kathryn C. & Plácido Luciano G. 1997. *Diccionario Chontal de Tabasco* (Serie de Vocabularios y Diccionarios Indígenas Mariano Silva y Aceves 36). 1st edn. Tucson, Arizona: SIL.

Kenesei, István & Robert M. Vago & Anna Fenyvesi. 1998. *Hungarian* (Descriptive Grammars Series). London, New York: Routledge. (doi:10.4324/9780203192238)

Kennedy, Graeme. 1998. *An introduction to corpus linguistics* (Studies in Language and Linguistics). 1st edn. London: Longman.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (Routledge Handbooks in Applied Linguistics), 487–500. London: Routledge. (doi:10.4324/9780203856949.ch35)

King, Gareth. 2003. *Modern Welsh: A Comprehensive Grammar* (Comprehensive Grammars). 2nd edn. London: Routledge. (doi:10.4324/9780203987063)

Köhler, Bernhard. 2016. *Form und Funktion von Fragesätzen in afrikanischen Sprachen* (Schriften zur Afrikanistik / Research in African Studies 25). Frankfurt am Main: Peter Lang.

König, Ekkehard & Peter Siemund. 2007. Speech Act Distinctions in Grammar. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*, vol. 1: Clause structure, 276–324. Cambridge: Cambridge University Press. (doi:10.1017/CBO9780511619427.005)

Kotek, Hadas & Michael Yoshitaka Erlewine. 2019. Wh-indeterminates in Chuj (Mayan). *Canadian Journal of Linguistics/Revue canadienne de linguistique* 64(1). 62–101.

Kratochvíl, František. 2007. *A grammar of Abui: a Papuan language of Alor*. Utrecht: LOT.

Lanz, Linda A. 2010. *A grammar of Iñupiaq morphosyntax*. Houston: Rice University. (PhD Thesis.)

LaPolla, Randy J. 2016. On categorization: Stick to the facts of the languages. *Linguistic Typology* 20(2). 365–375. (doi:10.1515/lingty-2016-0011)

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Jan Svartvik (ed.), *Directions in Corpus Linguistics*, 105–126. Berlin, New York: De Gruyter Mouton. (doi:10.1515/9783110867275.105)

Leskien, August. 1976. *Grammatik der serbo-kroatischen Sprache: Lautlehre, Stammbildung, Formenlehre* (Sammlung slavischer Lehr- und Handbücher I. Reihe, Grammatiken 4). 2nd edn. Heidelberg: Winter.

Levinsohn, Stephen H. 1972. The interrogative in Inga (Quechuan). *International Journal of American Linguistics*. Baltimore 38(4). 260–264.

Levinsohn, Stephen H. 1974. *Una gramática pedagógica del Inga I*. Lomalinda: ILV/MG.

Levinsohn, Stephen H. 1976. *Una gramática pedagógica del inga II*. Lomalinda: ILV/MG.

Levinsohn, Stephen H. & Luis G. Galeano L. 1981. *Inga yachaycusunchi (Aprendamos inga: Gramática pedagógica del inga)*. npl.

Levinson, Stephen & Sérgio Meira & The Language & Cognition Group. 2003. "Natural Concepts" in the Spatial Topological Domain-Adpositional Meanings in Crosslinguistic Perspective: An Exercise in Semantic Typology. *Language* 79(3). 485–516.

Levinson, Stephen C. 2012. Interrogative intimations: On a possible social economics of interrogatives. In Jan P.Editor de Ruiter (ed.), *Questions: Formal, Functional and Interactional Perspectives* (Language Culture and Cognition), 11–32. Cambridge: Cambridge University Press. (doi:10.1017/CBO9781139045414.003)

Levinson, Stephen C. 2016. Speech acts. *Oxford Handbook of Pragmatics*, 199–216. Oxford: Oxford University Press.

Levshina, Natalia. 2015. *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/z.195)

Levshina, Natalia. 2017. Online film subtitles as a corpus: an *n*-gram approach. *Corpora* 12(3). 311–338. (doi:10.3366/cor.2017.0123)

Levshina, Natalia. 2022a. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 26(1). 129–160. (doi:10.1515/lingty-2020-0118)

Levshina, Natalia. 2022b. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft* 41(1). 179–205. (doi:10.1515/zfs-2021-2043)

Lichtenberk, Frantisek. 2007. A typologically unusual interrogative word in Toqabaqita and other Oceanic languages. *Oceanic Linguistics* 46(2). 603–612.

Lin, Dong-yi. 2012. Interrogative Verbs in Kavalan and Amis. *Oceanic Linguistics* 51(1). 182–206.

Lindblom, Gerhard. 1914. *Outlines of a Tharaka Grammar: With a list of words and specimens of the language*. Appelberg.

Lindström, Eva. 1995. Animacy in interrogative pronouns. In Inger Moen & Hanne G. Simonsen & Helge Lødrup (eds.), *Papers from the XVth Scandinavian Conference of Linguistics*, 307–315. Oslo: Univesity of Olso.

Lundskær-Nielsen, Tom & Philip Holmes. 2011. *Danish: An Essential Grammar*. 2nd edn. London: Routledge.

Luo, Tianhua. 2016. *Interrogative Strategies: An areal typology of the languages of China* (Studies in Chinese Language and Discourse 5). Amsterdam: John Benjamins Publishing Company. (doi:10.1075/scld.5)

Ma, Seng Mai. 2012. *A Descriptive Grammar of Wa*. Chiang Mai: Payap University. (Master's Thesis.)

MacDonald, Archdeacon. 1972. *A grammar of the Tukudh language*. Yellowknife, N.W.T.: Curriculum Division, Dept. of Education, Government of the Northwest Territories.

Mackenzie, J. Lachlan. 2009. Content interrogatives in a sample of 50 languages. *Lingua* 119(8). 1131–1163. (doi:10.1016/j.lingua.2007.12.005)

Makino, Seiichi & Michio Tsutsui. 2008. *A dictionary of advanced Japanese grammar*. Tokoyo: The Japan Times.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4(3). 307–333. (doi:10.1515/lity.2000.4.3.307)

Matteson, Esther L. 1963. *The Piro (Arawak) language*. Berkeley: University of California at Berkeley. (PhD Thesis.)

Matthews, Stephen & Virginia Yip. 2013. *Cantonese: A comprehensive grammar*. London: Routledge.

Mayer, Thomas & Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 54–62. Avignon, France: Association for Computational Linguistics.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3158–3163. Reykjavik, Iceland: European Language Resources Association (ELRA).

Meyer, Charles F. 2008. Pre-electronic corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 1, 1–13. Berlin, New York: De Gruyter Mouton.

Miestamo, Matti. 2007. Symmetric and asymmetric encoding of functional domains, with remarks on typological markedness. In Matti Miestamo & Bernhard Wälchli (eds.), *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, 293–314. Berlin, New York: De Gruyter Mouton. (doi: 10.1515/9783110198904.5.293)

Miller, Carolyn P. 2017. Eastern Bru Grammar Sketch. SIL.

Miller, John D. 1964. Word classes in Brôu. *Mon-Khmer Studies* 1. 41–62.

Miyaoka, Osahito. 2012. *A grammar of Central Alaskan Yupik (CAY)*. Berlin, Boston: De Gruyter Mouton.

Moisl, Hermann. 2015. *Cluster Analysis for Corpus Linguistics*. Berlin, Munich, Boston: De Gruyter Mouton. (doi:10.1515/9783110363814)

Monguí, Raul & Stephen H. Levinsohn & Luis G. Galeano L. 1976. *Una gramática pedagógica del inga (segunda parte)*. Bogotá: Ministerio de Gobierno.

Montgomery-Anderson, Brad. 2008. *A Reference Grammar of Oklahoma Cherokee*. Ann Arbor, MI: University of Kansas. (PhD Thesis.)

Morales, Salomé Gutiérrez. 2015. *Vocabulario popoluca de la sierra - español - popoluca de la sierra*. Ciudad de México: Cultura-inali-sev-aveli.

Moskovsky, Christo & Alan Libert. 2006. Questions in Natural and Artificial Languages. *Journal of Universal Language* 7(2). 65–120. (doi:10.22425/jul.2006.7.2.65)

Muriungi, Peter. 2005. Wh-Questions in Kitharaka. *Studies in African Linguistics* 34(1). 43–104. (doi:10.32473/sal.v34i1.107332)

Mus, Nikolett. 2015. *Interrogative words and content questions in Tundra Nenets*. Szeged: University of Szeged. (PhD Thesis.)

Mushin, Liana. 1995. Epistememes in Australian languages. *Australian Journal of Linguistics* 15(1). 1–31. (doi:10.1080/07268609508599514)

Muysken, Pieter & Norval Smith. 1990. Question words in pidgin and creole languages. *Linguistics* 28(4). 883–904. (doi:10.1515/ling.1990.28.4.883)

Mwinlaaru, Isaac N. 2017. *A systemic functional description of the grammar of Dagaare*. Hong Kong: Hong Kong Polytechnic University. (PhD Thesis.)

Myers-Scotton, Carol & Gregory John Orr. 1980. *Learning Chichewa* (Peace Corps Language Handbook Serie). East Lansing: African Studies Center, Michigan State University.

Nagai, Tadataka. 2006. *Agentive and patientive verb bases in North Alaskan Iñupiaq*. Alaska: University of Alaska Fairbanks. (PhD Thesis.)

Nau, Nicole. 1999. Was schlägt der Kasus? Zu Paradigmen und Formengebrauch von Interrogativpronomina. *Language Typology and Universals* 52(2). 130–150. (doi: 10.1524/stuf.1999.52.2.130)

Naughton, James. 2005. *Czech: An Essential Grammar*. London, New York: Routledge.

Neijmann, D.L. 2001. *Colloquial Icelandic: The Complete Course for Beginners* (Colloquial Icelandic: The Complete Course for Beginners 1). London, New York: Routledge.

Newmeyer, Frederick J. 2007. Linguistic typology requires crosslinguistic formal categories. *Linguistic Typology* 11(1). 133–157. (doi:10.1515/LINGTY.2007.012)

Nida, Eugene A. & Charles Taber. 1982. *The Theory and Practice of Translation* (Helps for Translators). Second photomechanical reprint. Vol. 8. Leiden: Brill.

Nies, Joyce. 1986. *Diccionario Piro (Tokanchi gikshijikowaka-steno)* (Serie Lingüística Peruana 22). Yarinacocha: Ministerio de Educación and Instituto Lingüístico de Verano.

Oguro, Takeshi. 2015. WH-Questions in Japanese and Speech Act Phrase. *Linguistica Atlantica* 34(2). 89–101.

Olawsky, Knut. 2006. *A Grammar of Urarina*. Berlin, New York: De Gruyter Mouton. (doi: 10.1515/9783110892932)

Omar, Asmah H. 1969. *The Iban language of Sarawak: A grammatical description*. London: University of London. (PhD Thesis.)

Parker, Gary John. 1969. *Ayacucho Quechua grammar and dictionary* (Janua Linguarum: Series Practica 82). The Hague: Mouton.

Patz, Elisabeth. 2002. *A Grammar of the Kuku Yalanji Language of North Queensland* (Pacific Linguistics 527). Canberra: Australian National University.

Pease, Helen. 1968. *Parintintin Grammar*. Porto Velho, RO: Associação Internacional de Linguística SIL-Brasil.

Perkins, Revere D. 1989. Statistical Techniques for Determining Language Sample Size. *Studies in Language* 13(2). 293–315. (doi:10.1075/sl.13.2.04per)

Perkins, Revere D. 2001. Sampling procedures and statistical methods. In Haspelmath Martin & Ekkehard König & Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook*, vol. 1, 419–434. Berlin, Boston: De Gruyter Mouton. (doi:doi:10.1515/9783110194036)

Peter, Katherine. 1979. *Dinjii Zhuh Ginjik Nagwan Tr'iiltsaii: Gwich'in Junior Dictionary*. Anchorage: ANLA.

Petzell, Malin. 2008. *The Kagulu language of Tanzania: grammar, texts and vocabulary* (East African Languages and Dialects 19). Köln: Rüdiger Köppe Verlag.

Pittman, Richard Saunders. 1954. A Grammar of Tetelcingo (Morelos) Nahuatl. *Language* 30(1). 5–67. (doi:10.2307/522207)

Poppe, Nicholaus. 1951. *Khalkha-mongolische Grammatik: mit Bibliographie, Sprachproben und Glossar* (Veröffentlichungen Der Orientalischen Kommission / Akademie Der Wissenschaften Und Der Literatur 1). Wiesbaden: Steiner.

Preuss, Konrad Th. 1932. Grammatik der Cora-Sprache. *International Journal of American Linguistics* 7(1/2). 1–102.

Priest, Anne & Perry N. Priest. 1985. *Diccionario sirionó y castellano*. Cochabamba: Instituto Lingüístico de Verano.

Pulte, William & Durbin Feeling. 1975. Outline of Cherokee grammar. In Durbin Feeling (ed.), *Cherokee-English dictionary Tsalagi-Yonega Didehlogwasdohdi*, 235–355. Talequah: Cherokee Nation of Oklahoma.

Purvis, John B. 1907. *A Manual of Lumasaba Grammar*. London: Society for Promoting Christian Knowledge.

Resnik, Philip & Mari Broman Olsen & Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues.' *Computers and the Humanities* 33(1/2). 129–153. (doi:10.1023/A:1001798929185)

Rijkhoff, Jan. 2007. Linguistic Typology: a short history and some current issues. *Tidsskrift for Sprogforskning* 5(1). 1–18. (doi:10.7146/tfs.v5i1.529)

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314. (doi:10.1515/lity.1998.2.3.263)

Rijkhoff, Jan & Dik Bakker & Kees Hengeveld & Peter Kahrel. 1993. A Method of Language Sampling. *Studies in Language* 17(1). 169–203. (doi:10.1075/sl.17.1.07rij)

Robert, Stéphane. 2016. Content question words and noun class markers in Wolof: reconstructing a puzzle. *Frankfurt African Studies Bulletin* 23. 123–146.

Rounds, Carol. 2001. *Hungarian: an essential grammar*. London, New York: Routledge.

Rowan, Orland & E. B. Burgess. 1979. *Gramática Parecís*. Anápolis, GO: Associação Internacional de Lingüística SIL-Brasil.

Sadock, Jerrold & Arnold Zwicky. 1985. Speech act distinctions in syntax. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*, vol. 1: Clause structure, 155–196. Cambridge: Cambridge University Press.

Sadock, Jerry. 2012. Formal features of questions. In Jan P. de Ruiter (ed.), *Questions: Formal, Functional and Interactional Perspectives* (Language Culture and Cognition), 103–122. Cambridge: Cambridge University Press. (doi:10.1017/CBO9781139045414.008)

Sarlin, Mika. 2014. *Romanian grammar*. 2nd edn. Helsinki: BoD-Books on Demand.

Saxon, Leslie & Mary Siemens. 1996. *Tłı̨chǫ yatiì enı̨htl'è (A Dogrib dictionary)*. Rae-Edzo, NWT: Dogrib Divisional Board of Education.

Schachter, Paul & Fe T. Otanes. 1983. *Tagalog reference grammar*. Berkeley: University of California Press.

Schachter, Paul & Timothy Shopen. 2007. Part-of-speech systems. In Timothy Shopen (ed.), *Language typology and syntactic description*, vol. 1: Clause structure, 1–34. 2nd edn. Cambridge: Cambridge University Press.

Schoenhals, Louise C. & Alvin Schoenhals. 1965. *Vocabulario mixe de Totontepec* (Serie de Vocabularios Indígenas "Mariano Silva y Aceves" 14). Mexico: Instituto Lingüístico de Verano.

Schulze, Wolfgang. 2002. Gagausisch. *Lexikon der Sprachen des europäischen Ostens* (Wieser Enzyklopädie des Europäischen Ostens 10). Klagenfurt: Wieser Verlag.

Schulze, Wolfgang. 2007. Communication or memory mismatch?: Towards a cognitive typology of questions. In Günter Radden & Klaus-Michael Köpcke & Thomas Berg &

Peter Siemund (eds.), *Aspects of Meaning Construction*, 247–264. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/z.136.16sch)

Seiler, Wolf A. 2012. *Iñupiatun Eskimo Dictionary* (SIL Language and Culture Documentation and Description 16). Dallas, Texas: SIL International.

Shadeg, Norbert. 2014. *Tuttle Balinese-English Dictionary*. Berkeley: Tuttle Publishing.

Siemund, Peter. 2001. Interrogative constructions. In Martin Haspelmath & Ekkehard König & Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook*, vol. 2, 1010–1028. Berlin, Boston: De Gruyter Mouton. (doi:doi:10.1515/9783110194265-014)

Smith, Alexander D. 2017. *The languages of Borneo: A comprehensive classification*. Honolulu, HI: University of Hawai'i at Mānoa. (PhD Thesis.)

Sneddon, James N. 1996. *Indonesian: A Comprehensive Grammar* (Routledge Grammars Series). New York: Routledge.

Snell, Betty. 2011. *Diccionario matsigenka — castellano con índice castellano, notas enciclopédicas y apuntes gramaticales* (Serie Lingüística Peruana 56). Dallas: SIL International.

Snell, Betty E. 1998. *Pequeño diccionario machiguenga-castellano* (Documento de Trabajo 32). Edición preliminar. Lima: Instituto Lingüístico de Verano.

Somers, Harold. 2001. Bilingual parallel corpora and language engineering. *Anglo Indian Workshop "Language Engineering for South Asian Languages" LESAL*.

Song, Jae Jung. 2001. *Linguistic typology: morphology and syntax* (Longman Linguistics Library). Harlow, England, New York: Longman.

Song, Jae Jung (ed.). 2011. *The Oxford handbook of linguistic typology* (Oxford Handbooks in Linguistics). Oxford: Oxford University Press.

Soto-Ruiz, Clodoaldo. 1979. *Runasimi-Kastillanu-Inlis Llamkaymanaq Qullqa Ayakuchu-Chnka I Rakta / Quechua-Spanish-English Functional Dictionary Ayacucho-Chanka Vol I / Diccionario Functional Quechua-Castellano-Ingles Ayacucho-Chanka*. Vol. 1. Lima: CSR-PARWA.

Soukka, Maria. 1999. *A descriptive grammar of Noon, a Cangin language of Senegal*. United Kingdom: University of London. (PhD Thesis.)

Stassen, Leon. 2011. The Problem of Cross-Linguistic Identification. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology* (Oxford Handbooks in Linguistics), 90–99. Oxford: Oxford University Press. (doi:10.1093/oxfordhb/9780199281251.013.0006)

Stenson, Nancy. 2008a. *Basic Irish: a grammar and workbook*. London, New York: Routledge.

Stenson, Nancy. 2008b. *Intermediate Irish: a grammar and workbook* (Grammar Workbook Series). London, New York: Routledge.

Stevick, Earl W. & Raymond Setukura. 1965. *Kirundi basic course*. Washington DC: Foreign Service Inst., US Dept. of State.

Stolz, Thomas. 2007. Harry Potter meets Le petit prince – On the usefulness of parallel corpora in crosslinguistic investigations. *Language Typology and Universals* 60(2). 100–117. (doi:doi:10.1524/stuf.2007.60.2.100)

Stolz, Thomas & Levkovych, Nataliya & Urdze, Aina (eds.). 2017. *Spatial interrogatives in Europe and beyond: where, whither, whence* (Studia Typologica 20). Berlin, Boston: De Gruyter Mouton.

Strehlow, C. & Kenny, Anna & Inkamala, Rhonda & Inkamala, Mark & Henderson, John & Moore, David & Australian National University Press (eds.). 2018. *Carl Strehlow's 1909 comparative heritage dictionary: an Aranda, German, Loritja and Dieri to English dictionary with introductory essays* (Monographs in Anthropology Series). Acton, A.C.T: Australian National University Press.

Strehlow, T. G. H. 1942. Aranda Grammar (Continued). *Oceania* 13(2). 177–200.

Suazo, Salvador. 2002. *Conversemos en Garífuna* (Colección Lámpara). 3rd edn. Tegucigalpa, Honduras: Editorial Guaymuras.

Sundermann, Hermann. 1912. Der Dialekt der "Olon Maanjan" (Dajak) Süd-Ost-Borneo. *Bijdragen tot de Taal-, Land- en Volkenkunde van Nederlansche Indië* LXVI. 203–236.

Swift, Lloyd B. & A. Ahaghotu & E. Ugorji. 1962. *Igbo basic course*. Washington DC: Foreign Service Institute, US Department of State.

Tandioy Jansasoy, Francisco & Stephen H. Levinsohn & Alonso Maffla compilers Bilbao. 2006. Diccionario Inga (edición interina en el nuevo alfabeto). Comité de Educación Inga de la Organización "Musu Runakuna."

Taylor, Carrie. 1999. *Pronouns in Nomaande*. Yaoundé: SIL Cameroon.

The World Atlas of Language Structures Online. 2022. Zenodo. (doi:10.5281/ZENODO.7385533)

Thompson, Laurence C. 1965. *A Vietnamese Grammar*. Seattle: University of Washington Press.

Tiedemann, Jörg (ed.). 2011. *Bitext alignment* (Synthesis Lectures on Human Language Technologies 14). S.l: Morgan & Claypool.

Timyan, Judith. 1977. *A Discourse-based Grammar of Baule: The Kode Dialect*. New York: City University of New York. (PhD Thesis.)

Tindall, Henry. 1856. *A grammar and vocabulary of the Namaqua-Hottentot language*. Cape Town: Pike's Machine Printing Office for A.S. Robertson.

Tłįchǫ Community Services Agency. 2007. *Reading and Writing in Tł˛ıcho˛ Yatìì*. Canada: Library and Archives Canada Cataloguing.

Tompa, József. 1968. *Ungarische Grammatik* (Janua Linguarum: Series Practica 96). The Hague: Mouton. (doi:10.1515/9783111358628)

Topping, Donald M. 1980. *Spoken Chamorro*. Honolulu: The University Press of Hawaii.

Topping, Donald M. & Bernadita C. Dungca. 1973. *Chamorro Reference Grammar*. Honolulu: University of Hawaii Press.

Topping, Donald M. & Pedro M. Ogo & Bernadita C. Dungca. 1975. *Chamorro-English dictionary* (Pali Language Texts: Micronesia). Honolulu: The University Press of Hawaii.

Tran, Tri C. & Minh-Tam Tran. 2007. *Chào bạn! an introduction to Vietnamese*. Lanham, Md. [u.a]: University Press of America.

Tuggy, David H. 1979. Tetelcingo Náhuatl. In Ronald W. Langacker (ed.), *Studies in Uto-Aztecan grammar 2: Modern Aztec grammatical sketches* (Summer Institute of Linguistics Publications in Linguistics 56), 1–140. Arlington, Texas: Summer Institute of Linguistics and the University of Texas at Arlington.

Ulfa, Maria & Mulyadi Mulyadi. 2020. Interrogative Construction in Aceh Language. *Jurnal Arbitrer* 7(1). 45–50. (doi:10.25077/ar.7.1.45-50.2020)

Ultan, Russell. 1978. Some General Characteristics of Interrogative Systems. In Joseph H. Greenberg (ed.), *Universals of human language*, vol. 4: Syntax, 211–248. Standford: Stanford University Press.

Ulutaş, İsmail. 2014. An Overview of Relative Clause Constructions Modifying Nouns in Gagauz Syntax. *Tehlikedeki Diller Dergisi* 3(1). 85–91.

Vallejos Yopán, Rosa. 2010. *A grammar of Kokama-Kokamilla*. Eugene: University of Oregon. (PhD Thesis.)

van den Berg, René & Robert L. Busenitz. 2012. *A grammar of Balantak, a language of Eastern Sulawesi* (SIL E-Books 40). Dallas: SIL International.

van der Klis, Martijn & Jos Tellings. 2022. Generating semantic maps through multidimensional scaling: linguistic applications and theory. *Corpus Linguistics and Linguistic Theory* 18(3). 627–665.

van Haitsma, Julia Dieterman & Willard van Haitsma. 1976. *A Hierarchical Sketch of Mixe as Spoken in San José el Paraíso* (SIL Publication 44). Norman, OK: SIL.

van Schaaik, Gerjan. 2020. *The Oxford Turkish Grammar*. Oxford: Oxford University Press.

Varga, Dániel & Péter Halácsy & András Kornai & Viktor Nagy & László Németh & Viktor Trón. 2007. Parallel corpora for medium density languages. In Nicolas Nicolov & Kalina Bontcheva & Galia Angelova & Ruslan Mitkov (eds.), *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005* (Current Issues in Linguistic Theory 292), 247–258. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/cilt.292.32var)

Velupillai, Viveka. 2012. *An introduction to linguistic typology*. Amsterdam, Philadelphia: John Benjamins Pub. Co.

Wälchli, Bernhard. 2007. Advantages and disadvantages of using parallel texts in typological investigations. *Language Typology and Universals* 60(2). 118–134. (doi:doi:10.1524/stuf.2007.60.2.118)

Wälchli, Bernhard. 2010. Similarity Semantics and Building Probabilistic Semantic Maps from Parallel Texts. *Linguistic Discovery* 8(1). 331–371. (doi:10.1349/PS1.1537-0852.A.356)

Wälchli, Bernhard. 2018. 'As long as', 'until' and 'before' clauses: Zooming in on linguistic diversity. *Baltic Linguistics* 9. 141–236. (doi:10.32798/bl.372)

Wälchli, Bernhard & Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710. (doi:10.1515/ling-2012-0021)

Wąsik, Zdzislaw. 1982. Zur strukturellen Typologie der Fragen. *Language Typology and Universals* 35(JG). 466–475. (doi:10.1524/stuf.1982.35.jg.466)

Watkins, Mark Hanna. 1937. A grammar of Chichewa: a Bantu language of British Central Africa. *Language* 13(2). 5–158.

Watters, John Robert. 1981. *A Phonology and Morphology of Ejagham - with notes on Dialect Variation*. Los Angeles: University of California. (PhD Thesis.)

Weber, David. 1989. *A grammar of Huallaga (Huánuco) Quechua* (University of California Publications in Linguistics 112). Berkeley: University of California Press.

Weber, David. 1996. *Una gramática del quechua del Huallaga (Huánuco)* (Serie Lingüística Peruana 40). Lima: Ministerio de Educación and Instituto Lingüístico de Verano.

Webster, Donald H. & Wilfried Zibell. 1970. *Iñupiat Eskimo dictionary*. Fairbanks, Alaska: Summer Institute of Linguistics.

Wheeler, Max & Alan Yates & Nicolau Dols. 1999. *Catalan: a Comprehensive Grammar*. London, New York: Routledge.

Whitehead, G. 1925. *Dictionary of the Car Nicobarese Language*. Rangoon: American Baptist Mission Press.

Wilkendorf, Patricia. 1998. *Sketch Grammar of Nɔmaándé: Sections 1-4.* Yaoundé, Republic of Cameroon: Ministry of Scientific and Technical Research and SIL.

Wilson, Deirdre & Dan Sperber. 2012. Mood and the analysis of non-declarative sentences. In Dan Sperber & Deirdre Wilson (eds.), *Meaning and Relevance*, 210–229. Cambridge: Cambridge University Press. (doi:10.1017/CBO9781139028370.013)

Woollams, Geoff. 1996. *A grammar of Karo Batak, Sumatra* (Pacific Linguistics Series C, 130.). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Wurm, Stefan. 1951. The Karakalpak Language. *Anthropos* 46. 487–610.

Zariquiey, Roberto & Gavina Córdova. 2008. *Qayna, Kunan, Paqarin: Una introducción prática al quechua chanca*. Lima: PUCP.

Zeitoun, Elizabeth. 2007. *A Grammar of Mantauran (Rukai)* (Language and Linguistics Monograph A4-2). Taipei: Institute of Linguistics, Academia Sinica.

Zeshan, Ulrike. 2004. Interrogative Constructions in Signed Languages: Crosslinguistic Perspectives. *Language* 80(1). 7–39.

Zhao, Furong & Guoqing Chen. 2006. *WaYu JiChu JiaoCheng* (佤语 基础 教程) [Basic courses for Wa]. Beijing: ZhongYang MinZu DaXue ChuBanShe (中央 民族 大学 出版社) [China Minzu University Press].

Zhao, Yanshe & Zhao, Fuhe (eds.). 1998. *WaYu YuFa* (佤语 语法) [A grammar of Wa]. Kunming: YunNan MinZu ChuBanShe (云南 民族 出版社) [Yunnan Nationalities Publishing House].