A corpus-based contrastive analysis of modal adverbs of certainty in English and Urdu

Humaira Jehangir

A thesis submitted for the degree of
Doctor of Philosophy in Linguistics

April 2023

Department of Linguistics and English Language

DEDICATED

To my parents *Dr Malik Abdul Jabbar* and the late *Farida Abdul Jabbar Malik*: they encouraged me to dare to dream

To my granddaughters *Eshaal Abdullah* and *Zainab Abdullah*: I leave them my legacy to soar as high as they aspire

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated.

Abstract

This study uses the corpus-based contrastive approach to explore the syntactic patterns and semantic and pragmatic meanings of modal adverbs of certainty (MACs) in English and Urdu. MACs are a descriptive category of epistemic modal adverb that semantically express a degree of certainty.

Due to the paucity of research to date on Urdu MACs, the study draws on existing literature on English MACs for cross-linguistic description of characteristics of English and Urdu MACs. A framework is constructed based on Boye's (2012) description of syntactic characteristics of MACs, in terms of clause type and position within the clause; and on Simon-Vandenbergen and Aijmer's (2007) description of their functional characteristics including both semantic (e.g. certainty, possibility) and pragmatic (e.g. authority, politeness) functions. Following Boye's (2012) model, MACs may be grouped according to meaning: *high certainty support* – HCS (e.g. *certainly*); *probability support* – PS (e.g. *perhaps*); *probability support for negative content* – PSNC (e.g. *perhaps not*); and *high certainty support for negative content* – HCSNC (e.g. *certainly not*).

Methodologically, the framework identified as suitable is one that primarily follows earlier studies that relied on corpus-based methods and parallel and comparable corpora for cross-linguistic comparative or contrastive analysis of some linguistic element or pattern. An approach to grammatical description based on such works as Quirk et al. (1985) and Biber et al. (1999) is likewise identified as suitable for this study.

An existing parallel corpus (EMILLE) and newly created comparable monolingual corpora of English and Urdu are utilised. The novel comparable corpora are web-based, comprised of news and chat forum texts; the data is POS-tagged. Using the parallel corpus, Urdu MACs equivalent to the English MACs preidentified from the existing literature are identified. Then, the comparable corpora are used to extract data on the relative frequencies of

MACs and their distribution across various text types. This quantitative analysis demonstrates that in both languages all four semantic categories of MAC are found in all text types, but the distribution across text types is not uniform. HCS MACs, although diverse, are considerably lower in frequency than PS MACs in both English and Urdu. HCSNC and PSNC MACs are notably rarer than HCS and PS MACs in both languages.

The analysis demonstrates striking similarities in the syntactic positioning of MACs in English and Urdu, with minor differences. Except for Urdu PSNC MACs, all categories most frequently occur in clause medial position, in both independent and dependent clauses, in both languages. This difference is because *hō nahīṁ saktā* 'possibly not' is most frequent in clause final position.

MACs in both languages most often have scope over the whole clause in which they occur; semantically, the core function of MACs is to express speaker's certainty and high confidence (for HCS and HCSNC) or low certainty and low confidence (for PS and PSNC) in the truth of a proposition. These groups thus primarily function as certainty markers and probability markers, respectively. In both languages, speakers also use MACs short responses to questions, and in responses to their own rhetorical questions. HCS and PS MACs in clause final position may in addition function as tags which prompt a response from the interlocutor. When they cooccur with modal verbs, MACs emphasise or downtone, but do not entirely change, the modal verb's epistemic or deontic meaning. In both languages, all MACs preferentially occur in the *then*-clause of a conditional sentence.

Pragmatically, MACs are used for emphasis, expectation, counter-expectation and politeness. Additionally, HCS and HCSNC MACs are used to express solidarity and authority, and PS and PSNC MACs are used as hedges. Readings of expectation, hedge, politeness, and solidarity may be relevant simultaneously. Interestingly, reduplication for emphasis, common in Urdu, is only observed for one Urdu MAC, *żarūr* 'definitely', whereas all English MACs reduplicate for emphasis in at least some cases. Another difference is that, in Urdu, the

sequence *śāyad nahīṁ yaqīnān* 'not perhaps, certainly' expresses speaker authority within a response to a previous speaker, but no English MAC exhibits this behaviour.

Despite overall similarity, minor dissimilarities in the use of English and Urdu MACs are observable, in the use of MACs as replies to questions, and in their use within interrogative clauses. This analysis supports the contention that, cross-linguistically, despite linguistic variation, the conceptual structures and functional-communicative considerations that shape natural languages are largely universal.

This study makes two main contributions. First, conducting a descriptive analysis of English and Urdu MACs using a corpus-based contrastive method both illuminates this specific question in modality but also sets a precedent for future corpus-based descriptive studies of Urdu. The second is its inclusion of priorly considered distinct categories of modal adverbs of certainty and possibility in a single category of modal adverbs that are used to express a degree of certainty, i.e. MACs. From the practical standpoint, an additional contribution of this study is the creation and open release of a large Urdu corpus designed for comparable corpus research, the Lancaster Urdu Web Corpus, fulfilling a need for such a corpus in the field.

Acknowledgements

As I come to the end of my five-year PhD journey, I thank Allah (SWT) first and foremost, for creating this unique opportunity for me and providing me with strength and bringing me into contact with those who helped me throughout this tough but life changing journey – especially because I come from a society where it is unheard of a woman nearing her 50s to take up a PhD project. I would also like to thank all those who played a vital role in supporting me in various ways.

My thanks is due to my mentor and supervisor Dr Andrew Hardie, who guided me beyond the call of duty. During our regular discussions, both on campus and Skype meetings, I greatly benefitted from Andrew's wealth of knowledge, especially but not limited to Corpus Linguistics and English and Urdu grammar. In addition to thesis writing, and counselling on choice of coursework modules, he provided support, where needed, in compiling corpora. I am extremely grateful for his consistent, detailed and very constructive comments on my thesis drafts at every stage. I especially appreciate his acumen in handling even worst of my drafts and firmly steering me towards more coherent writing by channelling my strengths in the right direction. He was always there with a kind word and plausible solution in my hour of need, pulled me up with an encouraging word when he saw that I took some criticism to heart, and encouraged me to participate in conferences and to write papers. Most importantly, I am grateful for his support in helping me complete my thesis in time.

My especial thanks go to my external examiner Professor Karin Aijmer and internal examiner Professor Tony McEnery. I was privileged to engage with them in discussion and intense questioning at my final viva, especially regarding my methodology, analysis and contribution to a corpus-based contrastive analysis. Likewise, I offer thanks to Dr Dimitrinka Atanasova for arranging a smooth Viva event as a Chair of my thesis defence committee. I am grateful to the whole panel because due to technical handicap of internet connectivity (that suddenly erupted in my region as an aftermath of energy crisis), I was dangerously close getting my viva rescheduled. However, despite the technical handicap on my end, they all rose to the occasion (especially Dr Andrew Hardie who kept the communication lines open) and ensured that I could hear and understand their questions and the whole

session went smoothly without any further inconvenience due to the situation. I am eternally indebted to their kindness and professionally managing the situation.

Dr Daniel van Olmen's insightful remarks and helpful suggestions at my pre-Confirmation and post-Confirmation panel meetings proved inspirational for earlier and final drafts of my thesis, especially regarding choice of corpora for my analysis. Dr Andrew Wilson's invaluable advice at my Confirmation panel meeting helped me make informed decisions in refining my theoretical framework. I am also grateful to all my course instructors: Dr Mark Sebba, Dr Daniel van Olmen, Professor Uta Papen, Dr Jenefer Philp, Dr Veronika Koller, Dr Johann Unger, and Dr Diane Potts.

I am grateful to all my friends on the PhD programme at Lancaster University who frequently provided me with good advice, especially Dr Ekaterina Ignatova, Dr Prihantoro, Dr Andressa Gomez, and Orhan Bilgin. I am grateful to all my friends who were there with a kind word, especially Karen Davies, Lexi Webster, Dr Aqsa Isa, Helena Stakounis, Mariana De Luca, Dr Salomé Villa Larenas, and Maya Ahmed. I am thankful for the strong support network of Lancaster University that did not let me feel alone during the Covid times. In fact, we made the best use of those years. This includes Sam Holland and his team, who created an online module to teach Python, helping me in refining my Python script for corpus compilation. I am also grateful for initial guidance in the use of Python to Dr Kunal Mahajan (Columbia University, USA), who was introduced to me by my son Ali Jehangir (Columbia University, USA). I am especially thankful for all the help my son Abdullah Jehangir (NUST, Pakistan) provided when I initially got stuck writing Python scripts. I also wish a special thank you to Elaine Heron for always helpfully assisting with departmental and academic queries.

I am grateful to my family, especially to my father Dr Malik Abdul Jabbar, and my husband Col. (R) Muhammad Jehangir, who disregarded the stereotyping prevalent in our society and whose unstinting support (both moral and financial) has helped me achieve my PhD goal. And, not least, I am grateful to my sons, as well as my daughters-in-law Nayab Abdullah and Dr Fatima Qamar, who cheered me on all the way through.

Contents

List of Tables

List of Figures

List of Abbreviations used in glosses on Urdu examples

=	clitic boundary
1	first person
2	second person
3	third person
ACC	accusative (direct object)
CNT	contraction
DAT	dative (indirect object)
DEM	demonstrative
EMPH	emphatic
ERG	ergative
EXC	exclusive particle *hī*
F	feminine
FUT	future
GEN	genitive
IMP	imperative
INC	inclusive particle *bhī*
INF	infinitive
IPFV	imperfective
M	masculine
NEG	negator
OBL	oblique
PFV	perfective
PL	plural
POSS	possessive
PRS	present
PROG	progressive
PST	past
REFL	reflexive
SBJV	subjunctive
SG	singular
VALA	agentive nominalising enclitic

1 Introduction

1.1 Chapter overview

The topic of this thesis is a comparison of modal adverbs in Urdu and English. In this first chapter, I briefly introduce the topic of modal adverbs and their cross-linguistic study in section 1.2. This is followed by a short introduction to some basic facts regarding the Urdu language in 1.3. Then there is a short discussion on the notion of modality in Urdu in 1.4. I outline the research and methodological aims of this thesis in section 1.5. Finally, I provide an outline of the structure of the thesis in section 1.6.

1.2 Adverbs as modal expressions

This thesis addresses the use of those modal adverbs (e.g. *certainly, probably*) that are used to express specifically *epistemic* modality, that is, a speaker's perception of the truth value of a proposition.

Modality has been approached from various perspectives in theoretical and descriptive linguistics. *Modality* is defined as encompassing the semantic notions of *probability*, *certainty*, and *possibility* (epistemic modality); and *permission*, *volition*, and *obligation* (deontic modality) (Hoye, 1997, p.2). According to Simon-Vandenbergen and Aijmer (2007, p.1), despite extensive work in the area, there remain "several aspects of modality" to be addressed. Similarly, Hoye (1997, p. 1) observes that there remains a need for further insightful research on "the definition, description and analyses of this elusive and fundamental category of human language and thought".

Many studies on modality in English have limited themselves to the use of modal and semi-modal auxiliaries (e.g. *can*, *ought to*) to express modality. For example, Coates' (1983) and Palmer's (2001) studies discuss in detail how English modal auxiliaries are used for both epistemic and non-epistemic modal meaning, and how they create complex patterns with other elements to express additional specific modal meanings. Boye (2016, p. 118) says that in "English and related languages", modality can be expressed through different lexical means, including nouns, lexical verbs, adjectives, and adverbs. Researchers into modality consequently broadened their enquiry to include these modal expressions (e.g. Boye, 2016; Malchukov & Xrakovskij, 2016). Although some attention is given to modal adverbs in major English grammars (e.g. Quirk et al., 1985; Biber et al., 1999), there remained a need for extensive study of semantic and pragmatic function of modal adverbs. Only after a relatively recent increase in corpus-based studies in the area has research on English modal adverbs gained momentum (see 2.3).

Earlier studies of English adverbs that express epistemic modality (e.g. Greenbaum, 1969; Quirk et al. 1985) characterise them as expressions used to convey *conviction* or *doubt* in a proposition. Some later studies (e.g. Simon-Vandenbergen & Aijmer, 2007) classify certain modal adverbs as modals of *certainty* while others (e.g. Van der Auwera et al., 2005) classify certain modal adverbs as modals of *possibility*. Boye (2012) considers all those modal adverbs classified as certainty markers (e.g. *definitely*) or possibility markers (e.g. *perhaps*) a part of the same semantic field, that of expressing a *degree of certainty*. That is, modal adverbs that express an addresser's degree of certainty in the truth of a proposition form a category regardless of what that degree is (see 2.3.2.3).

1.3 Urdu: A brief discussion

Urdu is the national language of Pakistan. Urdu is widely spoken and understood by first language (L1) and second language (L2) speakers throughout Pakistan, India and the South Asian diaspora around the world. Table 1.1 presents demographic data for both L1 and L2 speakers of Urdu, according to the *Ethnologue*[1].

Table 1.1: Urdu language speakers (Source: Ethnologue)

Region	First language speakers	Second language speakers	Total
Pakistan (2018 census)	15,000,000	149,000,000	164,000,000
In all countries	69,006,470	161,045,800	230,052,270

Urdu is closely related to Hindi. Researchers of Hindi and Urdu (Ahmad, 2008; Bhatt et al., 2011; Kachru, 2008; Khan, 2006; Rahman, 2011) agree that, despite the different political and socio-cultural identities associated with these two languages in the present-day world, they have a common history. Both are part of the Indo-Aryan branch of the Indo-European family of languages. Both are variants of the New Indo-Aryan language spoken in and around Delhi during its time as capital of the Mughal Empire (11^{th} – 19^{th} century) and the British Raj (19^{th} – 20^{th} century). During the Raj, the names Urdu and Hindi became attached to this language as spoken by the Muslim and Hindu communities respectively. In fact, the present-day spoken forms of Hindi and Urdu are mutually intelligible, because they share phonology, morphology, syntax, and core vocabulary. As such, they are sometimes considered a single language, called Hindustani commonly or Hindi-Urdu in linguistics. They are distinguished in written contexts and in formal or technical speech by two factors. The first is orthography: the Perso-Arabic script is used for Urdu, and the Devanagari script for Hindi. Second, there exist differences in non-core vocabulary, especially technical, legal, and

[1] https://www.ethnologue.com/language/urd

religious. Speakers of Urdu borrowed many words of this kind from other languages spoken by predominantly Muslim communities, namely Arabic, Persian and to a lesser extent Turkish. In reaction, Hindi speakers borrowed words from Old and Middle Indo-Aryan, especially Classical Sanskrit (Farooqi, 2008).

In this thesis, due to the lack of much work to date on Urdu lexical modal markers, I have consulted research works on Hindi where relevant. I have not, however, followed the common practice among linguists of referring to the two languages simply as Hindi-Urdu. Rather, I refer specifically to Urdu because (1) nearly all the data to be considered is written, and in writing the two are unambiguously distinct; (2) many of the modal adverbs to be considered are among the vocabulary Urdu has borrowed from Arabic and Persian.

Urdu grammar, like that of other Indo-Aryan languages, is different from English and other languages of the western branches of Indo-European. Three main differences between English and Urdu are word order, gender and complexity of morphological marking. In English, the typical word order is subject-verb-object (SVO); in Urdu it is subject-object-verb (SOV).

Unlike English, where gender agreement is on the basis of natural categories, Urdu has a system of grammatical gender. Urdu has two genders – feminine (f) and masculine (m). All nouns are inherently classified as one or the other, and for some nouns this is explicitly evident in their morphology. Other categories such as adjectives and participles are gender-marked to agree with the nouns they qualify. Two formal categories of adjective exist: marked and unmarked. The former always carry a suffix indicating gender, number, and case (Schmidt 1999, p. 32); the latter are invariant. Meanwhile, all verbs are inflected to agree with their subject noun or pronoun in *either* number and person *or* gender and number (Schmidt, 1999, p. 111).

As this may suggest, Urdu is more morphologically complex than English. This is due to the need of many word categories to show agreement in gender, number, and case. Additionally, noun, pronoun and adjective declension involves markings for the nominative, oblique, and (marginally) vocative cases. Further distinctions of case (ergative, accusative-dative, ablative and so on) are marked by postpositions, which are generally considered to be clitics (Butt, 1995, p. 9). Some grammatical postpositions are also marked for number and gender, historically because of derivation from Old Indo-Aryan nouns or participles (e.g. nominative masculine singular *kā*, feminine *kī*, and masculine oblique or plural *kē*, all meaning 'of'). Finite verbs in Urdu may be marked for tense, aspect, and mood, but as in English (Schmidt, 1999), many such categories are expressed by complexes of a main verb with one or more auxiliaries, the most important of which is *hō* 'be'.

1.4 The notion of modality in Urdu

The area of modality has yet to be extensively researched in Urdu specifically, but modality as a grammatical phenomenon exists in all languages including Urdu. For Urdu, some research on modal verbs has addressed their syntactic placing and the meaning that they convey (Bhatt et al., 2011). However, the focus of these studies has been exclusively modal verbs or their use in combination with other auxiliaries.

Bhatt et al. (2011, p.2) say that Urdu has only two dedicated modal verbs: *cahiē* 'need/should' and *sak* 'can/be able to' (the former, but not the latter, is inflectionally deficient). Modality is conveyed through constructions that include these two modal verbs. Schmidt (1999, pp. 115-16) further classifies *pā* 'find' as a modal verb. In her account, the modal verbs *sak* 'can/be able to' and *pā* 'find' are used in constructions which convey epistemic possibility.

In these constructions, a nominative case subject is used with a root form main verb and modal verb to show possibility (Bhatt et al. 2011, p.2), as in (1).

1. Yasīn voh kar sakā
   Yasin DEM do can.PFV.M.SG

   "Yasin could do that" (Bhatt et al., 2011, p. 2).

Similarly, the construction ROOT + *pānā* expresses "the possibility of an action dependent on circumstances" (Schmidt, 1999, p. 116). Schmidt reports that this construction typically occurs in negative sentences conveying the improbability of some action, as in (2). Example (2) also illustrates how the future tense is used to express epistemic modality in the form of predictions.

2. Vahīd masrūf hai, kal=kī dāvat=mēṁ
   Wahid busy be.PRS.3.SG, tomorrow=GEN.F.SG party=in

   nahīṁ ā pāē gā
   NEG come find.SBJV.3.SG FUT.M.SG

   "Wahid is busy; he *can't* [*lit. won't manage to*]come to tomorrow's party" (Schmidt, 1999, p. 116).

While reviewing the literature on modality, I did not find any substantial research on adverbs as modal expressions in Urdu. Only reference and pedagogical grammars (e.g. Schmidt, 1999) note that Urdu speakers employ any other means than verbs to express modality. Schmidt (1999) shows that modal adverbs are one of the most common means of expressing modality in Urdu, but even she does not supply the much-needed detailed account of the various modal functions that adverbs can perform. As the following section will outline, the aim of this thesis is to investigate this phenomenon of modal adverbs in Urdu, where they have been much less studied than in English.

1.5 Rationale for choosing modal adverbs for analysis

In Urdu, there are two modal verbs: *cāhīē* 'want' and *saknā* 'can'. Some Hindi-Urdu researchers also count as modal other verbs such as *pānā* 'get' and/or *ānā* 'come' (Genady, 2005, p. 28; Schmidt, 1999, p. 116). Some count as modal devices certain nouns (e.g. *irādā* 'plan'), adjectives (e.g. *mūmkīn* 'possible'), and particles (e.g. exclusive *hī*). But there is no clear consensus on the operation of modality in Hindi-Urdu. Moreover, my personal experience of discussing modality in Urdu with linguistics researchers in Pakistan is that current understanding of the issue is generally unsatisfactory. Therefore, when I began my PhD, I opted to investigate modality in Urdu.

However, as I started my review of the existing literature, I came to fully understand how broad and diverse a topic research into modality truly is. Therefore, I decided that for a focused research aim, I needed to investigate just one category of element – modal adverbs – and one semantic type of modality – epistemic modality. I did not choose modal adverbs randomly. My initial work with English/Urdu parallel corpus data showed me that modal adverbs are pervasive in Urdu, although current reference grammars do not categorise them as modals and the minimal literature on Hindi-Urdu modal expressions focuses instead on the less abundant modal verbs. Only one work (Genady, 2005) has to some extent analysed the category of modal adverbs in Hindi-Urdu. By contrast, there has been considerable research on modal adverbs in English. This unfortunate state of the literature inspired me to conduct a contrastive analysis of English and Urdu modal adverbs. My goal became to identify both similarities and differences between the two languages in terms of how epistemic modal adverbs are used.

1.6 Aims of this study

In this section, I discuss both the main research aims and the methodological aims of my thesis. I postpone the statement of specific research questions to the end of the literature review (section 3.6), since they utilise many terms and concepts that are explained in my survey of the relevant literature.

1.6.1 Research aims

One of my main reasons for studying modal adverbs that express a degree of certainty is that, while studies of Urdu adverbs generally do exist, research specifically on such modal adverbs is almost non-existent. On the other hand, there is a substantial literature on modal adverbs conveying degree of certainty in English (see 2.3). The focus of this thesis project is to undertake a contrastive, corpus-based descriptive analysis of those English and Urdu modal adverbs that are used to express degree of certainty. Such modal adverbs express a speaker's estimation of the likelihood of the content of the proposition being true (Huddleston & Pullum, 2002, p.52). Alternatively, we may say that they express a speaker's judgement about the truth of some possible world (Simon-Vandenbergen & Aijmer, 2007, p.27).

The thesis's primary aim is thus to explore similarities and differences between English and Urdu *modal adverbs of certainty* (henceforth, MACs). Specifically, I will analyse parallel and comparable corpora (see 4.3) to examine these elements from a contrastive descriptive perspective. As a starting point, I will consult previous literature and investigate parallel corpus data to ascertain what lexical items and phrases constitute the corresponding sets of English and Urdu MACs. After that, I will mainly utilise monolingual comparable

corpora of English and Urdu. In particular, I will observe the syntactic environments of both English and Urdu MACs to understand their similarities and differences. This analysis will involve both comparison and contrast among the various MACs, *within* each language as well as cross-linguistically. On the basis of the observed syntactic behaviour of English and Urdu MACs (reported in Chapter 5), I will describe and analyse their semantic, rhetorical-pragmatic functions in Chapters 6 and 7. Collectively, these processes will answer my research questions, as laid out in section 3.6.

1.6.2 Methodological aims

In the field of linguistics in Pakistan, there is a growing interest in the study of features and characteristics of languages other than English. For instance, there are now studies on Urdu (e.g. Rahman, 2011) and other regional languages (e.g. Zaidi, 2010), including endangered languages of the region (e.g. Ali, 2015). Whilst these languages had previously attracted attention in the international contexts of field linguistics and language typology, amongst Pakistani linguists more focus had been given to matters such as English language teaching (e.g. Warsi, 2004) and English in the context of Pakistani society (e.g. Mansoor, 1993). But many contrastive studies of English and Urdu have now been published. Examples include a study of speech acts in Urdu and English (Akram, 2008), a contrastive analysis of the adpositional systems of Urdu and English (Bilal et al., 2013), and a study of meta-discourse markers in English and Urdu newspapers (Shafique et al., 2019). However, to my knowledge, no descriptive study to date contrastively analyses any grammatical feature in English and Urdu using corpus techniques. There is a need for such corpus-based cross-linguistic analyses to enrich the neglected area of empirical and comparative research on Urdu.

This study will fill this methodological and analytical gap. Thus, one of the objectives of this research is also to demonstrate the efficacy of corpus-based contrastive analysis of a grammatical category in English and Urdu. While this thesis targets the use and function specifically of MACs in English and Urdu, it is hoped that it will also provide a model for other corpus-based contrastive studies of this language pair in the future.

1.7 Outline of the thesis

This thesis comprises nine chapters.

Chapter 2 is a literature review on modality and modal adverbs. In sections 2.2 I review some background on modality, including definitions of modality and epistemic modality. This is followed by sections 2.3 and 2.4 on the formal and functional characterisation of English MACs that includes discussion on the semantic and pragmatic functions of MACs respectively. Section 2.5 addresses Urdu MACs specifically. Then in section 2.6, basing on the review of the existing literature on the formal and functional characteristics of MACs, I present my theoretical framework. Finally, in section 2.7, I summarise the research to date into characteristics of MACs.

Chapter 3 is a literature review on the corpus-based contrastive approach. In section 3.2 a brief overview of corpus linguistics and corpus-based research is given. This is followed by some discussion of corpus-based descriptive grammar in section 3.3. Then, section 3.4 presents a review of the literature on corpus-based contrastive studies of grammar. Section 3.5 outlines a framework for the analysis of MACs. Finally, in 3.6, I present the research questions which arise from the literature reviews in chapters 2 and 3, and which the remainder of the thesis addresses.

Chapter 4, the methodology, is comprised of two sections, dealing respectively with the data and the procedural steps used in this research. Section 4.2 introduces existing corpora of English and Urdu, as well as my compilation of new comparable corpora. The procedural steps described in section 4.3 include the use of parallel corpus data to identify English and Urdu MACs for this study, and the specific methods employed for quantitative and qualitative analysis of English and Urdu MACs. Having presented the methodology, I establish which English MACs and corresponding Urdu MACs will be analysed in later chapters (thus answering my RQ1).

Chapter 5, the syntactic analysis, answers my RQ 2. Through scrutiny of concordance lines from the comparable corpora, and calculation of frequencies, I describe the placement of English and Urdu MACs at various positions in different types of clauses. Sections 5.2. and 5.3 discuss clause level positions of MACs in independent clauses in English and Urdu respectively. Sections 5.4 and 5.5 discusses clause level positions of MACs in dependent clauses in English and Urdu respectively. In section 5.6, I present the combined data on clausal positioning.

Chapter 6, the semantic analysis, answers my RQ 3 parts I and II. By qualitative analysis of concordances from the comparable corpora, I investigate what meanings are conveyed by English and Urdu MACs, and how this is influenced by their occurrence at different clausal positions as well as their interaction with other elements of the clause.

Chapter 7, the pragmatic analysis, answers my RQ 3 part III. By investigating concordances (in extended context) from the comparable corpora, I present different pragmatic functions of English and Urdu MACs across sections 7.2 to 7.8.

Chapter 8, the discussion, answers my RQ 4 based on the syntactic, semantic and pragmatic analyses in chapters 4 to 7. In section 8.2, I discuss the distribution of English and

Urdu MACs across text types. Then, in section 8.3, I discuss the similarities and differences between clause positioning of English and Urdu MACs. In sections 8.4 and 8.5, I explore similarities and differences in the semantic and pragmatic functions of English and Urdu MACs.

Chapter 9 is the conclusion of the thesis. I summarise my findings and the answers they provide to my research questions (section 9.2). I also discuss the significance of this work for descriptive contrastive analysis of MACs in English and Urdu. In section 9.3, I discuss certain limitations of my thesis research. Finally, in section 9.4, I suggest some directions for future research, both in this area and in corpus-based contrastive analysis more generally.

2 Literature Review: Modal adverbs of certainty

2.1 Chapter overview

In this chapter, I review the existing literature on *modal adverbs of certainty* (MACs). The outline of the chapter is as follows. In section 2.2, I introduce the background on modality, including the definitions of modality, epistemic modality and MACs utilised in this thesis. In section 2.3, I review previous studies on the formal and functional characteristics of English MACs, including the prior literature on semantic categorisation of English MACs and on MACs' scope over negation. In section 2.4, I present the pragmatic functions of English MACs as described in these existing studies. In section 2.5, I give a necessarily short review of research to date on Urdu MACs. In section 2.6, I present the parameters (syntactic, semantic and pragmatic) of the theoretical framework that I will follow in my cross-linguistic description of the characteristics of English and Urdu MACs. This theoretical framework is based on the review of literature on English MACs (sections 2.4 to 2.6). Finally, in section 2.7, I summarise the chapter.

The English language examples used in this chapter (unless otherwise cited) are from the British National Corpus (BNC 1994) accessed via Lancaster University's CQPweb server[2].

[2]https://cqpweb.lancs.ac.uk/

2.2 Defining the notion of epistemic modality and modal adverbs

2.2.1 Modality

Modality can be defined, for present purposes, as a grammatical function that expresses the *attitude* and *stance* of a speaker (or of the subject of a clause) towards a *proposition* expressed in a clause or sentence (see Biber et al., 1999; Palmer, 2001). Biber et al. (1999, p.153) define *stance* as "the degree of certainty, or meaning such as obligation, necessity or giving or asking permission"; thus, modality encodes some human being's attitude towards the likelihood, unexpectedness, obligatoriness, and/or necessity (among other factors) of the proposition. Similarly, Lyons (1977, p.452) says that *modality* refers to a speaker's or writer's stance "about a proposition that a sentence expresses, or the situation that a proposition describes". A *proposition* is the primary information conveyed by a clause: an assertion of an event, general situation, or (set of) circumstances; another term used in discussions of modality for the primary content of a clause is the clause's *state of affairs*.

Expression of epistemicity is a core function of modality (Palmer, 1987, p. 8). The other two types of modality prominent in the literature are *deontic* and *dynamic modality*. As these two are not directly relevant to the present research, I introduce them only very briefly here. *Deontic modality* expresses the relationship between the state of affairs (SoA) and norms, expectations, and/or speaker (or clause subject) desires. Nuyts (2016, p. 36) defines deontic modality as "an indication of the degree of moral desirability of the state of affairs expressed in the utterance, typically but not necessarily on behalf of the speaker" though speakers may report on others' deontic assessments. By *moral desirability* Nuyts means both personal ethical criteria and societal norms. Additionally, deontic expectations can be in

terms of "institutional laws" (Kärkkäinen, 2007, p. 150). As traditionally defined, deontic modality expresses obligation (3) or permission (4) (Palmer, 2001, p. 10).

3. "You *must* decide" (BNC_A0N 1841).

4. "If you have any need of me you *may* enquire of my nephew the vicar" (BNC_A01 176).

Dynamic modality "relates to the ability or willingness, which comes from the individual concerned" (Palmer, 2001, p. 9). Dynamic modality relates to ability/capability as opposed to epistemic possibility: see example (5).

5. "As a result, the British *can* write novels and plays" (BNC_A0D 525).

2.2.2 Epistemic modality

Epistemic modality expresses the speaker's judgement of the truthfulness of a given proposition, or equivalently the speaker's estimation of the likelihood of the content of the proposition. The two basic epistemic divisions are *possibility* (6) and *necessity* (7) (Palmer, 2014, p. 50).

6. "Mary *may/can* come tomorrow" (epistemic possibility) (Palmer, 2001, p. 91).

7. "There *must be* several other reasons (epistemic necessity)" (Palmer, 2001, p. 72).

However, Palmer (2014, p.51) also says that epistemic modality cannot be limited to just the notions of possibility and necessity. Rather, it is a broader concept that includes "any modal system that indicates the degree of commitment by the speaker to what he says". Under the heading *modal system*, Palmer (2001) includes various lexical means (e.g. verbs, adverbs) that can be used to express a wide range of modal notions.

Epistemic modality can be expressed in English by expressions such as modal verbs (MVs), as in examples (3) to (6); modal adjectives (e.g. *possible*, *certain*); and modal adverbs (e.g. *certainly*, *perhaps*). As this study is focused on modal adverbs that express certainty, that is, MACs, the reminder of this literature review will consider formal and functional characteristics of MACs to the exclusion of other modal expressions.

A notion closely related to epistemic modality, with notable semantic overlap, is *evidentiality* (Nuyts, 2016, p. 51). *Evidentiality* is a grammatical category by which is expressed the source of the information presented in an utterance, that is, the nature of the speaker's evidence for believing that the proposition is so. The source of information can be the speaker's direct perception, inference, hearsay, or testimony (Sqaurtini, 2016, p. 58). For instance, in "I *saw* John playing soccer", the verb *see* asserts evidence acquired through direct perception, whereas in "*Apparently*, John has played soccer", the adverb *apparently* expresses indirect knowledge of the truth of the proposition (Squartini, 2016, p. 58). Thus, the meanings expressed by evidentiality include both direct evidence (e.g. perception) and indirect evidence (e.g. mental reasoning). The relationship between evidentiality and epistemic modality is a matter of some debate. Nuyts (2016, pp. 39-40) notes four approaches to this question. Some authors subsume evidentiality within the category of epistemic modality (e.g. Quirk et el., 1985). Some, on the other hand, treat them as distinguishable, but associate evidentiality closely with epistemic modality and declare the boundary to be fuzzy, such that both are part of a single "supercategory" (e.g. Hengeveld, 1989). Others see them as wholly separate categories of modality (e.g. Narrog, 2005). Yet others exclude evidentiality from the set of modal categories altogether (e.g. Aikendvald, 2004).For purposes of this thesis project I adopt the first of these approaches.

2.3 Modal adverbs of certainty (MACs)

MACs are those epistemic modal adverbs (e.g. *certainly, possibly*) that semantically express a degree of certainty regarding the truth of a given proposition (Simon-Vandenbergen & Aijmer, 2007). Simon-Vandenbergen and Aijmer (2007, p. 3) broadly define those elements as epistemic modals which express a speaker's attitude to the proposition. By Simon-Vandenbergen and Aijmer's definition, MACs are a type of epistemic certainty marker, because they mainly express a high or low degree of speaker's commitment to the truth of the proposition. Boye (2012, p. 158) says that cross-linguistically, MACs are polyfunctional, but among their functions, the most prominent meaning is that of epistemic modality.

Urdu, like English, possesses modal adverbs (e.g. *yaqīnān* 'certainly', *śāyad* 'possibly') that convey a speaker's or writer's degree of certainty regarding the truthfulness of a proposition. There have been studies of both formal and functional characteristics of English MACs. Formal features discussed in these studies include these adverbs' position at clause level and their placement in various dependent and independent clauses (see 2.3.2.2, 2.3.2.3). Meanwhile, functional characteristics of MACs include or relate closely to features such as scope (see 2.3.3.1), different semantic meanings (see 2.3.3.2 and 2.3.3.3) and negation (see 2.3.3.4). Before moving on to those issues, however, I will briefly consider the issue of MAC phrases that act as units of meaning.

2.3.1 MAC phrases

In traditional grammar, a *phrase* is a grammatical unit of one or more words that is one of the classes of constituent that form a sentence. Each phrase type is named after the

word class which plays the primary role in its structure, known as its head. Thus, in a noun phrase, the noun is the head and is an obligatory element, while all other elements are optional and modify the head (Leech, 2006, p. 50). For example, in *a beautiful child*, *child* (noun) is the obligatory head, while *beautiful* (adjective) and *a* (determiner) are both optional elements. Generally, a phrase performs the same grammatical function as its head, for example, a noun phrase has the grammatical function of a sole noun.

However, the term *phrase* also has a broader meaning. A phrase can be a sequence of adjacent words, regardless of the grammar, that possesses some specific meaning as a whole. I use the term *MAC phrase* to refer to a phrase of this type whose consistent meaning is the expression of epistemic modality, and which has adverbial function within the clause. A MAC phrase consists of a group of words that repeatedly co-occur in a specific sequence or pattern, and expresses the relevant modal meaning. Thus, MAC phrases have the same function as a single-word MAC, in that they convey some degree of certainty regarding a proposition. In most or all English studies of MACs, such phrases are treated identically to single-word MACs. For instance, Simon-Vandenbergen and Aijmer (2007, p. 156, p. 173, p. 207) refer to *of course* as a "word" and one of the adverbs of certainty that they address. Simon-Vandenbergen and Aijmer (2007, pp. 236-37) also call *no doubt* an adverb. However, when they discuss longer patterns that contain *no doubt*, they call such a longer combination of words a "sequence" . Thus, for instance, Simon-Vandenbergen and Aijmer (2007, p.236) say that there are several sequences that incorporate *no doubt* and express "epistemic judgment: *there is no doubt, no doubt about it, I have no doubt*".

Addressing the equivalent category of words in Hindi-Urdu, Genady (2005, p.181) says that when *hō* 'be' is used with existential meaning, and combined with *sak* 'can' in a

clause which functions as an adverbial within some larger clause, it has the contextual meaning of possibility, equivalent to the single-word MAC *śāyad* 'perhaps', as in (8).

8.	Hō	saktā	hai	maiṁ	hī	ā	jāōṁ
	Be.SBJV.3.SG	can.IPFV.M.SG	be.PRS.3.SG	I	EXC	come	go.SBJV.3.SG

"Perhaps I will come" (Genady, 2005, p. 146).

But to date, there has been no work specifically on MACs in Urdu.

2.3.2 Formal characterisation of MACs in English

2.3.2.1 Clause internal positioning of MACs

Before addressing the positions that MACs occupy in a clause, two terms used to discuss clausal elements must be clarified: *obligatory* and *optional*. Obligatory elements are those elements in a clause that are necessary to impart the clause's meaning, such as its subject or its main verb. Optional elements, on the other hand, are those "that can be omitted without seriously injuring the meaning" (Biber et al., 1999, p.79), such as adverbial prepositional phrases (e.g. *in the morning*). However, whether an element is obligatory or optional is dependent on the context it is used in. For example, in imperative clauses (defined by Leech, 2006, p. 51, as clauses used for commands or directives), the subject is an optional element (Huddleston & Pullum, 2002, p. 924).

In this study, I follow Biber et al.'s (1999, pp.770-71) practice of distinguishing three main positions that MACs may take at clause level: the *initial*, *medial* and *final* clause positions. A MAC is considered to be in initial position in a clause if it occurs prior to any obligatory element of the clause, that is, the subject (9) or finite verb (10). Any position between the subject and the last obligatory element in a clause is considered to be clause

medial, e.g. between the subject and main verb, as in (11) and (12), or between the verb and object, as in (13). A MAC is considered to be in clause final position when it occurs after all obligatory elements, as in (14). If a MAC is followed by optional element(s), it is not the absolute last element in its clause, but would still be considered to be in final position.

9. "*Possibly* a smell similar to pear drops will also be present on your dog's breath" (BNC_A17 779).

10. "*Certainly* knows a thing or two, that chap" (BNC_EEW 1347).

11. "I would *definitely* need to try to find some help from somewhere" (BNC_A0F 2374).

12. "She *maybe* floated in" (BNC_AD1 740)

13. "Out of that 4,000 I saw *maybe* a dozen or so women" (BNC_EG0 2091).

14. "What excuse could he use ? 'I just popped in to borrow a book' *perhaps* ?" (BNC_AVC 2450).

Numerous researchers (*inter alia*, Halliday & Matthiessen, 2004; Hasselgård, 2010; Simon-Vandenbergen & Aijmer, 2007) have observed that a MAC may perform different functions depending on which position it appears in. For example, Simon-Vandenbergen and Aijmer (2007, p.132) say that if *sure*ly appears in clause initial position, it has an intensifying effect, as in (15); whereas in final position *surely* may function as "inviting confirmation" from the hearer/reader, as in (16).

15. "*Surely* there is nothing easier than to imagine trees … in a park … and nobody by to perceive them" (BNC_ABM).

16. "He could be apprenticed to his mother's brother, *surely*" (BNC_CCM 1846).

A MAC's placement within the clause may also determine whether it has scope over one or more particular elements of the clause, or over the clause as a whole (see 2.3.3.1 for a definition of *scope*).

2.3.2.2 Distribution of MACs in different types of independent clauses

Boye (2012) says that, of the primary types of independent (or main) clauses, English MACs tend to occur in declarative clauses and interrogative clauses but not imperative clauses. Boye (2016, p. 137) reasons that MACs are incompatible with imperative clauses because imperatives do not have "truth-valued meaning" – they are not used to assert the truth of a proposition, as the awkwardness, or incorrectness, of example (17) illustrates. On the other hand, both declarative and interrogative clauses *are* used to convey propositions, because "declaratives signal the assertion of propositions", and "interrogatives signal that they [propositions] are questioned" (Boye, 2016, p.138), as examples (18) and (19) illustrate.

17. "*Be calm, *evidently*!"[3] (Boye, 2012, p. 201).

18. "Nero was *certainly/probably/possibly* a great musician." (Boye, 2012, p.199)

19. "Is Carlsberg *perhaps* the best beer in the world?" (Boye, 2012, p. 200).

2.3.2.3 Distribution of MACs in different types of dependent clauses

Boye (2016, p.138) observes that, of the various types of dependent clause, English MACs occur in adverbial clauses (defined by Leech, 2006 as dependent clauses which add extra meaning about the containing clause's state of affairs), as in (20)[4]; in parenthetical

[3] '*' is the means by which Boye marks this example as ungrammatical, as per standard practice.
[4] In examples 20-23, the MAC and dependent clause marker are both italicised.

relative clauses (also called non-restrictive or non-defining clauses, per Leech, 2006 these clauses provide additional, optional information on a nominal in their containing clause), as in (21); and in complement clauses (defined by Leech, 2006 as clauses which act as the complement of a verb, adjective or noun), as in (22). Complement clauses are often introduced by complementiser *that* (a *that*-clause), but can also be *wh*-clauses or non-finite clauses (*-ing*-clauses or infinitive clauses). *Wh*-clauses are defined by Leech (2006, p.124) as those dependent clauses which begin with a *wh*-element, as in the relative clause in (23).

20. "Ruth had to smile *because* it *certainly* didn't sound like her" (BNC_JY4 598).

21. "Keeping a pet means added responsibility, *which perhaps* is good, if you can cope with it" (BNC_BNL 2041).

22. "My anticipation is *that* we will *certainly* see a rise on Monday" (BNC_A1E 312).

23. "We have a drink here *which* is *definitely* non-alcoholic — it's called cherry brandy"(BNC_AR8 813).

Some researchers (Simon-Vandenbergen & Aijmer, 2007; Suzuki, 2015) have focused on placement of MACs in different clause positions, while others (Boye, 2012; Boye, 2016) go beyond clause-internal positioning and discuss the tendency of MACs to occur in different types of clauses. The position that a MAC takes within a clause, and the type of clause it appears in, help in establishing its functional characteristics (Boye, 2012, p. 100).

2.3.3 Functional characterisation of MACs

In this section, I discuss the semantic features of MACs. I first present the notion of semantic scope, including the scope of MACs over negation; and then the semantic

classification of English MACs based on degree of certainty as given in major reference grammars. Then I discuss categorisation of MACs by degree of certainty in more recent theoretical work. Finally, I discuss the notions of contraries and contradictories.

2.3.3.1 The notion of semantic scope

Boye (2016, p. 135) defines the *semantic scope* of an element "as the range of expressions or meanings to which it applies, semantically". A MAC having semantic scope over an element or a clause thus implies that its meaning, the expression of a degree of certainty, applies specifically to that element or clause, that is, "the latter meaning being in the scope of the former" (Boye, 2012, p. 70) – and not to any other element.

Due to their epistemic meaning, expressions like MACs normally "have a proposition as their implicit scope" (Boye, 2012, p. 185) since, as discussed above, epistemic modality relates to confidence or certainty in propositions. On the other hand, Simon-Vandenbergen and Aijmer (2007, p.82) say that repositioning a MAC within its clause shifts what elements are within its scope, to express different meanings. For example, Simon-Vandenbergen and Aijmer (2007, p.86) say that a MAC in medial position "has the whole of the proposition within its scope", and that its function there is to emphasise "the truth of the proposition as a whole", as illustrated in (24).

24. "No I think I **would** *certainly* want to live with someone that could understand one's own angst and anxieties (ICE-GB: S1A-056/244)" (Simon-Vandenbergen & Aijmer, 2007, p.86).

But according to Simon-Vandenbergen and Aijmer (2007, pp.87-88), there are instances where a MAC has scope only over the directly following element (e.g. the subject,

or an adverbial) and that in such cases it functions as a "focaliser". The scoped element is, in that case, a *marked theme*, as in (25).

25. "So they are setting a better example for the many school children that we have here in the crowd *certainly* **at uh this stage** (ICE-GB: S2A-010/138)" (Simon-Vandenbergen & Aijmer, 2007, p.88).

Simon-Vandenbergen and Aijmer (2007, p. 300) also point out that once a MAC is used as the focaliser of a particular scoped element, its scope over the whole clause weakens, that is, "the meaning of speaker's commitment to the truth is still there but it has weakened". In other words, Simon-Vandenbergen and Aijmer's (2007) claim is that when MACs put focus on a particular clause element, their scope over the rest of the clause is reduced but it does not end.

In certain cases, when a negator occurs after a MAC, the MAC has scope over that negator as well as the proposition. Squartini (2016, p. 64) says that modality and negation are two separate notions but have "potential semantic overlap". Boye (2012, p. 27) says that in addition to positive and neutral epistemic meanings covering certainty, probability, and possibility, there is also "negative epistemic meaning", and this is what is conveyed when negation is within the scope of some epistemic modal expression. Negation can appear both outside (26) and inside (27) the scope of a MAC; in the former case, it is the assessment of the proposition's truthfulness which is negated, not the proposition or element in scope.

26. "Nevertheless, it is *not certainly* fictitious" (BNC_HXX 1133).

27. "Who gives a damn, *certainly not* me" (BNC_A0L 1751).

Of these two, it is negation within the scope of a MAC that is relevant to the present study. According to Simon-Vandenbergen and Aijmer (2007, p. 90), negation within the

scope of a MAC tends to represent a speaker's/writer's negotiation of the truth value of the proposition. Negation within the scope of a MAC extends the scale of degrees of certainty and possibility that MACs may express beyond positive (or neutral) commitment to a proposition, to positive (or neutral) commitment *to the negative content* of a proposition (see 2.3.3.3 and 2.3.3.4; see also Boye, 2016, p. 117).

2.3.3.2 Degree of certainty categorisation of MACs in reference grammars

The classification of MACs in English varies widely across different authors. This can be illustrated by considering the treatment of MACs within three major reference grammars of English: Quirk et al. (1985), Biber et al. (1999), and Huddleston and Pullum (2002).

One of the earliest classifications of certain adverbs as adverbs of certainty was by Greenbaum. Greenbaum (1969, p. 94) classes adverbs of certainty as among the "attitudinal disjuncts" that "express the speaker's attitude to what he is saying, his evaluation of it. Or shades of certainty or doubt about it". Furthermore, Greenbaum (1969, p. 203) classifies adverbs of certainty into two groups according to the degree of certainty they impart: those that express some degree of doubt (e.g. *possibly, allegedly, supposedly*) and those that express some degree of conviction (e.g. *certainly, admittedly, surely*).

Quirk et al. (1985, p.440) define *disjuncts* as those adverbs that semantically "express an evaluation of what is being said" in terms of the way it is communicated or its meaning. Disjuncts are recognised as "the speaker's authority or comment on, the accompanying clause" (Quirk et al., 1985, p. 440). Quirk et al. (1985, pp. 620-21) refine Greenbaum's classification by distinguishing two categories of disjuncts: *content* disjuncts and *style* disjuncts. The *content* disjuncts, among which are the adverbs I label MACs, provide the speaker's comment on or attitude to the content of an utterance (Quirk et al., 1985, p. 615).

Quirk et al. (1985, pp. 620-23) sub-divide these content disjuncts according to whether they express a *degree of truth* or a *value judgement.* Among the *content disjuncts of truth* they identify a semantic division between *content disjuncts of conviction* (e.g. *definitely, undoubtedly*), *content disjuncts of doubt* (e.g. *perhaps, presumably*), and *content disjuncts of reality* (e.g. *apparently, fundamentally*). The first two of these, but not the third, are made up of MACs. Interestingly, Quirk et al. (1985, p. 622) mention some MACs (e.g. *certainly, obviously, of course, no doubt, really*) among the *value judgement* disjuncts as well. This creates some ambiguity because Quirk et al. do not actually explain how the MACs in question express value judgement, only explaining their use to imply certainty (see Quirk et al., 1985, p. 623).

Biber et al. (1999, p. 854) use a different typology. Among the group of *epistemic stance adverbials* is a category, *certainty/doubt*, for epistemic markers that express either certainty or "various levels of probability" (Biber et al., 1999, p. 854). They also include evidentials among the *epistemic stance adverbials* but in a separate sub-group for evidentials (i.e. *source of knowledge*) alluding "to evidence supporting the proposition" Biber et al. (1999, p. 855).

Huddleston and Pullum (2002, p. 768) term the adverbs that indicate the speaker's attitude towards the *likelihood* or *certainty* of a situation or action described in a sentence, such as *likely* or *definitely*, as *modal adjuncts*. However, unlike Quirk et al.'s (1985) and Biber et al.'s (1999) sub-categorisation of adverbs of certainty according to semantics, Huddleston and Pullum sub-divide their modal adverbs according to degree of certainty. This is somewhat similar to Quirk et al.'s (1985) distinction of conviction versus doubt among content disjuncts, but unlike Biber et al.'s (1999) treatment of adverbs of certainty and doubt as just one category. But where Quirk et al. (1985) have categories for *conviction* and *doubt*, Huddleston and Pullum (2002, p. 768) have four categories for level of "the speaker's

commitment to the truth of the proposition": *strong modals, quasi-strong modals, medium modals* and *weak modals*.

The three grammars also differ in their treatment of evidential adverbs (see 2.2.2). Those modal adverbs that Biber et al. (1999, p. 855) group separately as *evidentials* (see above) are distributed between the conviction and doubt groups by Quirk et al. (1985) and between strong modals and quasi-strong modals by Huddleston and Pullum (2002). I follow Quirk et al. and Huddleston and Pullum in considering evidentials to be a sub-group of epistemic markers, as does some other literature (see 2.2.2). Similarly following Quirk et al., later studies (Simon-Vandenbergen & Aijmer, 2007; Suzuki, 2015, 2018; Van der Auwera et al., 2005) use the categories of *modal adverbs of possibility* (e.g. *possibly, maybe, probably*) and *modal adverbs of certainty* (e.g. *certainly, of course, indeed*) for the items which I include under the cover term *modal adverbs of certainty*, MACs (see 1.2.). It is these adverbs which will be included in my analysis (see 4.3.2).

The discussion in this section is summarised in Table 2.1. This shows the categories used for what I call MACs within the typology of adverbs used by each of the three grammars. The table does not include any of the related subcategories (e.g. Quirk et al's *disjuncts of reality* or Biber et al.'s *reality or actuality* adverbs) made up of adverbs that I would not consider to be MACs. Thus, all the words given in Table 2.1 are among the adverbs that this thesis will analyse. Other non-MAC types of adverbs are passed over in silence henceforth.

Table 2.1 Degree of certainty categorisation of MACs in major English reference grammars

Quirk et al. (1985)	Biber et al. (1999)	Huddleston and Pullum (2002)
Content disjuncts **Degree of truth** ***Group 1 degree of conviction*** *admittedly, assuredly, avowedly, certainly, decidedly, definitely, incontestably, inconvertibly, indeed, indisputably, indisputably, unquestionably, undoubtedly, manifestly, evidently, clearly, obviously, patently, plainly, of course, no doubt, in fact, obviously, really* ***Group 2 degree of doubt*** *allegedly, arguably, conceivably, doubtless, , maybe, perhaps, possibly, presumably, purportedly, reportedly, reputedly, seemingly, supposedly, likely*	**Stance adverbial** **Epistemic stance adverbial** **i. Certainty/doubt** *probably, expectedly, likely, unlikely, perhaps, arguably, decidedly, definitely, incontrovertibly, indeed, no doubt, certainly, of course, undoubtedly, undeniably* **ii. Evidential** *Reportedly, evidently, supposedly*	**Modal adjuncts** **i. Strong modals** *assuredly, certainly, clearly, definitely, incontestably, indubitably, ineluctably, inescapably, manifestly, necessarily, obviously, patently, plainly, surely, truly, unarguably, unavoidably, undeniably, undoubtedly, unquestionably* **ii. Quasi-strong modals** *evidently, apparently, doubtless, presumably, seemingly* **iii. Medium modals** *likely, arguably, probably* **iv. Weak modals** *possibly, perhaps, maybe, conceivably*

2.3.3.3 Degree of certainty categorisation of MACs in recent research

2.3.3.3.1 A scale of epistemic support

Functionally, MACs as epistemic expressions communicate the speaker's degree of certainty. That is, they represent levels of strength of the speaker's commitment to the truth of a proposition. Thus, they allow a speaker to express either conviction or doubt about some state of affairs (Coates, 1983; Palmer, 2001, 2014; Van der Auwera & Plungian, 1998). Boye (2012, p. 2) uses the term *epistemic support* for the function of epistemic modality. Boye (2012, p. 22) develops this notion of *epistemic support scale* from Horn's (2001) view that the notion of epistemic support refers to a continuous *quantitative scale*. Horn (2001, p. 236) says there exists a "notion of correspondence between matched positive and negative scales". Thus, Horn (2001, p. 136) categorises *certainly, likely,* and *possibly* as expressing points on a positive epistemic scale, and places *uncertainly, unlikely,* and *impossibly* on a matched negative epistemic scale.

Following Horn, Boye (2012, p. 21-22) argues that the degree of a speaker's commitment to a proposition, or confidence, can be conceived as "a continuous quantitative scale: an *epistemic modal scale*" that includes *full support, partial support*, and *neutral support* for the proposition. Boye (2012, p. 22) says that *full support* means that a speaker has *high certainty* about the truth value of a proposition. Boye places *partial support* between *full* and *neutral support*. Boye (2012, p. 22) says that *partial* and *neutral support* are combined as "less than full support" in contrast to *full support* (Boye, 2012, p. 22). Alternatively, *full support* and *partial support* may be grouped together as "more than neutral support" in contrast to *neutral support* (Boye, 2012, p. 22). *Neutral support* can be characterised as representing ignorance, "complete lack of knowledge" on the part of a speaker (Boye, 2012, p. 25). However, at some points, Boye (2012, pp. 46, 135-36) does not distinguish *partial* and

neutral support, because certain languages do not explicitly express all these distinctions of epistemic support. For example, in Limbu an adverb of neutral support, *ti-ya* 'perhaps', can be used to express *partial support*, whereas its absence expresses full epistemic support.

Let us consider some examples of adverbs expressing the classes of epistemic support that Boye ultimately arrives at: *high certainty support* (HCS) and *probability support* (PS). Speakers and writers add HCS MACs to express high estimation of the likelihood of the proposition being true (Simon-Vandenbergen & Aijmer, 2007, p. 119), or alternatively to express aspects of meaning such as "emphatic assertion" that also affirm their commitment to the truth of a proposition (Simon-Vandenbergen and Aijmer, 2007, p. 105). For instance, in example (28), the speaker highlights "the qualification conveyed by *simple* by multiple intensification expressed by *really, very* and *indeed*", of which the first and third are MACs (Simon-Vandenbergen & Aijmer, 2007, p. 105).

28. "And what you can see is it's it's merely a kind of flat-backed shed which has been erected uhm the sort of thing that's *really* very simple *indeed* to build (ICE-GB:S2A-024/77)" (Simon-Vandenbergen & Aijmer, 2007, p. 104).

On the other hand, speakers/writers add PS MACs to express doubt, or a low estimate of the likelihood of the proposition being true, thus downtoning their certainty, as in (29).

29. "Rather surprisingly, *perhaps*, the war seems to have been popular" (BNC_E9V).

PS MACs may also be used by speakers to "present the state of affairs as a desirable possibility" (Pic & Furmaniak, 2012, p. 26), as in (30). Though arguably that meaning is only there because the MAC co-occurs with deontic "should".

30. "Then *perhaps* you should join a band of bell ringers, engaged in the grand old practice of ringing the changes." (Pic & Furmaniak, 2012, p.26).

2.3.3.3.2 MACs and negation

In addition to grouping together MACs that express certainty and possibility, Boye (2012, p. 22) categorises sequences of MAC + NEG as being, themselves, MACs (in my terminology, MAC phrases) that express a belief about a clause's negated content, i.e. "negative epistemic modal meanings". Parallel to epistemic support for a proposition is *epistemic support for the negative counterpart* (i.e. the stance that the proposition is certainly/probably/possibly not true).

Throughout this thesis, sequences of MAC + NEG will be treated as single units, such as *certainly not* or *perhaps not*, following Boye (2012, 2016). Boye's (2012) presentation of the scale including the negative epistemic modal meanings is inspired by Nuyts' (2001, pp. 21-22) epistemic scale, and is reproduced here as Figure 2.1.

Certainty that > > Neutral/Agnostic > > Certainty that not
(*certainly*) (*probably*) (*probably not*) (*certainly not*)

Figure 2.1: Boye's (2012, p. 46) presentation of Nuyts' (2001) epistemic scale

Boye's (2012, p. 36)own basic division of the notion of epistemic support is more detailed, as depicted in Figure 2.2.

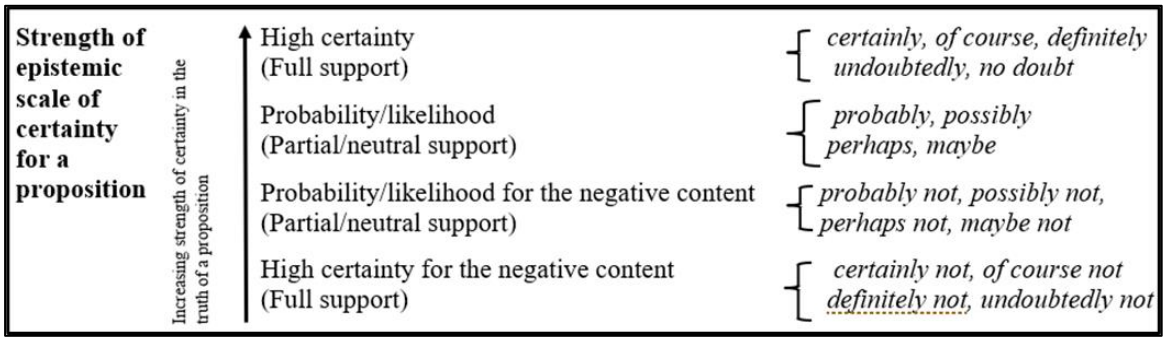


Figure 2.2: Boye's scale of epistemic support (adapted from Boye, 2012, p. 36)

Boye's (2012, p. 36) notion of epistemic scale division is that the epistemic modality scale has polarity, that is, it goes from certainty, through neutral possibility (on the positive side), to the negative side, where it goes "over neutral epistemic support to high epistemic support for the negative counterpart of a proposition" (Boye, 2016, p. 117) and thus populate the right-hand end of the continuum in Figure 2.1 (equivalently, the lower half in Figure 2.2).

On the scale in Figure 2.2, Boye (2012, p. 22) distinguishes four main classes of degree of epistemic support: *high certainty support* (HCS), *probability support* (PS), *probability support for negative content* (PSNC), and *high certainty support for negative content* (HCSNC). These are the categories that I utilise in my analyses; see Chapter 4.

It could be questioned why epistemic support for negative content needs to be expressed by combinations of MAC + NEG, and not derived MACs with explicit negating prefixes. Simon-Vandenbergen and Aijmer (2007) observe that certain HCS MACs do have the negative prefix *un-* (e.g. *undoubtedly*), but these do not form negative/positive pairs, except in the case of *arguably/unarguably*. For instance, the negative form of *certainly* is *uncertainly*. However, *uncertainly* is an adverb of manner, used to modify a verb (e.g. *He spoke uncertainly*) and not to convey lack of epistemic support. Simon-Vandenbergen and Aijmer thus say, with respect to *arguably/unarguably*, that the former means that there is a reason to support the proposition being true, whereas the latter means that there is no reason to refute it. (Simon-Vandenbergen & Aijmer, 2007, p. 61). The two thus express opposite modality. On the other hand, one cannot respond to an argument that something is certainly the case by asserting to the contrary that it is uncertainly the case. Simon-Vandenbergen and Aijmer also observe that negative-marked MACs, whether morphologically derived (e.g. *undoubtedly*, *undeniably*) or MAC phrases involving a negator or negative article (e.g. *no doubt*), all express a degree of certainty. Both Nuyts (2001; 2016) and Boye (2012; 2016)

note that when the sequences of HCS MAC + NEG and PS MAC + NEG are used, they convey epistemic support for the negative content of a proposition and function as the negative forms of the unnegated MACs (i.e., HCS and PS MACs). I follow Boye (2012) in categorising negative MAC phrases [MAC + NEG] as expressing epistemic support for the negated content.

2.3.3.3.3 Notions of contraries and contradictories

Boye (2012) distinguishes two types of negation within the scope of MACs: *contraries* and *contradictories*. *Contraries* of positive or neutral epistemic support are utterances with meanings such as "*I am certain that* [the proposition is not true]" and "*It is probable that* [the proposition is not true]" (Boye, 2012, p.28). Such utterances "can be characterised as *contraries* of positive or neutral epistemic modal meanings". The meanings of the relevant HCSNC and PSNC MACs gloss as "certainly not", "possibly not" and "probably not" (Boye, 2012, p.18). Boye (2012) also places *counterfactuals* (statements that express what has not happened or is not the case, also called *counterfactives* or *contrary-to-the-fact*) in the category of contraries. Unlike other contraries, counterfactuals do not assert a negative proposition explicitly, as examples (31-34) illustrate. In examples 32, 33, and 34 the clause "I own[ed] the house" is counterfactual.

31. It is *certainly not* my house. [Explicit certainty about negative content]

32. I pretend I own a house. [Counterfactual: implicit certainty]

33. I wish I owned a house.

34. If I had owned a house, I would be happy.

(adapted from Boye, 2012, pp.28-29)

Boye (2012, p. 28) says that, as *counterfactuals* do not directly express full support for the negation of a proposition, most scholars do not include them "under epistemic modality".

Boye (2012, p. 28) says that *contradictories* of positive or neutral epistemic support are those utterances with meanings of *I am not certain that P* and *it is not probable that P*, that is, unlike the contraries, in contradictories the negative meaning attaches to the epistemic support and not to the content of the proposition. Therefore, the meaning can be glossed as *uncertain* or *unlikely* (Boye, 2012, p. 29). For instance, Boye (2012, p. 29) says that if "uncertain" is used in a sentence (e.g. "I am *uncertain* that I own a house"), *uncertain* supports the notion of negative epistemic support. If "unlikely" is used (e.g. "it is *unlikely* I own a house") it can be described as a negative partial epistemic support (Boye, 2012, p. 29). However, all contradictories "have strength-equivalents that are not themselves contradictories" (Boye, 2012, p. 29). Thus, for instance, the meaning of *I am a little uncertain* is "approximately the same as" that of "*I am a little in doubt* or *it is unlikely* (Boye, 2012, p. 29).

2.4 Pragmatic functions of MACs

This section explores the functions of MACs as pragmatic markers. *Pragmatic markers* are lexical items which have come to express purely pragmatic functions, either as well as or rather than their original referential content, through the processes of grammaticalisation and pragmaticalisation. Hopper and Traugott (2003, p. 18) define *grammaticalisation* as the process through which lexical items and constructions come to serve grammatical functions. *Pragmaticalisation* is defined by Badan (2020, p. 314) as the process through which a lexical item becomes a pragmatic marker, that is, a lexical item that

expresses "pragmatic, interpersonal meaning". The *interpersonal*, according to Halliday's (1994, p.68) Systemic Functional Grammar (SFG), is a meta-function of language through which language users establish, negotiate, and position their stance as a part of communication.

Simon-Vandenbergen and Aijmer (2007) say that MACs have developed pragmatic functions such as *solidarity*, *politeness,* and *expectation*. MACs also function as *rhetorical* devices (Schwenter & Traugott, 2000; Simon-Vandenbergen & Aijmer, 2007). Rhetorical devices are linguistic tool that speakers/writers use to engage their audience or readers (Leach, 2000, p. 208). A speaker or a writer constructs an argument or develops an existing argument (using rhetorical device(s)) here persuade us to accept their viewpoint (Leach, 2000, p. 208). Simon-Vandenbergen and Aijmer (2007, p.42) say that MACs are "linguistic resources for [rhetorical] strategizing" that speakers use to accept, challenge, confront, manipulate, and persuade.

In 2.4.1, I discuss the notions of *indexicality* and *reflexivity*, which determine the pragmatic context. Then in 2.4.2, I discuss indexical stances (attitudes other than epistemic stance) whose expression may be a rhetorical-pragmatic function performed by MACs.

2.4.1 Pragmatic interpretation

Pragmatic interpretation of an utterance differs from semantic interpretation (Rühlemann & Aijmer, 2015). Pragmatic interpretation depends on certain aspects of the context that determine the meaning. These contexts include *indexicality*, *self-reflexivity*, and *socio-cognitive communication* . Understanding these contexts helps in understanding how lexical items function as pragmatic markers (Aijmer, 2015, p. 198; Rühlemann, 2019, p.85).

Aijmer defines *indexicality* as the use of certain words or phrases that signal a speaker's perspective or position in a particular discourse or conversation. These words or phrases, known as discourse markers, are used to signal shifts in topic, the speaker's attitude, or degrees of certainty. These discourse makers can thus provide clues about the social or interactional context in which the conversation is taking place. Thus, *indexical* indicators are those that "are used to index speakers' social or professional identity, social relations and activities" (Aijmer, 2015, p. 198). For instance, a speaker or writer's linguistic choices may differ in conversation as opposed to a research paper because they change their discourse identity with the changed communication domain (Flowerdew & Scollon, 1997, p.2). Simon-Vandenbergen and Aijmer (2007, p.44) say that MACs are indexical because they convey information about speakers/writers, implicit in context (e.g. information about authority and social identity). Therefore, MACs are "indexically related to variables in a social situation and are associated with types of social activity, with social roles and with power" (Simon-Vandenbergen & Aijmer, 2007, p. 5). Other indexical features of MACs include discourse functions and stance towards the information being expressed.

Schiffrin claims that writers or speakers are idea-oriented; therefore they evaluate ideas, present them neutrally, express commitment to them, or show their distance from them. Speakers' stance is dependent on how they act: "they may perform an action indirectly and thus deny responsibility for its consequences" (Schiffrin, 1987, p.27). Schiffrin (1987, p.31) defines discourse markers as "sequentially dependent elements which bracket units of talk". Schiffrin (1987, pp. 24-25) explains that in "exchange structures" (i.e. turn-taking, adjacency pairs) speakers may use MACs in order to make a claim to their turn or relinquish it. These turn-takings may signal feedback on the hearer's/reader's position, on some controversial issue, or on societal norms and culture. Simon-Vandenbergen and Aijmer (2007) say that such exchange structures also employ certain MACs (e.g. *of course, certainly*) as short

responses to questions. MACs used in short responses may be used to grant or refuse permission, to show solidarity (see 2.4.2.3), to conform to another speaker's stance, or as a politeness marker (see 2.4.2.8). To interpret such a MAC, then, we need to analyse the exchange structure in which it occurs, to determine whether any given instance is in fact performing that function.

Self-reflexivity is a "speaker-centered function" (Rühlemann, 2019, p. 85), the phenomenon whereby speakers/writers have a "metapragmatic awareness" of what type of interaction they are involved in (Aijmer, 2013, p. 4). Metapragmatic awareness means that the speaker/writer is aware of how to organize their communication and the hearer/reader is aware of how to "draw inference about the overall structure of the conversation" (Simon-Vandenbergen & Aijmer, 2007, p. 49; Verschueren, 2000, p. 444). Reflexivity refers to the potential of language to be used to reflect upon itself, that is, it is a metapragmatic function by which speakers/writers self-monitor their communication. For instance, the hearer can make inferences regarding whether the conversation is an argument, an exchange of information, or some speech act (Simon-Vandenbergen & Aijmer, 2007, p. 49). While it is true that we cannot access the internal processing of a speaker's/writer's mind, "pragmatic markers (and other devices) can emerge as overt indicators of ongoing metalinguistic activity in the speaker's mind" (Aijmer, 2013, p.4). Aijmer (2013, p. 109) illustrates self-reflexivity with an example where a speaker/writer uses *actually* to indicate apologetic tone in order to avoid face-threatening remark about an unexpected event as opposed to the interlocutor's assumption. The speaker's apologetic tone is additionally signalled by the use of *well* and *probably* in example (35). Use of *actually* overtly indicates politeness because E starts by agreeing with A, but then shifts to a contradictory stance – but one that is mitigated both by *actually* and *probably* (Simon-Vandenbergen & Aijmer, 2007, p. 109).

35. "A: The cat will attack anyone won't

E: Oh yes Well actually she'll run for it probably (S1A-019 162-166 FACE)" (Aijmer, 2013, p.109).

Socio-cognitive communication is a process of combining both social and cognitive factors in exchanging information and meaning between interlocuters via language and other forms of symbolic representation. From the pragmatic perspective, communication is influenced by social and cognitive contexts, which are manifested in the speaker's goals and the listener's expectations. Both social and cognitive factors are shaped by cultural norms and societal values; these factors in turn influence interpretation and understanding of the message. Nuyts (2001, p.4) introduces the notion of socio-cognitive communication in order to understand the functional factors that determine the selection of epistemic expressions, and to examine what cognitive processes are involved in linking epistemic notions to social-communicative skills. This idea is used by both Simon-Vandenbergen and Aijmer (2007) and Boye (2012). Simon-Vandenbergen and Aijmer (2007, p. 50) analyse how adverbs of certainty are used as social markers and to signal or guide organisation of meaningful messages in an evolving discourse. Simon-Vandenbergen and Aijmer (2007, p. 55) consider the link between the epistemic stance of a speaker and *indexicality* and *social cognition* to be associated with cultural and social factors. Similarly, Boye (2012, p. 7) says that language as a "functional-communication phenomenon" is "an integral part of human cognition" which is shaped by social contexts. Boye says that the ability to understand social-contextual meaning enables interlocutors to effectively communicate with each other by evaluating the meaning in the message.

2.4.2 Indexical stance and the rhetorical-pragmatic functions of MACs

Simon-Vandenbergen and Aijmer (2007) say that indexical stance is an expression of attitudes which may be conveyed by MACs. These attitudes include several pragmatic functions, such as *authority*, *concession*, *expectation* and *counter-expectation*, *hedging*, *politeness*, and *solidarity* (in addition to epistemic stance). MACs express these attitudes differently in different contexts. Another aspect of contextual meaning on which the function of a MAC may be dependent is "who is speaking/writing to whom for what purposes" (Simon-Vandenbergen & Aijmer, 2007, p.311). In this section, I discuss in turn each of the indexical stances that may be expressed by MACs functioning as pragmatic markers in context. Where Genady (2005) comments on the same functions in Urdu, that will be noted as well for completeness, prior to the more general survey to come in 2.5. I also primarily cite Simon-Vandenbergen and Aijmer (2007) in discussing these characteristics whose main work is on HCS MACs but the other research (e.g. Pic & Furmaniak, 2012) shows that most of PS MACs also exhibit these characteristics.

2.4.2.1 Authority

A speaker/writer may claim superior knowledge on some topic by using an HCS MAC, thus performing the power-oriented function of expressing *authority* (Simon-Vandenbergen & Aijmer, 2007). To express *authority* on a subject matter, speakers/writers express a high degree of confidence in the truth value of their claims on that subject matter; whatever inferences they may make are based, it is implied, on their existing knowledge of that area, and thus are to be assessed as highly probable. By expressing a claim to authority in some domain, a speaker/writer defines their own positioning as well as their hearer's/reader's. In this process of establishing their positioning, the speaker may align with

or diverge from an alternative position or voice. In sum, use of HCS MACs in this way by speakers/writers is intended to present "their judgments, opinions, and commitments" so as "to stamp their personal authority onto their arguments and step back and disguise their involvement" (Hyland, 2005, p. 176).

In some instances, MACs such as *of course* are used to endorse an *external* authority. In this case, the MAC in question may not be employed to explicitly "introduce external voice" but rather to "express a confirmation of an earlier statement" (Simon-Vandenbergen & Aijmer, 2007, p.307) attributed to some authority other than the speaker/writer. Another common practice when a speaker/writer asserts authority over their hearer/reader is to use *I think* before presenting their point of view (Aijmer & Rühlemann, 2015, p. 212). In Urdu specifically, a speaker/writer may overtly claim authority via phrases such as *maiṁ jāntā hūṁ ke* 'I know that' and *merā khayal hai ke* 'my opinion is that'. By adding these phrases, a speaker/writer asserts that their proposition is based on some evidence, or else is an inference based on their knowledge of the subject area (Genady, 2005, p. 110). Words for information (e.g. *mālumāt* 'knowledge/information') are especially used by Urdu speakers when claiming to possess "knowledge about, or acquaintance with, some specific state of affairs" (Genady, 2005, p.124).

In addition to the epistemic authority discussed above, MACs sometimes occur in contexts that shift the meaning from epistemic authority, expressing certainty about the speaker's/writer's knowledge, to deontic authority , expressing a speaker/writer having social power to issue a command to or to impose an obligation on the hearer/reader. Use of MACs to express deontic authority is explicit when they cooccur with certain MVs frequently used with deontic function (e.g. *should*). In English, in fact, speakers often use modal auxiliaries, such as *should* and *ought to*, to express the "speaker's authority" (Quirk et al., 1985, p. 227) to impose obligations on others. Use of MACs with deontic MVs implies that the speaker has

confidence that their recommendation/command/demand/suggestion will be carried out by the hearer/reader (or, sometimes, by the third person referent of the clause subject). Therefore, speakers/writers may add an HCS MAC to more strongly express confidence in the validity or utility of their recommendation.

Similarly, in an Urdu clause, speaker uses a MAC with modal auxiliary *cāhīē* 'should' to give suggestions or recommendations based on their "understanding of what is good and reasonable" (Genady, 2005, p. 111). Genady (2005, p. 111) says that speakers use such combinations (e.g. MAC + modal auxiliary) specific contexts as a "rational authority rather than a social authority" the reading will be deontic and not epistemic.

2.4.2.2 Emphasis

One of the most common pragmatic function of MACs is as *emphasisers* (Quirk et al., 1985, pp.583-84). This function is termed *intensification* by Hoye (1997). When MACs are used with this function in a clause, they do not alter the meaning of the clause's proposition; rather MACs as emphasisers "reinforc[e] effect on the truth value of a clause or part of the clause to which they apply" (Quirk et al., 1985, p.583). That is, the speaker/writer intentionally places a MAC next to the clausal element that they want to put emphasis on. Although MACs in general emphasise the speaker's degree of certainty regarding a proposition, in these cases the emphasis is more focused on some particular element(s) in the clause. In that case, MACs pragmatically function as emphasisers. In example (36), the speaker's use of *definitely* conveys certainty in their interpretation of the woman's apparent emotional state specifically.

36. "Behind the policemen was a middle-aged woman who looked *definitely* flustered. (ESPC:DF1)" (Simon-Vandenbergen & Aijmer, 2007, p. 102).

In some instances, HCS MACs perform the function of emphasiser when they cooccur with "the determiner *such* and the adverb *so*" used with "no accompanying correlative clause or phrase"; in this case the MAC is used by the speaker to reinforce their argument for "clarity and emphasis" (Quirk et al., 1985, p.1416). HCS MACs that cooccur with restrictive adverbials (e.g. *only*) similarly reinforce the emphasis on the proposition being true "in a way which expressly excludes some other possibilities" (Biber et al., 1999, p. 780). In example (37), the placement of a PS MAC right before *a feeling of shame* downtones the speaker's inference about the feelings of the person under discussion, but at the same time highlights their viewpoint, i.e. provides emphasis.

37. "Modi's reserve might have sprung from a desire to shield Jeanne from gossip, and *perhaps* a feeling of shame in front of a married man , at his own role in the affair" (BNC_ANF 1778).

2.4.2.3 Solidarity

Certain HCS MACs (e.g. *of course, indeed, obviously*) have an interactional function as solidarity markers, that is, they are used by the speaker/writer to signal to the hearer/reader that they have equality in knowledge. Simon-Vandenbergen and Aijmer's (2007) analysis shows that HCS MACs function as solidarity markers in two situations: when the speaker/writer assumes that the hearer/reader is uninformed and tries to redress the power balance, or when the speaker/writer believes that they and the hearer/reader have a shared world experience and wants to express a sense of belonging.

When the speaker/writer assumes that the hearer/reader is uninformed, they may add an HCS MAC to their assertion to indicate (insincerely) that they are just repeating what the hearer/reader already knows. By implying that *you know as well as I do*, the speaker/writer

imputes equality (of knowledgeability) between themself and the hearer/reader. For instance, *of course* functions as a solidarity marker in example (38) and thus protects the face (see 2.4.2.8) of the hearer/reader. This placement of *of course* implies that what the speaker/writer is telling the hearer/reader is not new information, and therefore disclaims any implicit claim to superior knowledge. This way, the speaker/writer both balances the power and marks solidarity with their hearer/reader.

38. "And it's telling us about a campaign in Georgia now *of course* the Soviet Union" (ICE- GB:S2A-059/23)" (Simon-Vandenbergen & Aijmer, 2007, p. 205).

Sometimes use of HCS MACs confirms a speaker's/writer's solidarity with the like-minded hearer/reader. That is, the speaker/writer uses an HCS MAC to express solidarity with the hearer/reader on the basis of an assumed shared world. When a speaker/writer assumes shared knowledge to signal solidarity, "it signals the speaker's awareness of a common background which includes knowledge of the facts or at least common expectations [see 2.4.2.3] about the state of affairs" (Simon-Vandenbergen & Aijmer, 2007, p. 312). For instance, with regards to the example given as (39), Simon-Vandenbergen & Aijmer say that the writer's use of *of course* with *BBC* refers to assumed shared preference (for BBC as opposed to other radio) by including the reader in the life-world of the writer.

39. "You're lying down, probably beneath a line of dripping wet clothing, the radio on (*BBC of course*), and underneath and around you lies … God only knows what ! (ICE-GB:W1B- 001/13)" (Simon-Vandenbergen & Aijmer, 2007, p. 210).

2.4.2.4 Expectation

Some HCS MACs function to signal "expectations of some kind, against which knowledge may be matched" (Chafe, (1986, p.270). Simon-Vandenbergen and Aijmer (2007,

p. 28) define these *expectation markers* as a means by which "the speaker evaluates the actuality of the situation not with regard to the source of information but with regard to whether it is expected or appropriate". HCS and HCSNC MACs which function as expectation markers convey certainty together with the meaning of "according to/in conformity with expectations" (Simon-Vandenbergen & Aijmer, 2007, p.172). That is, a speaker/writer expresses their commitment to the truth of the proposition on the basis that that the state of affairs given in their proposition is to be expected, as in (40).

40. But *of course* now they're going to send everyone aren't they regardless of whether you pay tax or not (ICE-GB:S 1A-007/272)" (Simon-Vandenbergen & Aijmer, 2007, p. 29).

2.4.2.5 Concession

Concession is a function employed by a speaker/writer who is engaged in persuasion or argumentation for some point of view. A concession is used in an utterance that contains (at least) two propositions, to establish that the first part of the utterance is not the main point which speaker/writer wants to convey and may therefore be taken for granted in the context of the argument at hand. Use of a MAC as a concessive marker is a strategy by a speaker/writer to foreground with emphasis the *second* part of the utterance, which expresses the speaker's "hypothetical alternative point of view" to build "into argumentation" or to reply to "critical remarks from interactants" (Simon-Vandenbergen & Aijmer, 2007, p.209). Hence, by conceding the earlier proposition, the speaker/writer backgrounds the alternative (real or hypothetical) argument(s) it expresses, allowing them to be considered assumed knowledge – as in (41), where the writer uses *of course* to concede, and then builds their argumentation after *but even so*.

41. "*Of course* republicanism isn't about the Royal Family failing as a tourist attraction *but even* so it is interesting to discover that it isn't a very great attraction." (ICE-GB:S2B-032/39)" (Simon-Vandenbergen & Aijmer, 2007, p. 178).

Sometimes, HCS and HCSNC MACs in concessive contexts function to contrast shared knowledge with some new (and important) information (Simon-Vandenbergen & Aijmer, 2007, p.183). In (42) the speaker first indicates a shared knowledge which then is contrasted with the new information given after *but*.

42. "then *of course* we can condemn them *but* we cannot take any particular legal action (TRIPTIC:DCER:06:01)" (Simon-Vandenbergen & Aijmer, 2007, p.183).

Pic and Furmaniak (2012, p. 26) observe that PS and PSNC MACs may also express concession when a speaker/writer admits to the possibility of a proposition being true, but considers it to be irrelevant to the state of affairs at hand, a point which will be mentioned explicitly in the following clause, as in (43).

43. "Goffman suggests a fairly rigid division between front and back regions, which was *perhaps* a feature of twentieth-century American homes *but* is less applicable to Georgian London" (Pic & Furmaniak, 2012, p. 26).

2.4.2.6 Hedges

Hedges are elements used by speakers/writers to express tentativeness in their assertion (Coates, 2015, p.88). Hedges express unassertiveness or are used to mitigate the force of an utterance to avoid unpleasantness (see 2.4.2.8). Coates says that hedges are mainly used in discussion of sensitive topics; in such contexts, they are an invaluable device for speakers/writers because they can mitigate the force of what is being said, but also as a

discourse strategy to "protect both speaker's and hearer's face" (Coates, 2015, p.90) (the notion of face is introduced in detail in 2.4.2.8). Various modal expressions, including MACs, are used as hedges with the purpose of conveying a tentative or vague proposition; as Salager-Meyer (1994, p.150) points out, such hedges are especially notable for the range of pragmatic functions they fulfil in the specific genre of medical (and more broadly, scientific) research writing. Speakers may utilise PS and PSNC MACs such as *possibly* to "mitigate criticism" (Diani, 2015, p.179).

PS MACs (e.g. *possibly*) are among the modal expression that may be used as hedges as a politeness device (see 2.4.2.8) (Holmes, 2013, p.64). PS MACs functioning as hedges also function as *adversatives* when they combine with adversative adverbs (e.g. *however*) so as to present an argument in opposition or contrast (see 2.4.2.7) (Aijmer, 2013, p. 82). However, in adversative dependent clauses MACs as hedges are used by speakers/writers to elaborate or to emphasise on their point of view (Aijmer, 2013, p.82), as in (44).

44. "If it could be shown that over the years there had been a major redistribution of wealth from the rich to the poor, this would indicate a reduction in class inequalities. *However*, wealth is *perhaps* even more difficult to measure than income and reliable data prove elusive" (BNC_FB6 1394).

2.4.2.7 Counter-expectation

Counter-expectation, a function also called "countering" (Simon-Vandenbergen & Aijmer, 2007, p.43), refers to the situation in which a speaker/writer "expresses beliefs or point of views contrary to his or her own or the interlocutor's expectations regarding the states of affairs under discussion" (Traugott, 1999, p.178). For instance, in the English conditional construction, the hypothetical use of the past tense within the conditional clause

"expresses what is contrary to the belief or expectation of the speaker" (Quirk et al., 1985, p.188). In Urdu, likewise, unfulfilled conditional sentences, also known as "contrary to fact" (Schmidt, 1999, p.101) or "contrafactive" (Genady, 2005, p. 23) express those conditions that lack any possibility of being fulfilled, because they express unrealisable action or are purely hypothetical.

A proposition that is a counter-expectation can be equally surprising for the speaker/writer and hearer/reader, because it runs contrary to what they know or expect. Therefore, these counters more often dis-align than align with the hearer's/reader's beliefs and expectations (Martin & White, 2005, p.121). Conversely, an expected proposition is based on speaker's/writer's and hearer's/reader's shared knowledge, and predictable according to their expectations (which results in solidarity; see 2.4.2.6, and for the link to politeness see 2.4.2.8). A proposition that asserts a counter-expectation strategically exploits this shared knowledge (Simon-Vandenbergen & Aijmer, 2007, p.322). In English, when MACs occur after conjunctions of contrast (*although*, *but*, *however*, or *even though*) then the MAC emphasises the counter-expectancy.

2.4.2.8 Politeness

Politeness is one of the most explored concepts in pragmatics (Culpeper, 2011, p. 394). It is concerned with those constraints that we follow to establish, maintain or violate interpersonal relationships. Politeness devices, MACs among them, help speakers/writers in maintenance of *face*. Face is each individual's understanding of their own public image: their prestige, their reputation, and their self-esteem (Brown & Levinson, 1987, p. 61). Brown and Levinson draw their notion of *face* from Goffman (1967, p. 5) who defines it as "the positive

social value a person effectively claims for himself by the line others assume he has taken during a particular contact".

Brown and Levinson (1987, p. 61) say that "face can be lost, maintained or enhanced and must be constantly attended to in interaction". Face is *lost* when a person is humiliated or embarrassed. Face is *maintained* when a person defends their own face. But speakers/writers are also expected to defend others' face if their hearer/readers feel threatened. Face maintenance, in any interaction, is based on mutual cooperation, because it is in the interest of every participant to protect the face of everyone involved (Brown & Levinson, 1987, p. 61). Two types of face are generally discussed: *positive face*, which is "the need to be liked and admired" (Coates, 2015, p. 105); and *negative face*, which is the need "to be free from impositions" (Sifianou, 2020, p. 43). Politeness is a set of communication devices used by speakers/writers as a face-redressive strategy, i.e. to counterbalance a face threat (Brown & Levinson, 1987, p. 27) by avoiding or mitigating speech acts which might otherwise create a threat to the face, positive or negative, of one or more discourse participants (see Sifianou, 2010, p. 48).

Both HCS and PS MACs are employed for politeness, whether *positive* or *negative* politeness. Positive politeness is when the speaker/writer redresses the hearer's/reader's positive face needs by treating the hearer/reader as a socially equal member of some relevant group (Culpeper, 2011, p. 403) (see 2.4.2.3). On the other hand, negative politeness is intended to redress threats to the negative face of the hearer/reader. For instance, a speaker/writer may signal that they do not want to interfere with a hearer's/reader's freedom of action by means of mitigating an imposition with hedges (see 2.4.2.6)

Positive politeness is conveyed by an HCS or HCSNC MAC when used by a speaker/writer to play down their superior knowledge, a strategy similar to the *solidarity*

function discussed earlier (see 2.4.2.3). Simon-Vandenbergen and Aijmer (2007, p. 301) say that an HCS or HCSNC MAC (e.g. *certainly*) used in place of *yes* in response to a question, as in (45), also functions as a discourse marker of politeness, because it logically implies that the speaker/writer does not think a simple *yes* would be an emphatic enough response.

45. "A: Can I ask you a question?

B: *Certainly*." (Simon-Vandenbergen & Aijmer, 2007, p.301).

When an HCS or HCSNC MAC is added as a politeness marker, its "face-attending function has taken over from the epistemic function" (Simon-Vandenbergen & Aijmer, 2007, p.302). Use of an HCS MAC in this way is evidence that the speaker/writer "respects the hearer's/reader's positive face" (Simon-Vandenbergen & Aijmer, 2007, p. 176), as in (46):

46. "*Of course* you are quite right (ICE-GB:W1B-003/168)"
(Simon-Vandenbergen & Aijmer, 2007, p. 176).

An HCS or HCSNC MAC is used as a negative politeness device by the speaker/writer to ward off any fear of face loss from hearer/readers "who are familiar with the information" being conveyed (Holmes, 1988, p. 57). This strategy is "associated with mitigation of face-threatening acts" (Sifianou, 2010, p. 48).

Regarding PS and PSNC MACs, on the other hand, Diani (2015, p. 185) says that certain MACs in these categories such as *possibly* and *probably* are used as politeness markers to soften or downtone a face threat or a disagreement (see 2.3.3.3). Diani reports this type of MAC "to be the most common realisation of politeness strategies" in her corpus.

2.5 MACs in Urdu

Platts (1874, p.189) was the first grammarian to label some Urdu adverbs (e.g. *śāyad* 'perhaps') as "adverbs of *probability* and *doubt*". Since then, however, MACs as a category have been discussed only marginally in Urdu grammars, including the widely-cited grammar of Schmidt (1999). There is only one study, by Genady (2005), that focuses on Hindi-Urdu modality, and in the course of so doing briefly mentions adverbs that express either *certainty* (e.g. *yaqīnān* 'certainly') or *possibility* (e.g. *śāyad* 'perhaps'). Genady (2005) is also the only work to date that discusses the use of MACs in Hindi-Urdu to convey such meanings as emphasis and counterfactuality. Though there has not been any study that specifically utilises Horn's (2001) or Boye's (2012) epistemic modal scale, Genady's discussion of MACs, albeit short, does show that Boye's framework can be applied to Urdu.

Genady (2005) also addresses what I term MAC phrases, explaining the use of *hō saktā hai* 'probably/lit. be can' as a complete adverbial clause which has become fixed and lexicalised when it acts as a modifier of a clause or a clause element. Specifically, Genady (2005, p. 181) points out that the "…verb *hō* (to be) in the existential function combined with *sak-* (*hō saktā hai*)" over time has become "an idiomatic expression [which] successfully substitutes the modal word *śāyad* (perhaps)…". That is, though *saknā* 'can' is an MV, within this adverbial clause, it forms a *possibility marker* MAC phrase. This is in contrast to *saknā* 'can' as the auxiliary of some lexical verb, where it retains its function as an MV (usually expressing dynamic modality). Genady (2005) also discusses the modal meanings of other aspects of the grammar system, such as indicative versus subjunctive mood. Genady's (2005) discussion mentions modal expressions other than MACs that convey degree of certainty, such *sabhav/mūmkīn hai* 'it is possible'. But, because his goal is to address all modal

expressions in Hindi-Urdu, Genady does not discuss modal adverbs (or idiomatic phrases that function as modal adverbs) in detail.

Linguists who have treated Urdu (or Hindi) modality in general (e.g. Bhatt et al., 2011) focus on *saknā* 'can' due to its status as one of only two modal auxiliaries in Urdu. They thus look at modal constructions involving this verb, such of its use governing a lexical verb root as in (47), while paying much less attention to non-verbal fixed expressions that convey modality, as Genady (2005, p.99) points out. *Sak* 'can' is used in the senses of both *ability* and *possibility* (Schmidt, 1999, p. 115).

47. Yasin voh kar sakā
Yasin DEM do can.PFV.M.SG

"Yasin could do that" (Bhatt et al., 2011, p. 2).

Genady (2005, p.3) acknowledges that his analysis of Hindi-Urdu modal expressions relies on Palmer's (2001) perspective on modality. Palmer's (2001) work is not corpus-based; neither, in the strict sense, is Genady's. Genady (2005, p.19) reports using a corpus of some literary works in addition to other sources, and most of his examples are drawn from this corpus. However, his investigation does not include the quantitative analysis that is an important component of corpus-based analysis. A very few other studies exist that briefly discuss a subset of the pragmatic functions of one or more Urdu MACs. For instance, Hassan and Said (2020) and Shahid et al. (2020) both report that the Urdu PS MACs *śāyad* 'perhaps' and *ġālibān* 'possibly' are used as hedging markers, but without investigating any other functions. However, the research to date on Urdu has yet to come to any systematic analysis encompassing the syntactic, semantic and pragmatic behaviour of MACs. Since there is so little published literature on Urdu MACs, I will utilise a theoretical framework based on the much more extensive literature on English MACs that I have reviewed earlier in this chapter.

2.6 A theoretical framework for analysing MACs

The literature establishes that MACs can be placed on a continuum of meaning from *high certainty* (HCS and HCSNC) to *probability* (PS and PSNC) (Boye, 2012; see 2.3.3.3). Boye (2012) considers MACs to be a typological category. *Typology* is defined by Gast (2015, p. 155) as the linguistic field which seeks to classify languages according to structural features that can be determined by comparing cross-linguistic data. However, Boye (2012, pp. 10-11) also says that these cross-linguistic (typological) categories have descriptive significance only. That means, according to Boye, that the set of MACs can be treated as equivalent, comparative categories in two or more languages. But the equivalence of descriptive categories does not warrant treating them as "exact conceptual or functional-communicative mirror-images" across languages (Boye, 2012, p. 11). Boye (2012, p. 9) says that the purpose of defining a crosslinguistic descriptive category is to generate crosslinguistic generalisations regarding that category. That is, such descriptive categories are "simply generalisation over coherent sets of linguistic phenomena" (Boye, 2012, p. 11).

The present study is not a crosslinguistic typological analysis as defined by Boye. It is, rather, a contrastive study of two particular languages. However, Gast (2012, p.1) states that, broadly defined, contrastive linguistics encompasses this special case of the comparative study of any pair of languages. As opposed to crosslinguistic analyses of a large set of languages, a contrastive typological analysis uses data from only two languages to investigate differences in meaning and/or usage of some linguistic feature in those two languages "against the background of similarities" (Gast, 2013, p. 154).

As the main focus of a contrastive analysis is descriptive (in the sense of Leech, 2015, p. 146), it therefore "provides an interface between theory and application" (Gast, 2013, p.

154). Gast (2013) claims that comparison of languages is sometimes driven by the objective of applicability in real life, as in translation studies and foreign language teaching, especially when the languages in question are in extensive use in a bilingual/multilingual community. English and Urdu are genetically related, albeit belonging to distant branches of the larger Indo-European language family (Kachru, 2006, pp. 2-3). As such, their comparison is of minor relevance to linguistic typology at large. But there exists a large bilingual community of Urdu and English speakers within Pakistan, India and the diaspora around the world, and so contrastive analysis of a specific linguistic feature in English and Urdu is justified by Gast's rationale. My reason for using English as the reference point in my contrastive analysis is that English has an extensive literature on MACs that can form a foundation upon which to build a cross-linguistically informed description of Urdu MACs (and, it is to be hoped, provide a framework for future corpus-based contrastive research on Urdu). Not least, a contrastive analysis of this specific descriptive category (MACs) in English and Urdu will help in mapping semantic and pragmatic similarities and differences in how the cognitive domain of certainty and possibility are represented by these two languages.

On that basis, bringing together the various threads of the literature reviewed in this chapter generates the theoretical framework that I will use in analysing my data. This framework is based on the formal and functional parameters, summarised in Table 2.2, that have been reported in the literature as characterising English MACs. These parameters constitute a roadmap that I can use to identify similarities and differences between English MACs and corresponding Urdu MACs. This will be accomplished using contrastive corpus-based descriptive analysis, the literature on which will be reviewed in sections 3.3 and 3.4. The parameters outlined in Table 2.2 are mainly inspired by the framework of Simon-Vandenbergen and Aijmer (2007). Simon-Vandenbergen and Aijmer (2007) study focuses on modality (and other allied meanings e.g. *evidentiality*) conveyed by HCS and HCSNC

MACs. But other existing research (e.g. Pic & Furmaniak, 2012) shows that Simon-Vandenbergen and Aijmer's (2007) parameters are compatible for the analysis of PS and PSNC MACs and can be applied to analyse modal meanings and pragmatic functions of MACs.

Within this framework, my research on MACs will begin by investigating their syntactic positioning, via an approach following Boye (2012, 2016). Then, I extend that evaluation to a comparison of semantic and pragmatic functions of MACs across the two languages, mainly following Simon-Vandenbergen and Aijmer's (2007) description of (HCS) MACs as epistemic certainty adverbs. Therefore, the general structure of my framework is adopted from Simon-Vandenbergen and Aijmer(2007). The modifications applied to this general structure, based not only on Boye's (2012; 2016) but also on Biber et al.'s (1999) and Quirk et al.'s (1985) research on the syntactic placement of modal adverbs and on their scope over other elements, are my own. Also, incorporation of Pic and Furmaniak's (2012) research on semantic and pragmatic functions of PS MACs, are my own. The structure of these parameters for analysis and their interrelations are given in Table 2.2.

Table 2.2: Parameters for a feature analysis of MACs

PARAMETERS		
SYNTACTIC FEATURES		
Realisation at clause level	Initial Medial Final	Scope over clause/clause elements
Types of clauses	Independent clauses	Declarative Interrogative
	Dependent clauses	Adverbial clause Non- restrictive relative clause Complement clause: *That* clause *Wh*-clause Finite nominal clause Non-finite nominal clause
SEMANTIC FEATURES		
Modal status (MACs in interaction with other elements)	Epistemic	Certainty Probability
	Negation	Certainty of negation Probability/neutrality of negation
	Suggestion	Emphasis Downtoning
	Tagging	Prompting
	Response to others	Short response
	Response to rhetorical question	Positive/negative Inference
PRAGMATIC FEATURES		
Indexical stance	Authority Solidarity Politeness	 Face-saving Face-attending Face-maintenance
Rhetoric-pragmatic functions	Emphasis Expectation Counter-expectation Hedging	 Adversative Concession Mitigation Tentativeness

2.7 Chapter summary

In this part of literature review, I have reviewed different theoretical perspectives on MACs. First, I discussed the general concept of epistemic modality. In the following discussion of MACs, I reviewed literature in order to establish the parameters on which my analysis will be founded. These parameters are the (formal) syntactic placement, and the semantic and pragmatic functions of each example of a MAC in context. With regard to syntactic placement of MACs, I reviewed discussion in the prior literature relating both to position within a clause (following Biber et al. 1999) and to the types of clause in which they tend to occur (following Boye 2012). I then reviewed literature on the semantic parameters of MACs and their categorisation as certainty and probability markers, and on the pragmatic and rhetorical-pragmatic functions that MACs express. Simon-Vandenbergen and Aijmer's (2007) comprehensive set of parameters for analysis of MACs has provided me a roadmap to follow in my own analysis. Collectively, all this has enabled me to develop and present a theoretical framework to be used in later chapters of this thesis for the analysis of this class of modal adverbs.

3 Literature review: Corpus-based descriptive and contrastive grammar

3.1 Chapter overview

Because this thesis project is a corpus-based contrastive descriptive study of Urdu and English MACs, it is necessary that it should include a literature review of both the corpus-based approach and some relevant descriptive studies. This chapter is structured as follows. I start with a brief introduction to corpus linguistics in section 3.2. Then, in section 3.3, I review literature on corpus-based descriptive grammar; in doing so I touch briefly upon pre-electronic corpora and reference grammars, before moving on to present day corpus-based descriptive studies. In section 3.4, I discuss corpus-based contrastive analysis and descriptive grammar studies, as well as prior research that *conjoins* the contrastive and corpus-based descriptive grammar approaches. At the end point of this second literature review , in section 3.5, I am able to present my completed framework for analysing English and Urdu MACs contrastively. Finally, in section 3.6, I lay out the specific research questions that my analysis will address in order to fulfil the aims of the research as introduced in Chapter 1.

3.2 Corpus linguistics and the corpus-based approach

Corpus linguistics is the study of language using corpora (the plural of *corpus*). McEnery and Wilson (2001, p.197) define a corpus as "a finite collection of machine-readable text, sampled to be maximally representative of a language or variety". Using machine-readable text means that we use computers to process and analyse large amounts of data stored on computer. It would be time-consuming for researchers to manually examine

such large data sets, in fact impossible in any feasible timespan. Corpus linguistics is often conceived to be a methodology[5] (e.g. Biber et al., 1999, pp.3-4; McEnery et al., 2006, p.7; McEnery & Wilson, 2001, p.197), meaning that corpus linguistics "focuses upon a set of procedures, or methods, for studying language" by using corpus data (McEnery & Hardie, 2012, p.1). The *corpus-based approach* is based on the view of corpus linguistics as a method or methodology.

Two defining characteristics of corpus-based analysis are adherence to the "principle of total accountability" (Leech, 1992, p.112), and the requirement that the corpus data to be explored "must be well matched" to the research questions (McEnery & Hardie, 2012, p.2). The principle of accountability in corpus linguistics means that the researcher must account for all instances of the linguistic phenomenon under examination, even if those instances contradict or undermine the original hypothesis. More specifically, McEnery and Hardie (2012, p.15) say that researchers should not filter out or ignore examples or statistical evidence from their corpus. Such manipulation may undermine the authenticity of the research data. Therefore, "there should be no motivated selection of examples to favour those examples that fit the hypothesis, and no screening out of inconvenient examples" (McEnery & Hardie, 2012, p.15).

Data and research question compatibility means that "a corpus is best used to answer a research question which it is well composed to address" (McEnery & Hardie, 2012, p.2). At the most obvious level, we may note that it is of no use to search for Urdu linguistic phenomenon in a corpus of English newspaper text instead of Urdu newspaper texts. Less trivially, if I am conducting contrastive research of a linguistic phenomenon in two

[5] Some researchers (see Sinclair, 2004, p.191; Teubert, 2005, p.2; Tognini-Bonelli, 2001, p.1) consider corpus linguistics to be a theory; but that view has generated criticism (see McEnery & Gabrielatos, 2006, pp.34-40; McEnery & Hardie, 2012, pp. 147-151). However, this debate is not relevant to the present study.

languages, I need data sets for each of those languages which are constructed so as to avoid the effects of any confounding variables on the phenomenon being contrasted.

A corpus may be monolingual (a corpus of a single language), bilingual (two languages), or multilingual (more than two languages). A bilingual or multilingual corpus can be *parallel* or *comparable*. Baker et al. (2006, p. 126) define a parallel corpus as one that "consists of two or more corpora that have sample texts and their translations". A typical parallel corpus will include, for some set of texts, each document in its original language plus that document's translation(s) into one or more other languages. A comparable corpus, on the other hand, is a bi- or multilingual corpus made up of different texts in each language, not translations but rather texts selected according to the same sampling frame (e.g. time periods, genres). Comparable corpora can also be sampled from different varieties of the same language (Hunston, 2002, p. 15). A corpus of Urdu and English short stories written in the last decade of twentieth century would be an example of a bilingual comparable corpus. Parallel and comparable datasets are used to compare "the lexis or grammar of different languages" (Baker et al., 2006, p.127) or dialects (Hunston, 2002, p. 15); their use in contrastive studies will be discussed in detail in section 3.6.

Sampson (2001, p.6) says that corpus linguistics is a form of empirical linguistics, because a corpus linguist makes appropriate use of tools such as concordancer software to examine real-life (attested) examples instead of relying on made-up (introspective) examples. A *concordancer* is a program that allows the user "to search a corpus and retrieve from it a specific sequence of characters of any length […] a word, a part of word, or a phrase" (McEnery & Hardie 2012, p.35). All instances of the queried item, and the immediate context of each instance, are displayed in one-example-per-line format (see Figure 3.1)(McEnery & Hardie 2012). McEnery and Hardie (2012, p.2) say that concordancers help corpus linguists to look at examples in context (i.e. in the environment of the surrounding text of the searched

word). McEnery and Hardie (2012, p. 35) say that the procedure of observing "*words* in their context" is "often called *key word in context* (KWIC) concordancing" . They further explain that the use of KWIC in concordancers is not necessarily limited to whole words. They cite an example from Baker (2009) in which, instead of whole words, a suffix is explored via concordancer, as in Figure 3.1 (reproduced from McEnery & Hardie, 2012, p. 36), which presents as a concordance the results of a query for English nominalising suffix *-ness*.

Exeter : the capital city of	**blandness**	ca n't break the chain TWO years ago the
after Yahoo warned it was seeing	**weakness**	in two of its biggest advertising segments .
advertising in the face of economic	**weakness**	. '' The warning , which was similar to
in the afternoon following news of more	**weakness**	in the US housing market , with housing
became a lifeline for Ross when his	**illness**	kept him confined to hospital and his home
Olympic bid . She was briefly married to	**fitness**	guru John Crisp . Her second marriage to
and are far less secure . " Ironically , this	**hyper-attentive-ness**	is actually having a damaging effect on our
thing is absolutely barmy . " And the	**madness**	continues with the goalposts constantly
notes led to a lifetime of passion and	**happiness**	. She was just 17 and he was 18 when they
so much , just to feel your warmth and	**warmness**	around me (lovely , lovely thought) . I 'm
Capt Alfred Bland MY only and eternal	**blessedness**	, I am never utterly miserable , not even
will and giving up a chance of supreme	**happiness**	with you . What more can anyone ask ?
made me during those days-just you , my	**sweetness**	. MY darling ... the nights here are weird .
range hens under threat as devastating	**illness**	nears Britain POULTRY farmers in
and has submitted it for inclusion in the	**Guinness**	Book of Records . 'Unfortunately , the

Figure 3.1 An extract of a concordance of words ending in *-ness* from the British English 2006 corpus (Baker, 2009), reproduced from McEnery and Hardie (2012, p. 36)

McEnery and Hardie note that concordancers are usually also able to produce one or more forms of frequency data, such as "a word frequency list, which lists all the words appearing in a corpus and specifies for each word how many times it occurs in that corpus" (McEnery & Hardie, 2012, p. 2). They say that analysis of concordances is a form of qualitative analysis and analysis of frequency data is a form of quantitative analysis, both of

which are equally important in corpus-based analysis. In line with that principle, this project will utilise both quantitative and qualitative methods (see 4.3).

3.3 Corpus-based descriptive grammar studies

This section deals with the two distinct beginnings of the application of the corpus-based approach within descriptive grammar studies: the early period of pre-electronic corpora prior to approximately 1960, and descriptive grammar studies in the 1970s and 1980s when linguists developed an interest in studying language in use (Johansson, 2008, p. 33). The latter period was when computers became more accessible to researchers in linguistics, making it easier for them to compile "frequency lists and concordances" (Johansson, 2008, p. 33). The availability of frequency lists and concordances facilitated a diverse range of analyses, such as description of specific grammatical features or of the behaviour of specific patterns or sequences of words. Though such analyses were done prior to the 1970s, analyses utilising corpora on the million-word scale were difficult, bordering onto impossible, using the less capable earlier computers (Johansson, 2008, p.33).

3.3.1 Pre-electronic corpora and descriptive grammar work

A *pre-electronic* corpus is one that is analysed manually without a computer. This form of analysis, the only way to utilise a corpus prior to approximately the 1960s, was often tedious and time-consuming (Meyer, 2008, p. 1). Two studies of grammar that used a pre-electronic corpus are Jespersen (1949) and Fries (1957). Their corpus-based work in descriptive grammar laid the groundwork for future research in that area. This section surveys Jespersen's and Fries' studies, and introduces the Survey of English Usage (SEU), a corpus

that was compiled before the period of computer-assisted corpus analysis, but was later digitised.

3.3.1.1 Jespersen's corpus-based grammar work

Jespersen's (1909-1949) grammar of English extends over seven volumes. He utilises a pre-electronic corpus that consists of quotations from the works of well-known authors across different written genres, such as poetry (e.g. Shelley), science (e.g. Darwin), philosophy (e.g. Locke), fiction (e.g. Woolf), including in each genre material from lesser-known authors as well (Jespersen,1949, vol. VII, pp.1-40). Jespersen (1949, vol. II, p. vi) says that he selected examples for addition to his corpus over many years through "both systematic and desultory reading". Jespersen indexes his paper corpus of quotations by adding the name of the author, the genre, the page number, and the line or sentence number at the start of each example.

Jespersen strongly believes that examples used in linguistic description should be from natural, real-life texts, instead of made-up examples. Jespersen cites an earlier scholar, Poutsma (1926),[6] who made similar use of examples from a pre-electronic corpus. While Jespersen (1949) does use some invented examples in his grammar, he avoids this where possible. The examples he cites are diverse, and whenever possible he "selected sentences that gave a striking and at the same time natural expression to some characteristic thought", supplemented with examples to present "grammatical peculiarities" of language in use (Jespersen, 1949, vol. II, p. vi).

A typical entry for a grammatical concept begins with a general discussion, and then quotes natural examples in support. For instance, Jespersen's (1949, vol. IV, pp. 237-240)

[6] Unfortunately, I was not able to locate a copy of Poutsma's original publication.

entry for *will* begins with the general comment that *will/would* as an auxiliary shares the "characteristic traits" of other auxiliary verbs such as *can,* for instance that it combines with an infinitive without *to.* Jespersen then quotes from his corpus to illustrate various uses of *will* as an expression of emphatic volition when "ascribed to lifeless things" (48), or of power or capacity (49), or of a habit as a consequence of character or disposition (50):

48. "[Mrs. Poyser] What is to be broke [=broken] *will* be broke" (Jespersen, 1949, vol. IV, p. 239, square brackets original).

49. "the Hall will seat five hundred" (Jespersen, 1949, vol. IV, p. 239).

50. "Many will swoon when they do look on bloud [blood] (Sh As IV. 3.159)" (Jespersen, 1949, vol. IV, p. 239).

Furthermore, Jespersen (1949, vol. IV, p.241) notes that certain patterns such as *will have it* have "no reference to future time" but rather are used to mean other things, in this case "[to] interpret, [to] apprehend", as in (51).

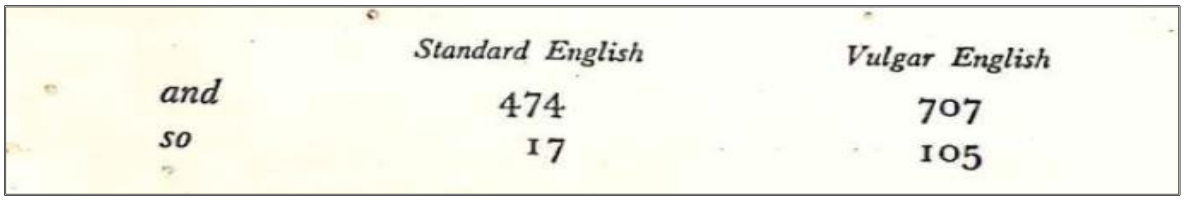
51. "Hope R 174 still she would have it that all men hailed him for their king".

3.3.1.2 Fries' corpus-based grammar study

Jespersen's (1949) corpus-oriented methodology is taken a step further by Fries (1957). Fries compiles a corpus of spoken data, rather than a written corpus like Jespersen's. Fries (1957, pp. ix-x) notes that this dataset is composed of "recorded conversations of speakers of Standard English in this North Central community of United States [Ann Arbor, Michigan]". Fries (1957, p. 3) explains that he recorded fifty hours of conversation, which were then "transcribed and roughly indexed" for reference. He further says that the recordings were made without the speakers knowing they were being recorded. Since the

Declaration of Helsinki in 1964, it is considered unethical to record a human subject without their prior knowledge, but in Fries' era, this was unexceptional practice. Fries (1957) mentions this to assure his readers that the spoken data is original, natural, and spontaneous, without any interference from the researcher. His purpose in using spoken data is to provide linguistics students with examples from "the language of the people", rather than examples limited to literature, which cover only one aspect of language use. Spoken data can assist in understanding "how a language works in fulfilling all the functions of communication in a particular social group that uses it" (Fries, 1957, p.4).

Fries' work is unique for its era because, in his descriptive analysis of various parts of speech in English, he not only uses corpus examples to illustrate his points, but also includes frequency information. While the frequency is mostly not given as explicit numbers, he often says that certain words or patterns are more frequent than others. Only in some instances does Fries compare the rate of occurrence of certain words numerically, or give percentages. For example, Fries (1957, p. 292) gives evidence in numbers for the greater frequency of English connectives *and* and *so* in "vulgar" (colloquial or non-standard) English than in Standard English; see Figure 3.2.



	Standard English	Vulgar English
and	474	707
so	17	105

Figure 3.2: Frequency of connectives in Standard and non-standard English (reproduced from Fries, 1957, p.292)

Both Jespersen's (1949) and Fries' (1957) grammars are now very old. Their non-computerised corpus-based analysis is basic in comparison to present-day studies. However,

this work is the precursor of present-day corpus-based grammar. Both Jespersen and Fries document language use via natural data, which is a core procedure for corpus-based studies.

3.3.1.3 The Survey of English Usage

The earliest multi-genre corpus to be made broadly available started out as a non-electronic corpus. This is the Survey of English Usage Corpus (SEU). Randolph Quirk founded the Survey of English Usage research unit at University College London in 1959[7]. Compilation of the SEU was undertaken under Quirk's direction; Quirk acknowledges as forebears descriptive grammarians including Jespersen (1949) and Poutsma (1926). However, Quirk (1974, p. 167) observes that these earlier efforts using pre-electronic data lacked organised categorisation of text types (genres) in their corpora. Therefore, their use of sources at times "leaves unclear the distinction between normal and relatively abnormal structures and the conditions for selecting the latter" (Quirk 1974, p. 167). Quirk implicitly criticises previous works such as Jespersen's descriptive entries (see 3.3.1.1), since Jespersen states that he uses both distinctive and common examples without clearly explaining the distinction in the examples given. Quirk's corpus design for the SEU (given in Figure 3.3) utilises a systematic structure of categories that became a benchmark for future corpus compilation.

[7] https://www.ucl.ac.uk/english-usage/about/index.htm

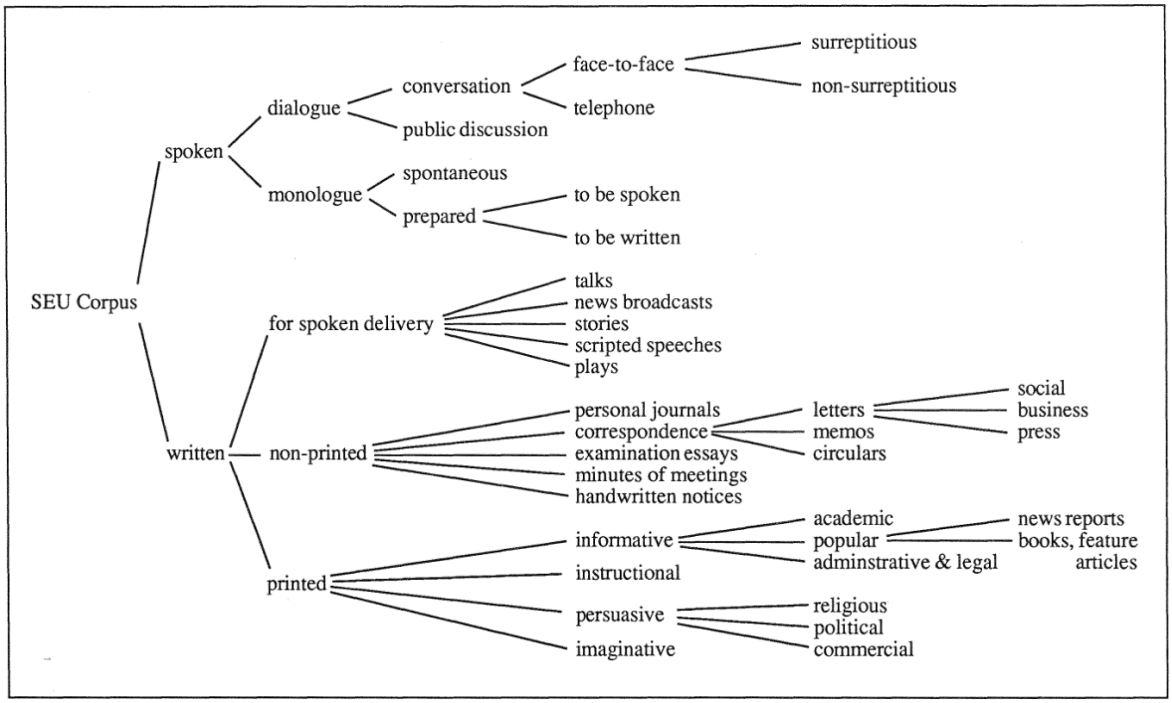


Figure 3.3: General structure of the SEU corpus (reproduced from Greenbaum & Svartvik, 1990, p.13)

The SEU spoken data was recorded using a tape recorder, with the audio stored on reel-to-reel tapes, and transcribed on paper (as this was still prior to the period of computer-based corpus linguistics). The transcribed data was stored in filing cabinets and indexed using paper cards; the written data was treated similarly (Meyer, 2008, pp.11-12).

The purpose of the SEU was to provide natural language data (both written and spoken) to ensure an accurate description of English grammar (Greenbaum & Svartvik, 1990, p.11). Greenbaum and Svartvik (1990, pp. 13-14) say that the SEU Corpus was grammatically analysed for "65 grammatical features, over 400 specified words or phrases, and about 100 prosodic paralinguistic features". These analyses were recorded in a comprehensive index that could then be used to access examples in the data. For example, index cards with references to all examples of noun phrases were stored in a dedicated noun phrase filing cabinet. Linguists could then build upon this data for their own grammatical

analysis. The SEU corpus included multiple different types of English (e.g. spontaneous conversations, newspaper reportage, technical writing, prose fiction). But it was difficult to study this data without any computerised database. Therefore, researchers mainly drew on the grammatically analysed texts for reference examples in their own research, rather than make more extensive use of corpus-based methods. For instance, Meyer (2008, p.12) says that in his own (1987) study, he quotes examples of "appositives in English (e.g. constructions such as *my friend, Peter*) based on slips of appositives in the SEU Corpus". Quirk et al.'s (1985) major reference grammar utilises examples from the SEU for illustrative purposes (in addition to other corpora, see 3.3.2.2). Because of its pre-computational storage, using this data during the first twenty years of its existence (from the late 1950s) required visiting the Survey in person (Meyer, 2008, p.12). Greenbaum and Svartvik (1990, p. 11) report that the SEU was later computerised, and some of its spoken material was made part of the London-Lund Corpus (LLC), along with data from a second project, the Survey of Spoken English (SSE), originated by Svartvik in 1975.

3.3.2 Reference grammars and corpora

This section discusses the so-called first-generation corpora, and overviews two well-known English reference grammars that have used corpora. The first-generation corpora were intended to "provide a source for an accurate description of the grammar of adult education of English" (McCarthy & O'Keefe, 2009, p. 1009). Their use in detailed corpus-based descriptions of the grammar for English language users is demonstrated by Quirk et al.'s (1985) and Biber et al.'s (1999) reference grammars. These two reference grammars, especially Biber et al.'s (1999), provide a good exemplar for later linguists to use corpora for description of some grammatical feature.

3.3.2.1 First-generation corpora

Johansson (2008, p.33) observes that although we can trace corpus-based studies back to pre-electronic times, as previously discussed, the real breakthrough in corpus studies started in the 1960s and reached fruition in the 1970s and 1980s. This was the time at which more than a minimal number of texts became accessible in machine-readable form. Text files could be "stored, transported, and analysed electronically" via computer. This enabled hitherto inconceivable analyses using large amounts of data, such as compiling frequency lists and generating concordances in moments.

The Brown Corpus, the Lancaster-Oslo/Bergen Corpus (LOB), and the London-Lund Corpus (LLC) are prominent among the English corpora that came to be known as "first generation corpora" (Kennedy, 1998, p. 23). The earliest machine-readable corpus of this modern kind is the Brown Corpus, a collection of Standard American English made up of one million words from fifteen different text categories (Kučera & Francis, 1967). Compilation of its British counterpart, LOB, began in the early 1970s and was led by Geoffrey Leech at Lancaster University (Leech & Leonard, 1974). Later this project was completed through a joint effort of two institutions in Norway, the University of Oslo and the Norwegian Computing Centre for the Humanities at the University of Bergen. The structure of the corpus was matched closely to that of the Brown Corpus in terms of sources and sampling size (Johansson et al., 1978). The composition of the two corpora is shown in Figure 3.4 Like Brown, LOB aimed to include samples of language from "a wide range of styles and varieties of texts, including both informative (A-J) and imaginative (K-R) prose" (Johansson, 2008, p.36). Quirk et al. (1985) utilised both Brown and LOB in their reference grammar of English.

Text categories		Number of texts in each category	
		Brown	LOB
A	Press: reportage	44	44
B	Press: editorial	27	27
C	Press: reviews	17	17
D	Religion	17	17
E	Skills, trades, and hobbies	36	38
F	Popular lore	48	44
G	Belles lettres, biography, essays	75	77
H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
J	Learned and scientific writings	80	80
K	General fiction	29	29
L	Mystery and detective fiction	24	24
M	Science fiction	6	6
N	Adventure and western fiction	29	29
P	Romance and love story	29	29
R	Humour	9	9
	Total	500	500

Figure 3.4: Composition of the Brown and LOB corpora (reproduced from Johansson, 2008, p.36)

3.3.2.2 Quirk et al.'s (1985) use of corpora

Quirk et al.'s (1985) *Comprehensive Grammar of the English Language* is a major reference grammar that draws data from three corpora: the SEU, LOB, and Brown. Data from elicitation experiments are also used. It has been called a "comprehensive description of present-day English" (Lindquist, 2013, p. 5951) and utilised as a reference point in many studies of English. Quirk et al. (1985, p.16) say that their grammar covers variation in English based on region, field of discourse, medium, attitude, and social variation; it is thus arguably comprehensive from a varietal perspective as well as a subject matter perspective.

Among regional varieties, Quirk et al. focus on the standard British and American forms over any others, not only describing but also comparing these dialects. The differences between British and American English that they identify are more often based on statistical, stylistic and pragmatic preferences than on categorical lexical or grammatical differences. For instance, Quirk et al. (1985, p. 395) explain that to convert a positive clause to a negative clause, "the clause negator *not*" is added between "the operator [e.g. *can, has, do*] and the predication [i.e. part of the clause following subject and operator]", as in *I have finished ~ I have not finished*, except that in British English (BrE) negative clauses with main verb *have* are usually formed using the participle *got* in informal style, rather than using operator *do* as in American English (AmE).

52. "He has enough money { He doesn't have enough money. <esp. AmE> / He hasn't (got) enough money. <esp. BrE>"

(Quirk et al., 1985, p. 776)

Quirk et al. do not take into account some of what would become established as core practices of corpus-based studies. They do not provide quantitative evidence from the corpora, except in a limited way explained below. Their grammar at times presents examples and usages in a prescriptive manner. They assert, however, that they mention prescriptive use of English not to promote that usage, but to inform readers on the accepted standard use of the language. Quirk et al. (1985, p.34) stress that their study is a "descriptive presentation of English morphology and syntax". However, they also point out uses of language in informal context as unacceptable. For instance, they say that if the operator (initial auxiliary) is the only verb in a verb phrase, then it is unacceptable to use a contracted form of the verb, as illustrated in examples (53) and (54).

53. “Will you be in tonight?

“*’ll you be in tonight?”(Quirk et al. 1985, p. 123).

54. “No, but I *will* tomorrow night”

“*No, but I’*ll* tomorrow night”(Quirk et al. 1985, p. 123).

Their statements on things being grammatically correct/incorrect, or acceptable/unacceptable, perhaps reflect a concern that theirs is a reference grammar that will be used by teachers and students of English as a second language as well as by linguists. This concern is evident from the first chapter (Quirk et al., 1985, p. 3) in which they discuss the importance of English as a means of communication.

Quirk et al. (1985) do attempt to give some frequency data in the manner typical of corpus-based analysis. However, the statistics that they provide are not based on their own corpus, but are quoted from earlier research (e.g. Coates, 1983; Granger, 1981; Hofland & Johansson, 1982), as discussed in “bibliographical notes” at the end of each chapter (e.g. Quirk et al., 1985, p.171). Thus, while using corpus examples and, to this limited degree, quantitative evidence, their grammar cannot be considered wholly a corpus-based work. Of course, Quirk et al. do clearly explain this in the preface, and acknowledge throughout the book their sources for examples and statistics.

Quirk et al. (1985) remains one of the most comprehensive reference grammars for contemporary English. Their description of syntactic forms and their meanings, inclusive of semantic and pragmatic information, stands as a good model for future descriptive grammar studies. The successor to this work, Biber et al. (1999), is more wholly situated in the corpus-oriented tradition of descriptive analysis.

3.3.2.3 Biber et al.'s (1999) corpus-based grammar

The *Longman Grammar of Spoken and Written English* is a corpus-based reference grammar of English by Biber et al. (1999). Biber et al. investigate how language is used in different ways across genres and in speech versus writing. For data, they use the Longman Spoken and Written English Corpus (LSWE) of 40 million words (37,244 texts). Their corpus is divided into three parts, as Figure 3.4 shows: core registers, American texts for dialect comparison, and supplementary registers.

	number of texts	number of words
core registers		
conversation (BrE)	3,436	3,929,500
fiction (AmE & BrE)	139	4,980,000
news (BrE)	20,395	5,432,800
academic prose (AmE & BrE)	408	5,331,800
AmE texts for dialect comparisons		
conversation (AmE)	329	2,480,800
news (AmE)	11,602	5,246,500
supplementary registers		
non-conversational speech (BrE)	751	5,719,500
general prose (AmE & BrE)	184	6,904,800
total Corpus	37,244	40,025,700

Figure 3.5 LSWE corpus composition, reproduced from Biber et al. (1999, p.25)

Biber et al. (1999) use the "core registers" (see Figure 3.4) for the analysis of register variation. Biber et al. (1999, p. 35) grammatically tag the entire data using an automated program. Automated analyses of the tagged corpus is complemented by manual checking of samples of concordances, followed by qualitative, functional interpretation of their quantitative findings.

Prior to their analyses, they give a detailed account of how they will interpret their quantitative data grammatically and functionally (Biber et al. 1999, pp. 35-43). Their description starts with small-scale units which they refer to as *elements*, and builds into larger, more complex units. An element is typically presented in a four-part arrangement: a description of its form and structure; corpus findings on distributional patterns of that feature; corpus frequencies presented as bar charts; and finally interpretation of the quantitative results in terms of the function(s) of the element. Biber et al. (1999, p. 41) say that their functional interpretation covers three major areas: "the **work that a feature performs** in discourse" (such as textual tasks, where words are used to form a coherent text); "the **processing constraints** that it reflects" (such as the use of contractions under time constraints); and "the **situational** or **social distinction** that it conventionally indexes".

Biber et al. also incorporate analysis of the interface between grammar and discourse pragmatics by looking at what Biber (1997) terms *lexical bundles*. Biber et al. (1999, p. 989) note that words "show a statistical tendency to co-occur" in longer sequences, which therefore are "extended collocations"; the term *lexical bundles* is used for such sequences. They note that there is a difference between lexical bundles and idioms, in that each word retains its meaning in a lexical bundle, whereas individual words in an idiom – in the sense of a fixed multi-word form with non-compositional meaning – do not. Moreover, Biber et al. report that lexical bundles are used differently in distinct registers. For example, common lexical bundles in conversation include *have a look at* and *I don't know what*, but in academic prose, common bundles include *is based on* and *as a result of* (Biber et al., 1999, p. 996).

Also notable is Biber et al.'s approach to differences between speech and writing, particularly those that concern the structure of spoken language. They say that the spoken medium has looser patterns than the structured patterns of written sentences (Biber et al., 1999, p. 998). Therefore Biber et al. (1999, p. 1069) propose *clausal* and *non-clausal units* as

the governing principle of conversation. Their model is supported with evidence from utterances in the corpus which they analyse by "segmenting dialogue into grammatical units", that is, dependent and independent clausal units and non-clausal units. Biber et al. (1999) define a non-clausal unit as a unit that is not a part of any clause, dependent or independent, in a conversation. In (55), for instance, the bold units are non-clausal according to Biber et al. (1999) (note that || is their unit-boundary symbol). They say that non-clausal units can also be found in written texts (e.g. newspaper headlines, essay headings), but that these units are more pervasive and extensive in spoken discourse than in written.

55. "A: || **No**, || I would wen give you that chair in there

B: || **Mm-** ||

A: || It came from Boston, by covered wagon. ||

B: || That's such a neat, || it's so nice to know the history behind it. ||

A: || **Yeah,** || **yeah**|| " (Biber et al. 1999, pp.1069-70)

We thus see that Biber et al. (1999) successfully use corpus data to describe how language varies according to the purpose of communication and the situations in which it is used. Their grammar's grounding in a systematically compiled corpus allows them to give clear and detailed description of language variation. This robust corpus-oriented model still influences corpus linguistic studies today. For instance, subsequent work on stance in spoken and written university registers by Biber (2006), in threatening communications by Gales (2010), on stance in texts by Hunston (2007), and in legal research articles by Adams and Toledo (2013) all builds on Biber et al.'s (1999, pp. 965-978) analysis of stance. Similarly, the lexical bundle approach has been utilised in work such as a comparative study of lexical bundles in conversation and academic prose by Conrad and Biber (2005) and research on connecting lexical bundles and rhetorical moves in research article introductions by Cortes

(2013). Researchers also extensively follow Biber et al.'s frameworks, especially lexical bundles, in research on teaching English as a second language (Biber et al. 2004; Conrad, 2010).

3.3.3 Present day corpus linguistics and descriptive grammar studies

Due to the easy accessibility of corpora in the present day, corpus methods are now widely used across various linguistic disciplines, for instance, corpus-based translation studies (Sutter & Lefer, 2020), corpus assisted discourse analysis (Baker, 2020), corpus-based language teaching (McEnery & Xiao, 2011), and corpus-based language description (Conrad, 2010). As this study's focus is descriptive grammar, I discuss what is entailed by corpus-based descriptive grammar studies in section 3.3.3.1. This is followed by a consideration of best practices in corpus-based grammatical research in section 3.3.3.2.

3.3.3.1 What is a corpus-based descriptive grammar study in the present day?

In this section, I briefly touch upon the difference between descriptive and prescriptive grammar, and the role corpus analysis has played in the development of descriptive grammar studies.

Since "the beginning of the electronic era", "studies of grammar have progressed and proliferated enormously through the development and exploitation of corpora" in numerous languages (Leech, 2015. p. 146). A corpus-based descriptive study of grammar is one which uses evidence of grammatical usage as observed in a corpus or corpora to present some form of "description of what the corpus attests about some area(s) of grammar" (Leech, 2015, p. 146).

Descriptive grammar is often contrasted with *prescriptive* grammar (Leech, 2015, p.146). The traditional, prescriptive perspective on grammar present forms in a language as acceptable or unacceptable, accurate or inaccurate, grammatical or ungrammatical. But this dichotomous view of grammatical accuracy plays a very minor role in contemporary grammar (Conrad, 2010, p.227). For instance, a descriptive grammar of English might note that an article (*a, an, the*) appears before a singular count noun (e.g. *a book*) in English. With few exceptions, one being "in locative preposition phrases" (e.g. *at school*), "this rule is absolute" (Conrad, 2010, p.227). Descriptive grammar does not assert that this form is *right*, or that not using it is *incorrect*, as prescriptive grammar, aimed at enforcing a fixed standard, would; descriptive grammar merely asserts as an observation that all speakers apparently do follow this rule. Descriptive grammar studies mainly aim to provide "quantitative and functional (qualitative) analyses" of language "as evidenced by usage" (Leech, 2015, p.147).

Leech (2015, p. 146) points out a misconception regarding corpus-based descriptive grammar. It is wrongly assumed that corpus linguists are content to "describe" which structures are found in a language and present their frequencies, rather than to express their findings "in terms of some theoretical framework(s) of how grammar works" (Leech, 2015, p. 146). In support of Leech's point, McEnery and Hardie cite a number of theoretical linguists who have incorporated "corpus evidence into their analyses" (McEnery & Hardie, 2012, p.175). For example, they cite Temperley (2003), who follows a functional theoretical approach proposed by Arnold et al. (2000), which asserts that there exists a relationship between formal phenomena (e.g. word placement in a clause) and functional phenomena (e.g. semantics) and that this relationship can be analysed by corpus investigation of sets of examples using quantitative analysis (i.e. statistical tests such as correlation, significance). Temperley (2003) investigates a specific type of syntactic ambiguity in relative clauses in English: a relative clause with a zero relative pronoun may lead to the first word of the

relative clause being interpreted as part of the main clause. Temperley (2003, p.473) gives the example of "*the biological toll that logging* can take" compared to "*the biological toll logging can take"*. In the latter, *logging* may be read as part of the main clause, before *can*, a modal verb, indicates that the word preceding it is a subject.

Temperley uses the Penn Treebank (Marcus et al., 1993) to identify and extract all examples of relative clauses where a zero pronoun is possible, "using a computer program especially designed for the purpose" (Temperley, 2003, p. 475). He tests two hypotheses: that zero relative pronouns are less common in clauses where syntactic ambiguity may occur; and that anaphoricity of the relative clause subject motivates use of zero relative pronouns (Temperley, 2003, p. 465). His statistical findings show that both hypotheses are correct. That is, zero relative pronouns are less common in clauses where temporary syntactic ambiguity occurs, but occur more readily in relative clauses whose subject is an anaphor. McEnery and Hardie (2012, p.175) argue that studies like Temperley's represent an effort by theoretical linguists to create "links between corpus linguistics and functional theory".

Before the availability of corpora for grammatical studies, "theoretically oriented grammarians" either tended to ignore what they considered uninteresting from a theoretical point of view, or else did not have resources to analyse "peripheral areas of grammar" in detail (Leech, 2015, p. 151). For instance, very little description of English adverbs and adverbial phrases is presented by early grammarians (Stroik,1992, p.375). Later corpus-based grammars, by contrast, dedicate entire chapters to adverbs (Biber et al., 1999, pp. 503-570; Quirk et al., 1985, pp. 399-653).

McEnery and Hardie say that there are also researchers who, as "corpus specialists", use corpus data to extend "functional theories in order to adequately characterise their data" (McEnery & Hardie, 2012, p.175). For instance, Hasselgård (2010, p. 12) provides an in-

depth quantitative and functional study of adverbs in English, in which she presents corpus evidence of how adverbs are realised in different clausal positions, and of how adverbs combine with other elements. Overall she characterises the "syntactic and contextual factors that govern adverbial usage and how they can be best described" (Hasselgård, 2010, p.12). Her analysis also involves "the description of usage" of adverbs in English as a discourse feature (Hasselgård, 2010, p.12).

3.3.3.2 Methodological considerations for the corpus-based study of grammar

Three considerations that are important in any corpus-based study, whether of grammar or not, are corpus design; in-depth, fine-grained analysis; and frequency analysis (Conrad, 2010, p. 228).

The *design of the corpus* must be representative of the language or variety under investigation, that is, the corpus compiled or chosen for analysis must match the intended research. For example, if the goal of the research is to analyse the use of *that*-clauses in English newspaper editorials, then a collection of English newspaper business articles will not represent the variety of language under analysis. Data selection can be further refined if a compiled corpus of editorials stores sub-types, say, British newspapers and American newspapers as separate text files. While designing a corpus, time frame may be an important consideration in addition to genre and/or dialect. For instance, if editorials from British and American English newspapers are compiled for comparison, then the time period sampled for the American English texts should match that sampled for the British English texts.

A grammatical analysis needs to be *fine-grained*. A *fine-grained* study is an in-depth examination of some linguistic element at both the structural (syntactic) and functional (semantic and pragmatic) levels. For instance, McCarthy and Carter (2001) provide a detailed

description of the contexts in spoken English in which subject ellipsis does and does not occur. They note, for instance, that in narratives, "the character and events are *displaced* in time and space" and thus in this context of speech those entities have "explicit references" when they are clause subjects, in contrast to other spoken genres where "subjects are ellipted" (McCarthy & Carter, 2001, p. 209). This fine-grained analysis would not be possible without computer-assisted analysis of large-scale data: corpus software allows the authors to extract both ellipted and non-ellipted constructions. This study demonstrates how differences among genres may result in a shift in speakers' linguistic preferences, and therefore provides corpus evidence for "the grammar of speech and of writing [being] very different" (Leech, 2015, p.153).

Frequency analysis allows us to observe whether an element or pattern under investigation is more or less frequently used in particular contexts. Frequency analysis may be simple, for example, presentation of percentages of occurrences; alternatively, it may be based on advanced exploratory or hypothesis-testing statistical calculations. Biber et al. (2004) argue that it is not sufficient in a descriptive analysis to present frequency counts, because they are not self-explanatory. Rather, we need to further examine and interpret frequency data in order to "identif[y] patterns that must be explained"; one especial virtue of frequency data is that "it identifies patterns of use that otherwise often go unnoticed by researchers" (Biber et al., 2004, p. 176). Thus, frequency information enables us to identify both typical and unusual examples of language use.

3.4 Contrastive linguistics and corpus-based analysis

3.4.1 Defining the terms contrastive linguistics and corpus-based contrastive analysis

As already explained, the corpus-based approach is a *method* (or methodology) for the study of language that allows a researcher to explore some linguistic phenomenon across copious amounts of data. Contrastive analysis, on the other hand, is the synchronic study of two (or more) languages, comparing corresponding items or structures to identify similarities and differences (Ebeling & Ebeling, 2013). This form of contrastive research often straddles theoretical and descriptive linguistics. Theoretical linguistics provides the frameworks for comparison of the languages; descriptive linguistics provides the informative description(s) of the languages under study.

The combination of these two approaches, a corpus-based contrastive analysis, is a comparative study of some corresponding linguistic feature(s) and/or structure(s) in two or more languages, one that aims to identify similarities and differences based on natural language data in corpora of those languages. This approach, joining corpus-based methods and contrastive analysis, provides the analytical framework for this thesis (see 3.5).

In the remainder of section 3.4, I start by considering the development of corpus-based contrastive studies and some current trends in 3.4.2. In 3.4.3, I briefly discuss parallel and comparable corpora from the perspective of corpus-based contrastive studies. Finally, in 3.4.4, I discuss views on the appropriate type of corpus for contrastive descriptive studies.

3.4.2 Corpus-based contrastive studies

Hasselgård (2020, p. 184) says that in the early 1990s, when multilingual corpora were introduced, they represented not only a "more solid empirical basis, new method and a general boost" relative to earlier monolingual corpora, but also a refinement in contrastive analysis. Until that point contrastive linguistics was an applied linguistics discipline associated with improving language teaching and translation through comparing similarities and differences between languages (Johansson, 2007, p. 1). But since that point, linguists of varied theoretical persuasions now use corpora to complement evidence from other sources (Kennedy, 1998, p. 8).

Mair (2008, p.10) says that the arrival of multilingual corpora redefined the aims of contrastive analysis by shifting its focus from the study of a "decontextualised system of choices" to the descriptive comparison of "language use in context". By "language use in context", Mair (2008, p.10) refers to Biber's (1994) explanation of the long-established notion that speakers make linguistic choices based on the situation in which they are communicating, and that this leads to linguistic variation dependent on communicative context, that is, register variation (see 3.3.2). Barlow (2008, p. 103) claims that by using multilingual parallel corpora we can understanding the relation between grammatical structures of linguistic systems and instances of use. To exemplify this, Barlow (2008, p. 103) uses a part of the French and English data from the European Parliament corpus (*Europarl*: Philipp, 2005), the analysis of which "provides information on usage and hence can lead to insight into the nature of grammar". Barlow (2008, p. 105) says that by using relevant frequency data, corpus-based contrastive analysis helps achieves "a more objective picture of the degree of correspondence of patterns" between languages. Also using French and English data, Barlow (2008) applies his tool *ParaConc* to identify recurrent patterns involving the

English verb *go* and the corresponding French *aller* 'to go'; ParaConc is a concordancer able to locate translation equivalents and the constructions they occur in. Barlow (2008, pp. 120-121) also stresses the importance of determining equivalent collocations, and the distribution of these collocations in different text-types. In a retrospective discussion on research of this kind, Gonzalez et al. (2008, p. xvii) say that use of empirical evidence in corpus-based contrastive studies represents "a step forward in terms of testability, authenticity and general empirical adequacy" for contrastive analysis.

The current trend in corpus-based contrastive analysis, since multilingual corpora became broadly available in the 1990s, is a shift towards greater diversity within individual studies, in terms not only of language pairs under scrutiny, but also of the areas of grammatical description that analyses address. Xiao (2008, p. 435) says that newer multilingual corpora comprise diverse text types (such as newspapers, novels, online forums) as compared to earlier such datasets, noting the heavy reliance on literary translations in many parallel corpora of the 1990s. The availability of diverse multilingual data has increasingly supported varied types of cross-linguistic investigation of genre or register. Research of this type to date has included descriptions of similarities and differences in lexis, in syntactic patterns, and in the semantic and pragmatic factors that motivate different usages in two or more language (Kenning, 2010, p. 492).

Egan and Dirdal (2017, pp. 7-8) observe that early studies in contrastive analysis focused heavily on verbs (e.g. Engel, 1999) and adverbs (e.g. Hasselgård, 2004). For instance, Engel (1999) compares present perfect forms (present tense of *have* plus past participle) in English and the equivalent French *passé composé*. Engel's corpora consist of randomly selected "talk shows and news bulletins" from British and French radio. The analysis is qualitative rather than quantitative, focusing on the present perfect's contrasting functions in the two languages. Engel finds that the *passé composé* is mostly used by talk

show and news bulletin presenters to create links between items being discussed during a talk show, but it is not the only "past tense of narration", because other forms are used by the French speakers for that purpose. On the other hand, Engel (1999, p.266) suggests that the tendency of English speakers to use indirect report while referring to past events in English triggers the replacement of the present perfect with the simple past tense. When adverbials such as *fairly recently* are used within narration of past events, the simple past tense is used. In both languages, the present perfect is used "in introducing and summarising" an event, and "is linked with certain temporal adverbials" (Engel, 1999, p. 267). Engel concludes that sub-genres (in Biber & Conrad's 2019 terminology, *registers*) exist within the genre of radio talk, and that these reflect linguistic choices made by speakers in particular communicative contexts (Engel, 1999, p. 271). Therefore, Engel (1999, p.271) says, like "newspaper language", radio talk shows and news bulletins cannot be considered to be a "unified genre but a collection of subgenres".

Later corpus-based contrastive analyses have covered parts of speech other than verbs/adverbs, such as pronouns (Coussé & Van der Auwera, 2012) and prepositions (Egan, 2013). Furthermore, corpus-based contrastive analysis has over the years increasingly addressed the semantic and pragmatic functions of lexical items (e.g. Simon-Vandenbergen & Aijmer, 2007). This can be seen in Pešková's (2019) contrastive study of pro-drop (subject pronoun omission) in Czech and Spanish. Pešková examines passive-like generic constructions with generic pronouns *uno* 'one' (Spanish) and *člověk* 'man, human being' (Czech), and finds that pro-drop occurs in a wider range of contexts in Czech than in Spanish. In Czech pro-drop appears, for instance, in echo yes/no questions and with auxiliary-drop (omission of the finite auxiliary). An example of Pešková's (2019, p.320) contrastive pragmatic-discourse analysis is her account of the neuter demonstrative pronoun in Spanish (*ello* 'it') and Czech (*to* 'the/that/it'). She says that Spanish *ello* has negative connotations, its

use being associated with speakers of a lower educational level, whereas Czech *to* is emotive, used to express "surprise, joy, disappointment, or warning" (Pešková, 2019, p.320) and without class association. However, Pešková (2019) acknowledges that her study lacks systematic quantitative analysis to support her qualitative findings.

Aijmer and Lewis (2017, p.3) say that researchers in the field of cross-linguistic contrastive analysis have realised that "different social and cultural practices" influence and generate linguistic patterns "in the compared languages". Consequently, contrastive researchers now increasingly focus on cross-linguistic similarities and divergences in registers, as defined by Biber (examples include Aijmer & Lewis, 2017; Biber, 2014; Kunz et al., 2018; McEnery & Xiao, 2010; Neumann, 2013).

Let us consider one example of contrastive analysis in detail. Neumann's (2013) study exemplifies corpus-based cross-linguistic analysis of registers. She uses a corpus of English and German texts and translations from the CroCo Corpus (Neumann, 2005). Neumann (2013, pp.84-86) draws on eight registers to investigate three types of cross-linguistic variation: between English and German, across registers, and between original texts and the corresponding translations. She identifies linguistic features which vary across registers. For instance, she examines translation-equivalent epistemic markers, such as modal verbs *must* in English and *müssen* in German, and adverbs *possibly* in English and *vielleicht* in German. Neumann finds that among informational registers, instructional manuals convey a higher degree of writer authority than, say, scientific articles. On the other hand, scientific articles are among the registers that make prominent use of hedges to express lack of certainty regarding assertions (Neumann, 2013, pp. 142-46; p.197). Neumann then contrasts these identified elements on the basis of a parallel corpus, observing that "in sentence-aligned German-English texts" (Neumann 2013, p. 295), the majority of personal pronouns have direct one-to-one matches, with few exceptions. But she also observes that personal pronouns

may be lost in translated texts because they get translated by other words or phrases. For instance a specific use of German nominative *ich* 'I' may correspond to "a circumstantial prepositional phrase *in the leading role*" in English translations (Neumann, 2013, p. 295). Neumann observes that such loss of personal pronouns constitutes "translation shift", which is typical in translated texts "regardless of register and translation direction". In most of these analyses, Neumann combines qualitative analysis with quantitative analysis to determine the extent of linguistic variation across registers and between languages.

3.4.3 Parallel and comparable corpora

As discussed earlier (see 3.2), the two types of multilingual corpus are parallel corpora and comparable corpora. These are fundamentally different. A parallel corpus is a collection of the same set of texts in two or more languages. Each sample collected for the corpus consists of an original text in one language (the source language) and translations of that original text into one or more other languages (the target language or languages). Parallel corpora can be unidirectional (with a fixed source language), or bi-directional (the source and target languages switch places for some fraction of the texts). Typically, the translated texts are *aligned* at sentence level (see 4.2.1). Major uses of parallel corpora are in translation studies and comparative linguistics (Aijmer & Altenberg, 2013, p. 4).

A pioneering parallel corpus from the early 1990s is the English-Norwegian Parallel Corpus (ENPC) (Johansson & Hofland, 1994). Johansson and Hofland (1994) refer to parallel corpora simply as *multilingual corpora*, a practice which for clarity I will not follow. Johansson and Hofland (1993, p. 25) say that parallel corpora enable "text-based contrastive studies" to take the place of traditional comparisons of languages considered as abstract systems "not connected to real texts". Johansson and Hofland (1993, p.36) conclude that if

parallel corpora are properly compiled and used, they can enrich comparative studies of languages.

A comparable corpus, on the other hand, is not a compilation of translated texts. Rather, it consists of collections of texts in two or more languages which are comparable based on the parameters according to which they have been collected. These comparable parameters, or sampling criteria, are determined by the corpus compiler and may include, among other things, content/subject matter, medium, time of creation, and formality level. Comparable corpora are used in translation studies to identify differences and equivalences between languages (Hunston, 2002, p.15). An example of a comparable corpus of sample texts from different varieties of the *same* language is the *Corpus de Referência do Português Contemporâneo* ('Reference Corpus of Contemporary Portuguese', CRPC), approximately 334 million words of multiple geographical varieties of Portuguese from around the world (Do Nascimento et al., 2006, p. 1791).

An example of translation research using comparable corpora is De Cock and Goossens (2013). These authors use comparable corpora of business news in a contrastive analysis of quantity approximating devices, such as quantifiers (e.g. *many*, *some*), nouns (e.g. *loads*), precise numbers (e.g. *23*), and imprecision (e.g. *about 50*), across English and French. As a starting point, they located numbers in their two English and French comparable corpora of business news reporting (De Cock & Goossens, 2013, p. 142). Then, they manually scanned the concordances of the retrieved examples. They use these examples to compare the semantic and grammatical characteristics of quantity approximators in English and French (De Cock & Goossens, 2013, p. 143). De Cock and Goossens identify both similarities and differences. For instance, while the languages share eleven grammatical categories of quantity approximator, determiners as a means of expressing quantity approximation are a feature of English but not French. Other findings include the result that, in French but not

English, verbs are the preferred means of expressing quantity approximation (De Cock & Goossens, 2013, p.153).

De Cock and Goossens' (2013) study exemplify how corpus-based contrastive analysis of grammatical feature(s) in two or more languages can help identify elements used similarly or differently to perform some function across the languages. With corpus tools, it is much easier to analyse each instance of language use in context than it would be if the only possible method was manually going through data.

3.4.4 The appropriate type of multilingual corpus for a contrastive study

The disciplines of contrastive studies and translation studies both use parallel *and* comparable corpora, but with different aims. In translation studies, the focus is not on comparison and description of the languages, as it is in contrastive analysis. Rather, translation studies "often focus[es] on the translation process, features of translated texts and/or the application of findings to translation practice" (Hasselgård, 2020, p. 186). As the present thesis is a corpus-based contrastive study, there is no need for further discussion of corpus selection for translation studies.

In corpus-based contrastive studies, both parallel and comparable corpora are likewise used. In the mid-1990s, when multilingual corpora were first introduced to contrastive linguistic studies, there was a strong preference for use of parallel corpora (Aijmer & Altenberg, 2013, p.1) over comparable corpora. Later, following Johannsson and Hofland's (1994) use of the ENPC parallel corpus (see 3.3.2.2), there was a shift towards integrating both kinds of corpus (Aijmer & Altenberg, 2013, pp.2-3). Yet discussion of which of the two types of multilingual corpora is better for contrastive study has continued. I present

arguments in favour of and against parallel corpora in 3.4.4.1; then arguments in favour of and against comparable corpora in 3.4.4.2; and the resolution of the problem in 3.4.4.3.

3.4.4.1 Use of parallel corpora in contrastive studies

Those who favour the use of parallel corpora in contrastive analysis reason that such a corpus documents native speaker intuitions arising from their proficiency in the correct use of their language (Aijmer, 2008, p.278). For instance, Sinclair (1996, p. 174) says that he uses parallel corpora as "indirect evidence" of the knowledge and expertise of multilinguals as expressed in their translations. Parallel corpora, according to Sinclair (1996, p.174) constitute "large repositories of the decisions of professional translators, supplied together with the evidence they had for those decisions". Parallel corpora can therefore be examined to substantiate linguists' strong suspicions that certain words and phrases get translated differently in different contexts. Another advantage, according to Noël (2003, p.759), is that parallel data makes elicitation experiments comparing native speakers to multilinguals a superfluous process in contrastive grammatical studies. This is because translation activity leads the speakers involved to "evaluate meaning relations between expressions without doing so as part of some kind of meta-linguistic philosophical or theoretical reflection, but as a normal kind of linguistic activity" (Noël, 2003, p.759). The result of the translators' evaluations is "manifest in observable relations between texts" and therefore usable within a study of semantic descriptions that may be of interest in a wider linguistic context (Noël, 2003, p.759). In sum, then, using a parallel corpus in contrastive analysis is an excellent means of identifying translation equivalents in source and translated texts, but may also furnish empirical evidence for semantic (and at times pragmatic) claims. Aijmer (2008) observes that sometimes a contrastive analysis *cannot* be based on a parallel corpus if there exists insufficient or no parallel data for the two languages under analysis. In that case, using

comparable corpora is a better option than made-up examples (Aijmer, 2008, p.278). However, Aijmer also says that using a comparable corpus is not the preferred practice for contrastive analysis, because the comparability of the texts in the different languages is not guaranteed. Observing comparable corpora, a researcher will get "a less clear picture of the correspondences of a lexical item or construction than [using] parallel corpora" (Aijmer, 2008, p. 278). Aijmer (2008, p. 279) further says that parallel corpora have the advantage of readily highlighting similarities and differences in the use of lexical items or constructions across languages. This is because "translations are ways of tapping native speaker's intuitions" and translators have the necessary knowledge of both languages to judge which words or constructions will produce meaningful translations (Aijmer, 2008, p. 278). She gives the example of English modal auxiliary verbs largely corresponding with modal adverbs in Swedish.

Aijmer (2008, p. 280) says that it is easier to extract corresponding examples in different languages from aligned parallel corpora (a corpus in which the chunks that directly correspond to one another as source/target language pairings are annotated at paragraph, sentence, or word level) than from comparable corpora. These extracted correspondences for lexical items or constructions then help in establishing translation correspondences and in using them as parameters for translation comparisons. Once corresponding items are extracted from translated texts, the correspondences can be tested by studying monolingual corpora (Aijmer, 2008, p. 278). Another advantage is that using aligned parallel corpora can also help in identifying and establishing contrastive "lexical-semantic fields" (Aijmer, 2008, p.278), that is, groups of lexical items which are a part of a certain collection or network of similar meanings. This can be accomplished by "going back and forth between translations" (Aijmer, 2008, p.278). However, Aijmer goes on to argue, researchers must keep in mind that though words translated from one language into another may have the same value, they may

not be "translation equivalents". That is, although translated items may appear to be "more or less interchangeable in some contexts", they have distinct and separate meanings in different contexts (Simon-Vandenbergen & Aijmer, 2007, p. 247). Simon-Vandenbergen and Aijmer use a parallel corpus to study overlap and differences in the use of translated adverbs. They find that, in different contexts, a single adverb in the original language (e.g. English) can be translated by multiple adverbs in the other language (e.g. Swedish). In their analysis, they construct networks or clusters of translated words and the words they are used to translate in different contexts. For instance, Simon-Vandenbergen and Aijmer (2007) use parallel corpora to identify each English modal adverb of certainty that is translated by various Swedish modal adverbs (e.g. in Swedish *säkert/säkerligen* 'certainly', *definitivt* 'definitely'). They then build a network for each adverb to demonstrate their polysemous nature. For instance, the Swedish adverb *säkert* 'certainly' is most closely associated with English *certainly*. But Simon-Vandenbergen and Aijmer demonstrate how, by building a network, it becomes evident that *säkert* is also strongly related to *no doubt* and is sometimes used to translate *surely* (Simon-Vandenbergen & Aijmer, 2007, pp. 249-50). These networks exemplify the kinds of analyses that only parallel data can support.

On the other hand, those corpus linguists who favour, or at least do not reject, use of comparable corpora in contrastive studies (e.g. Ebeling, 2000; McEnery & Xiao, 2010; Gilquin & Granger, 2010) point out drawbacks in relying exclusively on parallel corpora. The first is the interlanguage properties of translated language. *Interlanguage properties* are defined by Granger (2008, p. 259) as characteristics of a language that tend to be produced by learner or non-nativelike second language speakers of a language but which are either absent from or less common in the usage of native speakers. Interlanguage properties arise from the tendency of translators to simplify the language they translate for clarity (Aijmer, 2008). Baker (1993, p. 249) points out that sometimes translations lead to "unusual distribution of

[linguistic] features" in the target language. By *unusual distribution* Baker (1993, p. 249) means that words or other features that are more frequent in a source language than their equivalents in the target language may be overrepresented by translators in translated texts. For example, Shamaa (1978, pp.168-71) demonstrates that the frequency of common words such as *day* and *say* in English is significantly higher than that of their direct translations in Arabic. In consequence, in translations of Arabic texts into English, these common words are often underused relative to rarer synonyms that directly translate the Arabic words used, changing the frequency distribution relative to native English texts. The English of the translation thus has varietal features specific to its nature *as a translation*. Practices like this result in "the third code" (or *translationese*), defined by Frawley (1984, p. 168) as a variant of the target language which differs from both the source and target language. Moreover, stylistic decisions by translators may be influenced by personal preference, publishers' requirements, or further factors *other than* features of the source text. Such choices may not represent typical use of the target language (Lauridsen, 1996, p.67).

Ebeling (2000, pp.25-26) says that even if we consider a translator's work to be a native speaker's intuition-based use of their own language, certain constraints cannot be overlooked. Due to the translationese phenomenon, translations may become distorted, or (some of) the intended meaning of the original text may be lost. One straightforward piece of evidence for this is the commonplace observation that translations of the same text by different translators are not identical, and indeed, often very far from identical. Each translator works with particular motivations, in particular circumstances (Ebeling, 2000, pp.25-26), and since these are typically not known to the analyst, they are in effect unpredictable.

The same holds true if we consider Malmkjær's (1998) suggestion that, instead of relying on one version of a translation of a text into the target language, we could include as

many versions of the same translated text as possible. McEnery and Xiao consider this suggestion not operationalisable for a contrastive analysis, because different translations of source text make "the comparison of these versions less meaningful" (McEnery & Xiao, 2007, p. 5). The reason is that the texts which could be utilised in this way are "literary works where multiple translations of the same work are available", which are typically "non-contemporary and different versions of translation", spaced decades apart (McEnery & Xiao, 2007, p. 5). In McEnery & Xiao's (2007) view, a comparison of such multiple translated works is more useful for translation studies than contrastive analysis.

Beyond the question of interlanguage properties, a further issue is that parallel corpora are usually small and represent a restricted range of genres (Aijmer, 2008, p.278). One reason for this is that when compiling parallel texts, linguists may consider themselves obliged to exclude texts if it is difficult to assess which is the translated language and which the original (Lauridsen, 1996). This may be the case, for example, with parallel texts of European Union official documents (Aijmer, 2008, p.282), otherwise a plentiful source of multilingual data. Aijmer (2008, p. 282) reasons that parallel corpora may also be restricted because of unavailability of translations of texts that would otherwise be included (as, after all, not everything gets translated), and because of limited availability of bi-directional translations. As an example of the latter, many English non-fiction works are translated into Swedish, but hardly any such texts are translated from Swedish into English (Aijmer, 2008).

To design a parallel corpus, then, a number of issues have to be taken into account, including the availability of text types and of bi-directional translated texts (Aijmer, 2008). Ideally, parallel corpora for contrastive studies would comprise a "large number of text categories" (Aijmer, 2008, p.282). However, non-availability of a sufficient range of texts across genres or registers may lead to either a much smaller or a much less diverse corpus than would be desirable. Using a too-small parallel corpus risks not finding examples of all

possible correspondences for some element, construction or "range of distinctive features of translated language" (McEnery & Xiao, 2010, p.7). Another issue with parallel corpora raised by McEnery and Xiao (2010) is the difference in the original and the translated content of a language that makes the translated texts unsuitable for a contrastive study. They quote other studies (e.g. McEnery & Xiao, 2007; Xiao, 2010) that show that certain features in translated texts are "characterised beyond the lexical level, by normalization, simplification, explication, and sanitization" – features which are common in English and many other languages. But this makes the parallel corpus unrepresentative of the actual target language, and it therefore "cannot serve alone as a reliable basis for contrastive studies" (McEnery & Xiao, 2010, p. 8). The issue of the restricted quantity or range of parallel data can be a major reason to avoid complete reliance on parallel corpora for a contrastive analysis.

Yet another reason to avoid such reliance is that in the process of translation, some linguistic element(s) may not get translated from original language, because sometimes in translations "loss of a specific feature does not entail a 'gain' at different level" which leads to "loss of register profile" (Neumann, 2013, p.282) (see 3.4.2). This process of adaptation in translated texts is due to linguistic and socio-cultural factors inherent in the translation process (Neumann, 2013, p. 307).

The two principle drawbacks, of interlanguage properties and the restricted range of genres available for parallel corpora, alongside related issues also explored in this section, lead many corpus linguists to the conclusion that parallel data is suitable for studying translation norms, but if used for a corpus-based cross-linguistic descriptive analysis, a parallel corpus will often either be too small to capture the full range of possible equivalents for some feature, or else lead to a high risk of inadvertently analysing the third code rather than the target language or the source language or the difference between them. In sum, using

only parallel corpora for contrastive analysis can lead to a "misleading conclusion" (McEnery & Xiao, 2007, p.139).

3.4.4.2 Use of comparable corpora in contrastive studies

A solution to non-availability of a large or unreliable parallel corpus is compilation of "monolingual corpora of different native languages which are created using the same sampling techniques and similar balance and representativeness" (McEnery & Xiao, 2010, p. 8). Gilquin and Granger (2010) say that using comparable corpora in contrastive studies has two advantages. The first is that comparable texts are widely available. There are now a plethora of sources of electronic text available for many languages in addition to English (from such sources as online newspapers).

The second advantage is that comparable corpora represent spontaneous use of language by native speakers. Therefore, these texts are free of the influence of an underlying source text that would induce the production of third code rather than authentic instances of the language. Gilquin and Granger acknowledge that this claim is not true in its entirety because traces of other languages (usually English, whose influence on other languages extends beyond texts translated *from* English) can be seen in some kinds of texts. Yet even so, comparable corpora are "therefore arguably more reliable, especially to assess frequency and patterns of use" (Gilquin & Granger, 2010, p. 6).

However, some problems with exclusive use of comparable corpora in contrastive analysis should be noted. The first drawback is "text type comparability" (Gilquin & Granger, 2010, p.5). The translated texts in a parallel corpus are the same in content as the originals, and therefore the source and target language texts are semantically and pragmatically comparable. On the other hand, in comparable corpora, the match in content

between languages is much less exact. To start with, comparable corpus texts will not be as directly similar as in a parallel corpus because they are not the *same* texts – even if the register and genre selections match precisely across languages, the texts' content necessarily differs merely by virtue of them being *different documents*. Moreover, again even if the registers and genres match exactly, the internal mixture of sub-types, domains, and/or topics within bundles of texts collected for a particular text category in different languages will be less than identical if for no other reason than pure happenstance in the text collection process. More critically, though, across languages the actual register and/or genre categories that exist, and on which the design of the comparable corpora is to be based, may not be the same due to cultural differences between the speaker communities of the two languages (Gilquin & Granger, 2010, p. 5). For example, McEnery and Xiao (2004, p. 1176) had difficulty finding data for the category "Western and adventure fiction" while designing a Chinese corpus according to the sampling frame of Brown, LOB and similar corpora, because there is no such genre of fiction in Chinese. Therefore, McEnery and Xiao replaced this FLOB category with "adventure and martial arts fiction". This illustrates how it is not always possible match registers exactly across languages.

Another drawback of using comparable corpora instead of parallel corpora is that cross-linguistic comparison requires "sameness" (James, 1980, p.169) as a point of departure. This is also referred to as *tertium comparationis*, defined by Ebeling and Ebeling (2020, p. 97) as "an objective background of sameness that ensures that we compare like with like", that is, the "background of sameness against which differences can be viewed and described" (Johansson, 2007, p.39). Otherwise, "how do we know what to compare?" (Johansson, 2007, p.3). Ebeling and Ebeling's (2020) study shows that there are different *tertium comparationis* for the two types of multilingual corpora. For a parallel corpus, the background of sameness is the presence of the precise same documents across languages, so that each individual text

and its translation(s) are exactly "matched by topic, function, meaning and style" (Ebeling & Ebeling, 2020, p. 101) (see 3.2). For a comparable corpus, the corresponding parts for different languages "are usually matched by period, genre and/or domain to enable contrastive analysis" (Ebeling & Ebeling, 2020, p. 100), a very different "background of sameness".

The import of this issue is illustrated in an example discussed by Ebeling and Ebeling. Ebeling and Ebeling (2020, p. 106) investigate two patterns, "*for * sake* in English originals vs. *for * skyld* in the Norwegian originals", via concordances from "the comparable part of the ENPC+". They observe in this data that different meanings are conveyed by these patterns in English and Norwegian despite their formal similarity, *skyld* literally meaning 'sake'. In English, *for * sake* is most commonly used in expletives (e.g. *for God's sake*), whereas in Norwegian *for * skyld* is normally used to express purpose (e.g. *for syns skyld* 'for the sake of sight'), and only secondarily to form expletives (Ebeling & Ebeling, 2020, p. 113). These results show that analysis of bidirectional parallel corpus data can lead to a "deeper understanding of items compared" than analysis using a unidirectional parallel corpus or comparable corpus. A comprehensive contrastive account of the feature(s) under investigation is possible with parallel corpora because elements in the source and target languages correspond directly by virtue of one translating the other (Ebeling & Ebeling, 2020, p. 98). The use of bi-directional parallel corpus helped in identifying "congruent [formally similar] and non-congruent [formally dissimilar] correspondences of patterns" in the data (Ebeling & Ebeling, 2020, p. 109). Follow-on analysis using comparable corpora can be helpful in identifying additional semantic and pragmatic uses of these patterns, which may not be fully represented in a parallel corpus built on a range of texts restricted by issues of availability (Ebeling & Ebeling, 2020, p. 113).

3.4.4.3 A viable solution for use of corpora in contrastive analysis

The foregoing discussion shows that different linguists lean, more or less strongly, towards different verdicts on this issue, some being in favour of parallel corpora and others in favour of comparable corpora. Most contrastive linguists who have worked with both types of multilingual corpora (Aijmer & Altenberg, 2013; Ebeling & Ebeling, 2020; Hasselgård, 2020; McEnery & Xiao, 2010) do not recommend *exclusive* use of either type of multilingual corpus. Instead, they recommend concurrent use of both parallel *and* comparable corpora, because each has its own strengths and weaknesses, such that when they are used together, they complement each other (Aijmer & Altenberg, 2013, p. 2; Johansson, 2007, p.31; p. 182; McEnery & Xiao, 2007, p.139; Granger, 2010, p. 7). For instance, Ebeling and Ebeling (2020, p.113) conclude that because a bidirectional parallel corpus excels in identifying corresponding elements and patterns, such a corpus can appropriately "serve as a starting point for further investigations that draw on much larger comparable monolingual corpora". According to these and other authorities, then, the best practice is (1) to consult parallel data at the initial stage of cross-linguistic analysis, to identify the items that are translation equivalents of one another and thus relevant to investigation, and based on those findings, (2) to use a comparable corpus of the languages under comparison, or monolingual corpora sufficiently similar to be judged comparable, for further contrastive research on their typical behaviour and meanings in non-translational texts where the problem of third code or translationese is not present.

Corpus-based contrastive studies have been an established trend in linguistic research for more than twenty-five years. As one pioneer of corpus-based contrastive analysis has observed, "if used with care and imagination, multilingual corpora lead us beyond what we

know or did not see so clearly. This is the essence of the cross-linguistic perspective" (Johansson, 2012, p.65).

3.5 A framework for analysing MACs in English and Urdu

The purpose of this section is to pull together various points from throughout chapters 2 and 3 to explain which precedents I follow in my corpus-based contrastive analysis of modal adverbs of certainty (MACs). I use a descriptive analytical framework, in that my approach mainly follows that of present-day descriptive English grammars such as Quirk et al. (1985) and Biber et al. (1999) (see 3.3.2). I follow these grammarians not only because they provide grammatical descriptions for English, but also because they are prominent among the proponents of introducing corpus-based methods to descriptive studies. I also draw on Genady (2005) and Schmidt's (1999) Urdu descriptive grammar (see 2.5). To date, there is no descriptive study specifically of Urdu MACs. Therefore, I will apply the general principles of the tradition represented by Quirk et al. (1985), Biber et al. (1999), and Leech (2006) in devising my analytical approach for Urdu as well as English. Thus, the terminology I utilise is mainly based on Leech (2006) for English and Schmidt (1999) for Urdu. The theoretical framework of my thesis is mainly built on Simon-Vandenbergen and Aijmer's (2007) analysis of MACs (see 2.3.3, 2.4.2 and 2.6).

For this project, I follow the example of analysts who have relied on corpus-based methods to contrastively analyse languages, most notably McEnery & Xiao (2010). As my analysis is a contrastive study of MACs in two languages, English and Urdu, I follow those corpus linguists (Aijmer & Altenberg, 2013; Ebeling & Ebeling, 2020; Hasselgård, 2020; McEnery & Xiao, 2010) who recommend use of more than one type of multilingual corpus (see 3.4.4.3). These authors all propose initially using a parallel corpus to identify a

descriptive element in the two languages, and once it is identified, using large monolingual corpora that are representative of the respective languages but also composed of as similar text types as possible (see 3.4.4.2). The use of examples from comparable corpora can help in further and more comprehensive investigation of the use of MACs in the two languages. (see 3.2). Thus the two kinds of corpus complement one another in a comparative analysis.

McEnery and Xiao's (2010) work also demonstrates the benefits of synergy between corpus linguistics and contrastive linguistics in a cross-linguistic descriptive analysis (see also Ebeling & Ebeling, 2020). A contrastive analysis of MACs using comparable monolingual corpora will help me in identifying to what extent English and Urdu MACs exhibit similar or differing semantic and pragmatic functions in the two languages (see 3.4.4.2). This study will be a benchmark in contrastive studies in Urdu because to date, the practice of applying comparable monolingual corpora to analyse a descriptive category has not been applied to Urdu.

3.6 Research questions

My research focuses on four main aspects of MACs: what exactly the MACs in English and Urdu are which correspond to one another; the behaviour that they exhibit in their clause-level placement; their scope and semantic status; and the pragmatic functions they express. My research questions on these issues are in congruence with my research aims (see 1.5.1) and methodological aims (see 1.5.2) discussed in detail in chapter 1.

The research questions are thus:

RQ 1 On the basis of previous literature and corpus investigation, which lexical items and phrases constitute the set of MACs in English and Urdu?

RQ 2 Is the placement of (English and Urdu) MACs in different clauses similar to what previous literature shows?

- In dependent clauses: complement, relative, adverbial?
- In independent clauses: declarative, negative, interrogative, imperative?

RQ 3 In what ways do (English and Urdu) MACs in different clausal positions influence the surrounding elements?

- How does the placement of MACs affect their semantic scope over other clause elements, clauses or sentences?
- What are the modal semantics of MACs at clause level?
- What pragmatic functions do MACs perform in English and Urdu?

RQ 4 To what extent are the behaviours of the Urdu and English MACs (as established by RQs 2 and 3) similar, and in what ways are they distinct?

Answering these research questions will fulfil both my descriptive and methodological aims. They focus on the aforementioned four main aspects of MACs so as to direct a detailed investigation of the similarities and differences between MACs in the two languages. Syntactically, the placement of MACs is operationalised as their flexibility to occur in clause initial, medial and final positions (see 2.3.2.1). Moreover, MACs can occur in different types of clauses, both dependent and independent (see 2.3.2.2, 2.3.2.3). While it is known that positioning of MACs indicates their scope over the other clause elements (see 2.3.3.1), and that the different positions may also indicate which of the MACs' possible modal semantic values and/or discourse and pragmatic functions is being utilised (see 2.3.3.3, 2.4.2), answering the above research questions will allow it to be determined empirically and in detail how this works in natural language data.

Use of the existing literature and of both parallel and comparable corpora to answer my research questions also fulfils my methodological aim (see 1.5.2). In this way, I provide a detailed 'showcase' demonstration of corpus-based contrastive analysis, which is novel to this thesis and which, it is to be hoped, may influence future corpus-based analysis of Urdu grammar to its benefit.

3.7 Chapter summary

In this second and concluding part of the literature review, I have reviewed corpus-based approaches relevant to the kind of contrastive analysis with which this thesis is concerned. I first discussed corpus linguistics and the corpus-based approach in general. Then I reviewed corpus-based descriptive studies from the pre-electronic corpora era; the use of corpora by English reference grammars; and present-day corpus-based descriptive studies. Moving on to corpus-based contrastive studies, I reviewed the nature of this approach in general before considering in detail arguments regarding the use of parallel as opposed to comparable corpora in this area. This literature review, together with that in the previous chapter, inform my thesis's analytical framework and the research questions that it will be deployed to answer (as described across 3.5 and 3.6). The methods to be used to accomplish this are the topic of the next chapter.

4 Methodology

4.1 Chapter overview

This chapter is divided into two main parts. Section 4.2 presents the data used in this thesis; section 4.3 discusses the methods and procedures applied. In 4.2.1-4.2.4, I explain which corpora I use in my analysis and the reason why I have chosen those corpora. I also explain how I present examples from the corpora (in 4.2.6). In section 4.3, I explain the steps taken in the analysis and discussion chapters to answer my RQs. In section 4.4, I answer RQ 1; this answer establishes which English and Urdu MACs will be further analysed to answer RQs 2,3 and 4.

4.2 Data

4.2.1 Parallel corpora

The parallel corpora that I utilise are modified versions of the Urdu-English corpora created by the EMILLE[8] (Enabling Minority Language Engineering) project (Baker et al., 2002). Jawaid and Zeman (2011) modified these corpora by correcting typographic issues in the original Urdu texts, and then aligning the English and Urdu data. *Alignment* creates links between segments of the translated texts (target) and the corresponding segments of the original (source) language corpus. In this case, the English and Urdu texts are linked at sentence level. Alignment of texts is helpful in searching for an item and its translation (see McEnery & Xiao, 2007). Jawaid and Zeman's (2011) modified corpus is a part of their larger

[8] Details of the EMILLE project can be found at http://www.emille.lancs.ac.uk/

UMC005 corpus. UMC005 means *Urdu monolingual corpus* – but this part is bilingual, because it comprises parallel texts in English and Urdu. In addition to the EMILLE corpora, UMC005 includes further parallel texts representing the Bible and Quran in English and Urdu. Jawaid and Zeman (2011) aligned the corpora using the GIZA++ alignment toolkit (Och & Ney, 2000). I was given access to the modified EMILLE parallel corpora by Dr Jawaid in 2017.

4.2.2 Pre-existing comparable corpora

In addition to parallel corpus data, I use comparable corpora. The impetus to use comparable corpora in my study is twofold: to supplement my small parallel corpus, and to follow similar contrastive analysis in the literature. Specifically, I follow McEnery and Xiao (2007, 2010), who illustrate the value of both parallel and comparable corpora in translation studies and contrastive descriptive studies. McEnery and Xiao (2007, p.135) say that a parallel corpus is useful as a starting point for contrastive research but that such corpora "alone serve as a poor basis for cross-linguistic contrast because translations (i.e. L2 texts) cannot avoid the effect of translationese". On the other hand, they advise that carefully chosen comparable corpora can overcome the issue of translationese (defined by Ebeling & Ebeling, 2013, p. 42 as distorted translations in target language because of source language influence) and are a useful resource for a contrastive analysis "when used in combination with parallel corpora" (see also 3.4.3).

In an initial pilot analysis, I extracted English examples from the written part of the British National Corpus 1994 (XML Edition; BNC 1994) and Urdu examples from the Charles University Urdu Monolingual Corpus 2014 (CUUMC). I accessed both BNC1994

and CUUMC via Lancaster University's CQPweb server[9]. Both are monolingual, non-specialised corpora, that is, they incorporate a range of fiction and non-fiction sources and are not restricted to any specific genre. I therefore presumed them sufficiently comparable for my pilot, but evaluated their suitability subsequent to this exploratory analysis.

4.2.3 The need to develop new comparable corpus

The CUUMC data used in the pilot analysis is derived from diverse web sources. However, the compilers have split the data into sentences or clauses and then scrambled the order of these units throughout the whole corpus. Their reason for doing this is that the texts were downloaded from the internet without copyright clearance. The compilers obscure the data via this scrambling to avoid any intellectual property rights violation and ethical issues when giving others access to the corpus. An issue during descriptive analysis thus arises when the split is at a clause level, or the searched item needs to be looked at in a context beyond a clause. For example, in CUUMC, there are six instances of PSNC *ġālibān nahīṁ* 'perhaps not' where it occurs as a phrase on its own, as in (56). That is, there is nothing before or after this phrase because of split data. .

56. ġālibān nahīṁ yaqīnān.
perhaps not certainly.

"Not perhaps, certainly" (CUUMU_15 18997).

Example (56) seems to be a reply to a preceding sentence. By using *nahīṁ* 'not', the speaker actually negates *ġālibān* 'probably' to claim that what the previous speaker expressed as a possibility is in fact a certainty. However, this is my intuitive interpretation. Due to the clause-level scrambling, the preceding text is not linked, so there is no evidence for what the

[9] https://cqpweb.lancs.ac.uk/

speaker is replying to. Similar examples (of short replies by the addressee) exist in the English data, but as the BNC 1994 is not scrambled, I could observe the context in which such examples occur.

Considered as English-Urdu comparable corpora, CUUMC and BNC 1994 satisfy the criterion of size, that is both corpora are vast[10], and there is an overlap of domains (such as news, blogs). However, their time periods are not the same. Moreover, the BNC 1994 has well-defined, organised, and indexed categories, whereas the CUUMC compilers did not index the text into categories/domains, in order to anonymise the data "for academic research purposes only" (Jawaid et al., 2014, p.2939). As Egbert et al. (2022, p.13) point out, the "design, compilation, and use of a corpus for different kinds of linguistic analysis" is a concern for any researcher using an existing corpus. This is because to analyse it linguistically, the researcher needs to understand whether that corpus actually represents the "range of texts or text types" in the relevant domain of language use, and thus whether analysis of that corpus addresses their research questions sufficiently (Egbert et al., 2022, p. 12).

On the basis of the above-mentioned issues, the concern is raised that even though there is some degree of similarity in the text type composition of the BNC 1994 and CUUMC, the differences between them will make the validity of the comparative analysis uncertain and indeterminable. Therefore, I decided not to continue using these corpora beyond my pilot explorations, but rather to compile comparable corpora of English and Urdu from accessible sources on the web. This approach, detailed in the next section, not only

[10] The written part of BNC 1994 is approximately 90 million words in size; the CUUMC is approximately 95 million words in size.

satisfies all parameters for compiling comparable corpora, but also, for Urdu, allows me to evaluate concordance examples in context beyond the clause.

4.2.4 Novel English-Urdu comparable corpora

4.2.4.1 Compilation parameters for the English-Urdu comparable corpora

As mentioned above , my priority was to compile comparable corpora that are alike according to as many parameters as possible to fulfil the condition of *sameness* or *tertium comparationis* (TC) (see 3.4.4). For this purpose, I looked at the available web sources and settled on two kinds of source – news websites and chat forums. I included chat forums because it is the type of writing on the web likely to be most similar to informal conversation Initially, I wanted to include novels and short stories to diversify the register coverage. I identified many suitable sources of such texts. However, all Urdu novels and short stories that I found online turned out to be stored as scanned images of the original print versions (in PDF files). Some experimentation showed that the available optical character recognition (OCR) apps either do not convert scanned Urdu text into machine readable text at all, or are still not well-developed to produce a quality converted text. Thus the inclusion of fiction had to be abandoned.

Having decided on the source categories, I looked for Urdu news websites. I selected three: Daily Express News[11], Jang[12], and NawaiWaqt[13]. Correspondingly, I selected three UK-based English news websites: The Daily Mail[14], The Guardian[15], and The Independent[16].

[11] https://www.express.pk/
[12] https://jang.com.pk/
[13] ttps://www.nawaiwaqt.com.pk
[14] https://www.dailymail.co.uk/
[15] https://www.theguardian.com/
[16] https://www.independent.co.uk/

Based on these sites' structures, I categorised the texts I would collect across five divisions: news, editorials, opinions, features, and blogs. During collection of the Urdu data, I realised that after a certain period, varying from three days to two months according to the newspaper, the actual texts of stories are removed from the websites, leaving only scanned images of the newspapers. Therefore, for most categories my data collection covers September to November of 2020. For UK newspapers, there is no such constraint on dates because they preserve complete text archives but for the sake of comparability I sampled from same dates (i.e. September to November of 2020).

I searched for publicly accessible chat forums from where I could retrieve the data without any restriction. For Urdu, I identified a site called *Urdu Mehfil Forum*[17]. I selected certain topics within this forum (e.g. education and teaching) that have considerable discussion and retrieved all threads under those topics. For English, I opted to use data from Reddit, but targeting UK specific forums, to standardise on British English. Initially, I collected data from three such British-centric pages. But this generated massively more data than I had obtained for Urdu. Therefore, I limited the use of Reddit data to a single forum, AskUK[18]. AskUK appears to encompass similar discussion topics to Urdu Mehfil Forum. Because forum threads may contain posts from across wide time periods, I did not take into account their dates when collecting the threads as texts.

Because these forums are online social media platforms that anyone can join, their users may come from various geographical locations. They cannot be guaranteed to be of UK or Pakistani origin, although the source were selected in the expectation that most users would be. The resulting potential for a mixture of dialects is not a major problem for my research, however, because the purpose of this research is to describe how MACs are used

[17] https://www.urduweb.org/mehfil/
[18] https://www.reddit.com/r/AskUK

cross-linguistically based on corpus evidence and not a comparison of socio-cultural norms reflected in these languages. To respect privacy, I anonymise any names in the Urdu data, by replacing them with Roman letters (e.g. X, Y, Z) in the examples cited in this thesis. This is not necessary for Reddit because its usernames are not linked to real-world identities.

By these means, I have compiled two corpora matched in genre in English and Urdu. For the news data, there is sameness in time span, and domains; for the chat data, there is sameness in content in that the forums selected were generic rather than addressed to specialised topics.

4.2.4.2 Automated downloading of texts

Both English and Urdu data were downloaded from the source sites using web scraping techniques. Rather than a general scraping tool, I used specially created scripts in PHP and Python[19].

Once the data was compiled, it was automatically annotated with POS tags[20]. The corpora were then indexed and made accessible for analysis on Lancaster University's CQPweb server[21]. The Urdu corpus has open access for all, but the English corpus has restricted access[22]. The English corpus is named the *English Comparable Corpus* (ECC) and the Urdu corpus is named the *Lancaster Urdu Web Corpus* (LUWC).

[19] I am indebted to Dr Andrew Hardie for his immense help in formulating the scripts and his constant monitoring while I compiled the data.
[20] For Urdu POS tagging, I am grateful to Dr Sarmad Hussain's team at the Centre for Language Engineering-University of Engineering and Technology (CLE-UET) in Lahore, Pakistan, especially Asad Mustafa and Ehsan ul Haq.
[21] I owe thanks to Dr Andrew Hardie for conducting the whole lengthy and technical process of cleaning, removing duplicates, indexing the raw data, and finally providing space on cqpweb.lancs.ac.uk for Urdu and English comparable corpora.
[22] CQPweb has now the feature *install your own corpus*, but the English corpus is too large to be self-installed. Therefore, it was centrally installed but with access restricted to myself and CQPweb administrators.

4.2.5 Descriptive statistics on the corpora

The sizes and composition of the corpora are detailed in Tables 4.1, 4.2 and 4.3.

Table 4.1: Statistics on the English and Urdu corpora

	Size	
Parallel corpora	English tokens	Urdu tokens
EMILLE (UMC005)	136,073	196,921
Comparable corpora		
English monolingual corpus (ECC)		97,237,115
Urdu monolingual corpus (LUWC)	-	24,033,770

Table 4.2: Composition of the English and Urdu corpora

Corpus	**Corpus sources**
Parallel English-Urdu (EMILLE)	Written text, formal: UK Government documents[23] (original English translated to Urdu)
Monolingual English (ECC)	Written text (web sources), formal and informal: news website texts classified as editorials, opinion articles, features (including blogs and articles); chat forum
Monolingual Urdu (LUWC)	Written text (web sources) formal and informal: news website texts classified as editorials, opinion articles, features (including blogs and articles), chat forum.

Table 4.3: Text types in the English and Urdu comparable corpora

Text type	**Tokens in ECC**	**Tokens in LUWC**
Blogs	206,365	67,285
Editorials	41,189	3,648,051
Features	29,221,622	595,418
News	50,083,775	611,635
Opinion	3,025,359	1,591,502
Online chat	14,058,805	17,519,879
Total	**97,237,115**	**24,033,770**

[23] "Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions" (Baker et al., 2002, p.76).

4.2.6 Rationale for comparability of the corpora

The corpora discussed above all represent present day English and Urdu in the written form, but are dissimilar in terms of the type of communication they represent. The parallel corpus is composed of UK government leaflets, whereas the comparable corpora have more diverse sources (see Table 4.2). By contrast the English and Urdu comparable corpora are similar in that they represent a wide range of online chat media and news sub-genres including same sources (i.e. editorials, feature articles, news, opinion articles). As the two types of corpora are different in their composition, their compatibility for the analysis may be questioned. In fact, the answer is straightforward, because of the difference in the analyses for which each type of corpus will be used. The analyses are interlinked, but the parallel corpus is used as a starting point only, whereas the comparable corpora are used to conduct a contrastive analysis (see McEnery & Xiao, 2007; 2010).

The use of the parallel corpus is limited to (a) identifying occurrences of English MACs (as listed in grammars); and (b) identifying which particular items in Urdu have been used to translate those instances of English MACs. Therefore, it does not matter, for instance, whether, say, the frequencies of the MAC semantic categories in the parallel corpus are skewed from more general usage. Moreover, this is not an exercise in Translation Studies (for example, analysing the stylistic choices of translators), for which the balance of the parallel corpora would be a concern. Once the MACs are identified in the parallel data, I will use the comparable corpora for further analysis: to compare and contrast the frequencies of occurrence, functions and usage of various MACs.

4.2.7 Presentation of corpus examples

Examples in the remainder of this thesis are drawn mainly from the comparable corpora, with only a few from the parallel corpus. Every example is enclosed in quotation marks, and followed by a corpus reference. In the parallel corpora, these have the form *123_En_EMILLE*, where the first portion is a line number, and the middle portion indicates English (En) or Urdu (Ur). For ECC and LUWC examples, source references have the form *ECC_ABC* or *LUWC_ABC*, where *ABC* is the text ID for the source text in ECC/LUWC. Text IDs derive from (elements of) the original URLs. The occasional examples from the BNC 1994 have corpus references in this format as well.

I present English examples as they appear in the ECC. I present Urdu examples from LUWC with accompanying interlinear morpheme-by-morpheme glosses, composed according to the Leipzig Glossing Rules (Comrie et al., 2008). The purpose of these glosses is to "give information about the meanings and grammatical properties of individual words and parts of words" (Comrie et al., 2008, p.1). Each example of this kind has three lines. Line 1 is Romanised Urdu, which I derive from the original text by applying the transliteration rules of ISO 15919:2001, ISO Romanisation for Devanagari, adapted very slightly for Perso-Arabic script (see Appendix A). Line 2 contains the morpheme-by-morpheme gloss for the example. The list of grammatical abbreviations used in Urdu examples are listed on page (xx). Line 3 contains a free translation of the complete Urdu phrase or sentence. Examples from the Urdu parallel data use the same format, except that line 3 is omitted as superfluous (since the corresponding sentence in the English parallel data is always supplied). To respect privacy, I anonymise any names in the Urdu data, by replacing them with Roman letters (e.g. X, Y, Z) in the examples cited in this thesis. This is not necessary for Reddit because its usernames are not linked to real-world identities.

4.3 Procedure of analysis

4.3.1 Introduction to procedural steps

My contrastive analysis of English and Urdu MACs is more oriented towards description of the data than to any one specific theoretical framework (see 2.6). My analytical procedure for description of MACs relies mainly on Biber et al. (1999) and Boye (2012) for syntactic analysis (see 2.3.2.1); on Van der Auwera et al. (2005), Pic and Furmaniak (2012), Simon-Vandenbergen and Aijmer (2007), and Suzuki and Fujiwara (2017) for semantic analysis (see 2.3.3) and pragmatic analysis (see 2.4.2). My semantic and pragmatic description also utilises Boye's (2012) notion of *socio-cognitive communication* (see 2.4.1). For my corpus-based contrastive descriptive analysis of English and Urdu MACs, I rely on McEnery and Xiao's (2010) examples of contrastive grammatical description of two languages (see 3.5). While McEnery and Xiao (2010) analyse genetically unrelated languages (English and Chinese), and I analyse (distantly) genetically related languages (English and Urdu), this does not affect the applicability of procedures.

I will first identify English MACs and the items by which they are translated in Urdu using the parallel corpora. From among the translations found, I will identify and separate out the Urdu MACs. I will then extract the thus identified English and Urdu MACs, calculating their frequencies of occurrence in the comparable corpora. This data will be presented, together with the distribution of these items across various domains, in sections 4.4-4.6 of this chapter, thus answering my RQ 1.

I will then calculate the frequencies of MACs in different types of clause (independent and dependent). I will also break down the percentages with which various MACs occur in different clause positions (initial, medial, and final). This data will inform an

understanding of the preferred positions exhibited by the different MACs; the frequency of a linguistic item definitionally reflects linguistic choices made by speakers or writers, such that comparative frequency sheds light on preference. These quantitative observations will constitute the basic empirical evidence for a subsequent qualitative analysis that aims to characterise meaning and function. The combination of these analyses, presented in Chapter 5, will answer my RQ 2.

Based on concordance analysis of English and Urdu MACs across different clausal positions, I will answer RQ 3 parts 1 and 2 in Chapter 6. Finally, in Chapter 7, I will contrastively evaluate the behaviour of English and Urdu MACs, as shown collectively by my analyses, in order to establish the extent to which English and Urdu MACs are similar and different in use. This will answer RQ 4.

In the remainder of section 4.3, I explore in depth all the procedural steps of the analyses just outlined.

4.3.2 Step 1: Identifying English and Urdu MACs using the parallel corpus

The first procedural step is to identify those lexical items and phrases that constitute the set of MACs in English and Urdu. First, I will locate English MACs in the parallel corpus by searching for all items listed as MACs in the reference grammars and other studies that I reviewed in Chapter 2. Moreover, I will place these lexical items into semantic categories. In this initial step, I identify and note down any Urdu translation of an English MAC as an equivalent element (see Table 4.9).

I will search for each English MAC in the parallel data, one by one, using the advanced text editor Notepad++[24]. For each English example, I will then locate the corresponding sentence in the parallel Urdu text. The parallel corpus is aligned at sentence level (see 4.2.1), as indicated by line breaks. Thus, line numbers can be used to identify corresponding stretches of text, as illustrated as in Figures 4.1 and 4.2.

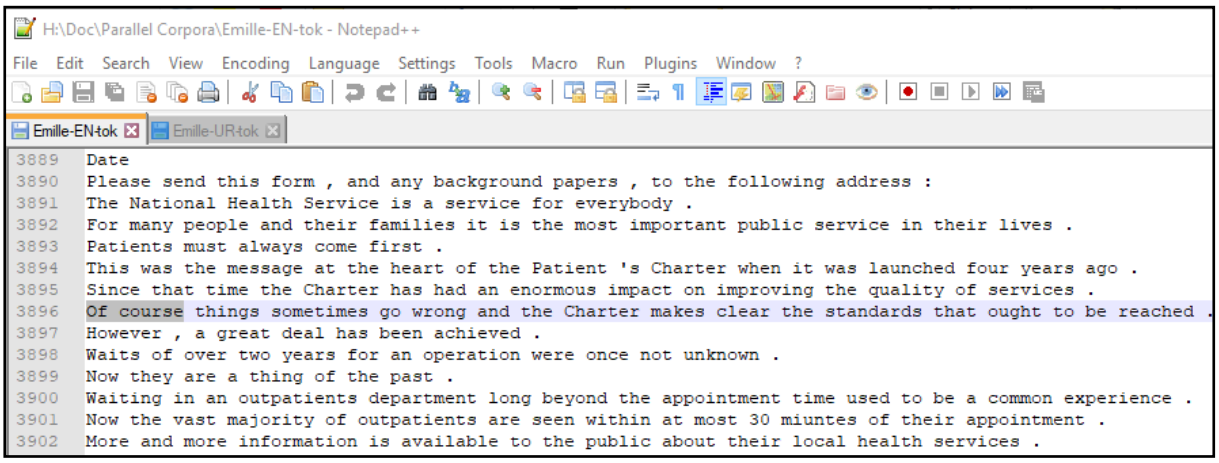


Figure 4.1: An English MAC in the parallel corpus, shown in Notepad++

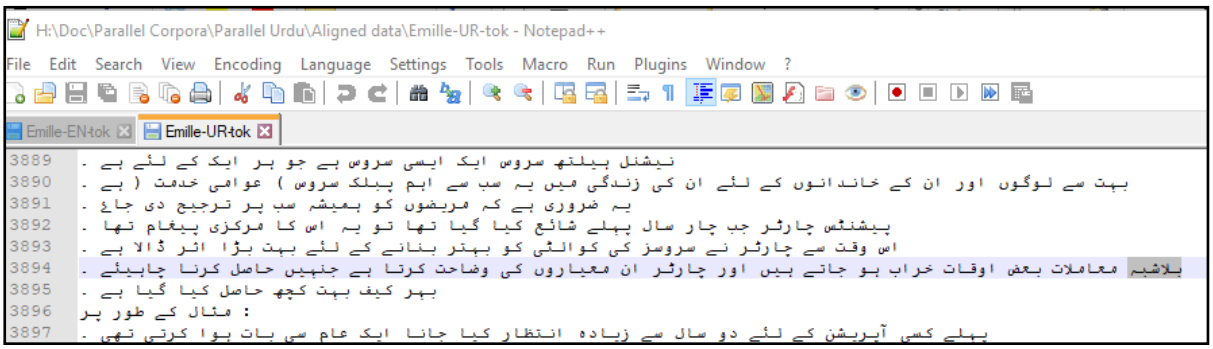


Figure 4.2: Urdu text corresponding to the English example in Figure 4.1, shown in Notepad++

In the parallel Urdu sentence, I will identify the part that corresponds to (is a translation of) the English MAC. Figures 4.1 and 4.2 exemplify *of course* (English MAC) being translated as *bilāśubha* (Urdu MAC). As in English, different classes of lexical items in

[24] https://notepad-plus-plus.org/

Urdu may convey modality (see 2.5). Consequently, in many instances, the English MAC is not translated as an Urdu MAC, but rather as some other part of speech, such as a noun (e.g. *mūmkīn* 'possibility') or as a phrase (e.g. *yaqīnī tor par* 'certainly'). In other instances, the English MAC is not translated by any single identifiable item. I will note down the thus identified Urdu word or phrase, along with the line where it occurs. I will repeat the above process for the whole list of English MACs. Then, I will separately list the Urdu MACs and non-adverb translations along with the English MACs that they translate, and their frequencies normalised per one hundred thousand words. These instances will be recorded manually in an Excel spreadsheet and then later presented in tabulated form in the analysis.

In tabulations of my results, I will group the lexical items into semantic categories. As noted in 2.6, I follow Boye (2012), who places adverbs that express certainty or uncertainty on a continuum rather than sharply distinct categories, but still groups them according to position on the continuum (see 2.4.2.3). The groups are: *high certainty support* (HCS MACs); *probability support* (PS MACs); *probability support for negative content* (PSNC MACs); and *high certainty support for negative content* (HCSNC MACs) (see 2.3.3.3 .

4.3.3 Step 2: Calculating frequencies of MACs in the comparable corpora

I will present relative frequencies (per hundred thousand words) in the monolingual comparable corpora of each of the English and Urdu MACs identified in the prior step, extracted using CQPweb queries. It is pertinent here that certain words have two spellings in Urdu (e.g. *śāyad* and *śāid* 'probably'), where although one is the standard form, people consistently use both. In such cases, I will combine the frequencies of both spellings.

4.3.4 Step 3: Examining distribution of MACs across corpora

Next, I will extract and present the distribution of these items across the various text categories within the comparable corpora. The purpose of examining these distributions is to gain insight into the frequency of English and Urdu MACs in particular kinds of text. I will use CQPweb's *distribution* function to generate this data (see Figure 4.3).

Distribution of hits for query "بے شک" returned 846 matches in 426 different texts
Currently displaying distribution across text classifications.

Display distribution of:	Text categories	Show as:	Distribution table
Cross-tabulating against:		Other actions:	Choose action...

Based on classification: *Source website*

Category [↓]	Words in category	Hits in category	Dispersion (no. texts with 1+ hits)	Frequency [↓] per million words in category
Daily Express	4,043,249	43	40 out of 6,242	10.64
Urduweb	17,519,879	763	354 out of 8,232	43.55
Daily Jang	2,284,364	28	25 out of 3,132	12.26
Nawa-e-waqt	186,278	12	7 out of 496	64.42
Total:	24,033,770	846	426 out of 18,102	35.20

Based on classification: *Text type*

Category [↓]	Words in category	Hits in category	Dispersion (no. texts with 1+ hits)	Frequency [↓] per million words in category
Online chat	17,519,879	763	354 out of 8,232	43.55
Editorials	3,648,051	37	32 out of 5,366	10.14
Features	595,418	15	13 out of 424	25.19
Opinion	1,591,502	31	27 out of 1,714	19.48
Total:	23,354,850	846	426 out of 15,736	36.22

Figure 4.3: CQPweb distribution for Urdu MAC *beśak* across various text categories in LUWC

Of the two distribution analyses shown in Figure 4.3, "text type" is the relevant one for present purposes. I will run same process for each English and Urdu MAC. Graphs of this data will illustrate how the frequency of MACs differs across various text domains in each language.

4.3.5 Step 4: Examining the placement of English and Urdu MACs in clauses

4.3.5.1 English and Urdu MACs clausal positions to be investigated

I will calculate the frequencies with which each English or Urdu MAC occurs in initial, medial, and final clause positions in independent and dependent clauses, following Boye (2012; 2016). These quantitative measurements of the preferred positions of MACs will eventually help in analysing the semantic and pragmatic functions expressed by each MAC in different syntactic contexts.

Negative clauses in English and Urdu are marked by negators: *not* or *no* and *nahīṁ* or *nah* respectively. Boye (2012) categorises a MAC immediately followed by a negator (e.g. *certainly not*) as forming a single, phrasal MAC, one that expresses negative epistemic support (see 2.4.2.3; Figure 2.1). I follow Boye (2012) in considering MAC + NEG constructions as negative counterparts to the MAC in question considered on its own (see 2.4.2.4), and thus as MACs themselves.

In the following sub-sections, I explain each step I take in the calculation of frequency of occurrence of each English and Urdu MAC in initial, medial, and final positions (see 2.4.1 for the definitions of initial, medial, and final positions that I employ here) in independent and dependent clauses. Prior to explaining the procedure, I define the terms that are relevant to this part of the methodology, and note the software tools to be utilised.

4.3.5.2 Terminology for corpus analyses

Three terms, *node*, *collocate*, and *collocation*, are recurrent terminology in the study of words used in combination. McEnery and Hardie (2012, p. 123) define a *collocation* as "a co-occurrence pattern that exists between two items that frequently occur *in proximity* of one another – but not necessarily adjacently or, indeed, in any fixed order". These co-occurrence patterns are not random; rather they are frequently recurring sequences or combinations of words (Harris, 2006, p. 126).

Stubbs (2009, p. 29) defines the *node* as a lexical item or word that is being examined. In this study, all English and Urdu MACs are treated as *nodes* which I investigate: first for syntactic positioning, subsequently for usage and functions. In many cases, I also look at the *collocates* of the node word. A *collocate* is a lexical item that frequently cooccurs with a given node in the corpus being investigated. The node and collocate pair constitutes a collocation. My methodology involves examining specific collocates of MACs (including, for instance, clause markers), but not for the usual reason of studying their phraseological behaviour. Rather, I use these collocates a part of the process of identifying what clause position each MAC example occurs in, as will be explained in 4.3.5.4.

Collocates are determined within specific *spans*. A span is "the number of lexical items on each side of a node that we consider relevant to that node" (Sinclair, 2004, p. 304). A span, then, is the maximum distance between the node and collocate, stated as a number of word tokens on either side of the node. Seretan and Wehrli (2007, p.75) note that in computational linguistics, it is common to consider a span of +/-5 words around the node, in contrast to the span of +/-4 words adopted by most corpus linguists working on English following Sinclair et al. (2004, p. 5) and Stubbs (2002, p. 29). I adopt the former practice here. I use one-way spans (just + or just -) in cases where I seek a particular collocate only on

one side of the node. This is because, during my initial inquiry into my corpora, I found that using a five-token span retrieved the most possible instances of a particular collocate of a given word within the syntactic structures being sought in both English and Urdu. That is, the larger span prevents any instances of a given MAC being missed or mis-classified.

4.3.5.3 Software used in this research

I will mainly use CQPweb, a web-based concordancer, for my analysis. I utilise Lancaster University's CQPweb server[25], which hosts a large number of corpora in a range of languages. Because of their huge size (see Table 4.1), my comparable corpora were installed for analysis on CQPweb. CQPweb includes sophisticated corpus analysis tools (Hardie, 2012); I will use it to extract and examine concordances and collocations, and to calculate frequencies of items under analysis. I will also use Microsoft Excel to record clausal positions of English and Urdu MACs as observed in concordances.

CQPweb offers a choice of three query modes for use in generating concordances: simple query (ignore case), simple query (case-sensitive), and CQP syntax query. A simple query uses a query language designed to be easy for non-specialists. It makes accessible the most commonly used features of CQP via simple codes (e.g.? * + - {}), without the complexity of CQP syntax. Notable among simple query codes is the *wildcard* asterisk (*), which matches anything; this can yield different possible patterns occurring in a sequence (e.g. *certainly * that* allows any number of words, including zero, between *certainly* and *that*). Case sensitive simple queries treat upper and lowercase forms of one letter as distinct (so a search for *Word* will not match any instances of *word*), whereas simple queries that ignore case treat upper- and lower-case letters as equivalent. CQP syntax[26] is a language

[25] https://cqpweb.lancs.ac.uk/
[26] https://cwb.sourceforge.io/files/CQP_Manual/

without any of the simplifications of the simple query language. This makes writing queries a considerably more specialised task: all patterns are regular expressions, and searches can refer to *any* of the data in the corpus index, not just the parts made easy to access in simple queries.

CQPweb possesses a *collocations* tool, which can be applied to the results of any query. It controls are shown in Figure 4.4. Particularly, it allows a single collocate to be specified for retrieval, a feature which I use to look for specific lexical items, including clause markers, collocating with a MAC as node, as shown in Figure 4.4.

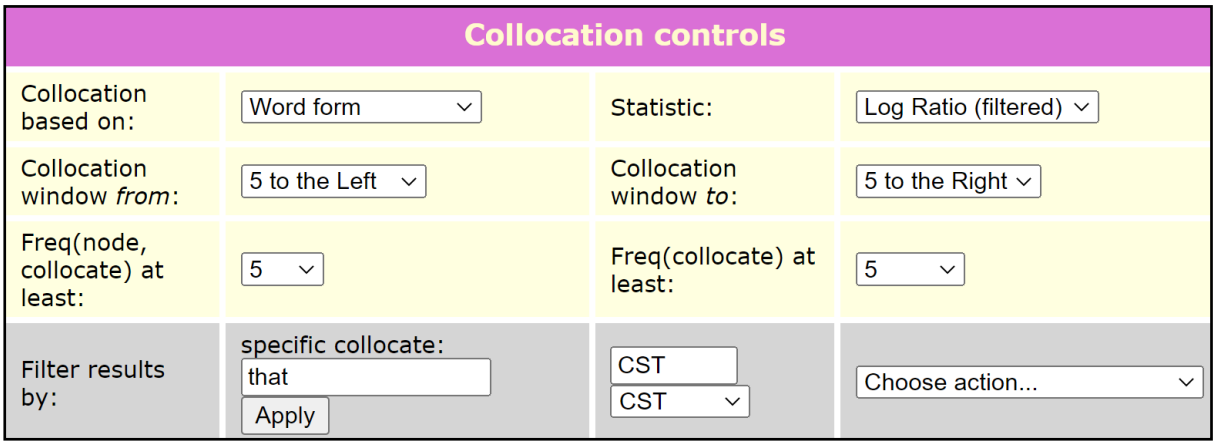


Figure 4.4: Control panel of the collocation tool in CQPweb

Figure 4.4 shows how a specific word (here clause marker *that*) can be specified as the collocate. The collocation tool has number of statistical measurement options, including MI3 and Log Likelihood, but defaulting to *Log Ratio*. According to CQPweb's online help text, the Log Ratio measures "how big the difference is between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus". Log Ratio (*filtered*), the version of Log Ratio used in the collocation tool, requires that collocates

must pass a Log Likelihood test to appear on the list; Log Likelihood measures the amount of evidence for a collocational link.

4.3.5.4 Distinguishing MACs that occur in independent and dependent clauses

I will first generate concordances for each English and Urdu MAC in ECC and LUWC. I will download the concordance lines and manually check (with assistance from automated searches, as explained below) each example of a MAC to determine whether it occurs in clause initial, medial, or final position, and whether it occurs in an independent or dependent clause. By calculating the frequency with which the English and Urdu MACs occur in these three positions, across the two major clause types, I will be able to observe the preferred position of each MAC. Following existing literature (see 2.4.2.3), I include both declarative and interrogative clauses, and both affirmative and negative clauses; likewise, I include all of complement, adverbial, and relative clauses, in my search for MACs in independent and dependent clauses.

As previously noted, I could use three tools in CQPweb to carry out this part of analysis: the standard query tool, in either simple query or CQP syntax mode, or the collocations tool. One approach to creating a more specific concordance would be to include in the query more conditions on words around the MAC (using CQP syntax where necessary). However, I found it more straightforward to use the collocations tool for this purpose. Using the collocations tool helps me to identify syntactic positions efficiently for all straightforward examples (necessary due to the size of the complete concordances). These automated searches help in later manual classification of MACs. Therefore, I use CQPweb's collocations tool to automatically generate smaller, more specific query results that group together examples likely to exhibit the same clausal positioning/clause type. Once identified, I mark the examples found by each specific query as having the position/clause type in

question in my main concordance data. Initial experimentation showed that this process did not classify every example with 100% accuracy, and some examples were found within multiple specific queries. However, the automated queries made the subsequent process of manual checking for classification much less time-consuming. The specific queries I used tried to identify those MACs that occur shortly before (clause final) or shortly after (clause initial) a word which marks a clause boundary. This was done by looking for those clause markers as collocates of the main query for a given MAC, and extracting the examples alongside the clause markers as separate queries. The type of clause marker used as collocate was used to infer dependent versus independent clause.

The question arises: how does this work? Rarely does an independent clause occur as the single clause of a simple sentence. Rather we see them alongside other independent clauses linked by coordinating conjunctions (e.g. *but*) to form compound sentences (Leech, 2006, p.24), as in (57). Moreover, independent clauses occur alongside dependent clauses in complex sentences (Leech, 2006, p. 23). Linked independent and dependent clauses may occur in either order, connected by a variety of types of element, as illustrated in (58) and (59), in which the clausal links as well as nearby MACs are highlighted.

57. "I couldn't be certain how much painkilling medication she had taken, *but* it was almost *certainly not* enough" (BNC_ABS 2017).

58. "Her endorsement potential is *definitely* suffering *because* of her Japanese face and her Japanese name" (BNC_AHU 1160).

59. "Although these ensembles are undoubtedly very varied, it is *perhaps* the professional mixed choirs and the vocal consorts, usually performing a cappella, who dominate the Europeans' sense that there is a distinctively English force at work in the early music revival" (BNC_J1A 1441).

Because of this wide variety of sentence configurations, I search for MACs occurring to both the left and right of each selected clause marker to ensure that I cover all instances of MACs in all types of independent and dependent clauses. Tables 4.4 and 4.5 present the clause markers that I search for as collocates of English MACs to identify specific instances as occurring within independent or dependent clauses.

Table 4.4: Subordinate markers retrieved as collocates of MACs in English

	Dependent clause type	Subordinate markers
1.	Complementiser *that*-clauses	*that*
2.	Relative clauses	*when, whose, who, which, whom, that*
3.	Adverbial clauses:	*although, though, even though, because, for, since, as, so that, whereas, nevertheless, while*

Table 4.5: Coordinate markers retrieved as collocates of MACs in English

	Coordinate markers
1.	yet, but
2.	So
3.	And
4.	or

In Tables 4.6 and 4.7, I present the subordinate and co-ordinate markers that I search for as collocates of Urdu MACs to identify specific instances as occurring within independent or dependent clauses.

Table 4.6: Subordinate markers retrieved as collocates of MACs in Urdu

	Dependent clause type	Subordinate markers
1.	Complementiser *that*-clauses	*ke* 'that'
2.	Relative clauses	*jō* 'that/which/who', Oblique forms of *jō*: *jin.*OBL.3.PL */jinne.*OBL.3.PL 'who', *jinhōṁ.*OBL.3.PL*/inhōṁ.*OBL.3.PL */unhōṁ.*OBL.3.PL 'who', jis.OBL.3.SG */jisse.*OBL.3.SG 'who
3.	Adverbial clauses	*agarce* 'although', *hālāṁke* 'even though', *kīūṁke* 'because', *cūṁke* 'since', *cunāṁce* 'therefore', *tā keh* 'so that', *is liē* 'therefore', *jabke* 'whereas'

Table 4.7: Coordinate markers retrieved as collocates of MACs in Urdu

	Coordinate markers
1.	*magar/lekin/par* 'but'
2.	*tō* 'so'/'then'; *tō phir* 'so then'
3.	*aur* 'and'
4.	*yā* 'or'

My purpose here is not to generate or analyse collocate lists for the MACs, but rather, as explained, simply to identify instances of MACs cooccurring with certain clause markers, and by that process to deduce whether those MACs occur in clause initial, medial, or final position. Since the statistical association between MAC and collocating clause marker is not any part of my analysis, I retain the default settings of the CQPweb collocations tool for statistical measurement (i.e. Log Ratio filtered).

However, vicinity to an explicit clause marker is not always directly indicative of clause initial or clause final position. For this reason, I also generate concordances of specific node-collocate pairs, which I can examine manually to determine each example's position. For each MAC, I generate concordances for the following:

i. Coordinate clause markers to the left of the MAC

ii. Coordinate clause markers to the right of the MAC

iii. Subordinate clause markers to the left of the MAC

iv. Subordinate clause markers to the right of the MAC

I use these concordances to classify the main concordance, but then manually check the classification to ensure its accuracy. Then I calculate the frequencies with which the MAC occurs in the various clausal positions from the checked full concordance.

To exemplify my procedure, I will describe in detail the process for one sub-type of complement clause, i.e. *that*-clauses. Each type of dependent clause is usually marked by some grammatical word; *that*-clauses are marked by complementiser *that* (Leech, 2006, p.17). Similarly, in Urdu, complementiser *ke* 'that' marks complement clauses. In some cases, including identification of complement clauses, the corpus annotation is of use, and is usable in CQPweb. For example, *that* may be tagged for part of speech either as determiner (tag DD1) or complementiser (tag CST)[27]. Only examples of *that* as complementiser are relevant for determining the clause position of a nearby MAC. Hence, when identifying examples of a given MAC nearby to *that*, I search only for *that* as CST.

To execute this step, I search for collocations between *that* (as complementiser) and the MAC under analysis, co-occurring within +/- 5 tokens. CQPweb can generate a breakdown of the relative positions of examples of the collocate, as shown in Figure 4.5.

[27] https://ucrel.lancs.ac.uk/claws6tags.html

Collocation information for node of query "certainly" collocating with *that* with tag restriction *CST* (880,711 occurrences of *that* in the whole corpus)			
Statistic	**Value (for window span -5 to 5)**		
Conservative LR	-0.508		
Within the window -5 to 5, *that* with tag restriction *CST* occurs 638 times as collocate in 513 different texts (expected frequency: 839.4)			
Distance	**No. of occurrences**	**In no. of texts**	**Percent**
-5	44	42	6.9%
-4	42	40	6.6%
-3	56	50	8.8%
-2	65	58	10.2%
-1	29	26	4.5%
1	1	1	0.2%
2	145	120	22.7%
3	86	81	13.5%
4	81	78	12.7%
5	89	82	13.9%

Figure 4.5: Breakdown of relative positions for *certainly* collocating with *that* within a span of +/- 5

Using the links in the position breakdown, I generate concordances for each position from -5 to +5, as exemplified in Figure 4.6.

Solution 1 to 29	Page 1 / 1	
more potent substances if that 's all they can get ...	**that**	certainly increases risks of harm . " Biondo says th
to anger and a little more willing to forgive , and	**that**	certainly applies to me . I am a Christian work-in-p
probably wants to tighten up some of the 404(b) issues	**that**	certainly are ripe for tightening , " said Brennan , w
da Cidade Online hide people , businesses and entities	**that**	certainly act in their desired attempt to install left-v
da Cidade Online hide people , businesses and entities	**that**	certainly act in their desired attempt to install left-v
ever really going to be the new Chelsea unleashed , and	**that**	certainly was n't the case in terms of approach . It v
The shift means more empty calories and high-fat diets	**that**	certainly pack on the pounds , but do n't do much t
vith the phone-a-friend part writ large . It was one hour	**that**	certainly felt like 408 . And as for the wheel landin
being very hopeful with the family , that it 's something	**that**	certainly was possible . ' To explain what the twins
's watch and branding Barr 'unfit for office' . 'I believe	**that**	certainly there has been an enormous amount of lav

Figure 4.6: Some concordance lines for *certainly* occurring after *that* at a distance of -1

These instances of *certainly* (shown in Figure 4.6) are, therefore, candidates for being in clause initial position in a *that*-clause. Actually, only seven out of ten examples are genuinely of *certainly* occurring initially within dependent clauses: one example in a complement clause and six in relative clauses. In the other three examples, *that* functions as a

determiner (the POS tagger has incorrectly marked it as complementiser). Figure 4.6 substantiates the usefulness of manually checking the automatically generated results for classifying clausal position.

For any given distance between the tokens of *that* and *certainly*, a single concordance may include instances of *certainly* in different clause positions and types. This can be seen in Figure 4.7.

It 's an unexpectedly heartening thought and one **that** the stock markets	**certainly**	drew comfort from . Wher
a few big surges . Health experts stress **that** official data almost	**certainly**	under-reports both infecti
' Ivey wrote in a letter . 'For **that** , they most	**certainly**	deserve a sincere , heartfe
Doctors have privately admitted to the boys **that** they are almost	**certainly**	not infectious , and that th
end . And I think , too , **that** the Liberal-National Party	**certainly**	did n't put forward a very
Harris added : 'There is something about Covid-19 **that** remains . 'I	**certainly**	would n't disagree a propo
in the general population . ' This means **that** it was almost	**certainly**	random that all the people
about it thoroughly . I can just add **that** we do not	**certainly**	have to force a player to s

Figure 4.7: Some concordance lines for *certainly* occurring after *that* at a distance of +4

Manually checking the concordance of which an extract is given in Figure 4.7 shows that the example of *certainly* is not necessarily in the dependent clause initiated by the clause marker; nor is it necessarily in clause initial or medial position, as might be assumed from the distance of +4 tokens. In two examples in Figure 4.7, *certainly* does indeed occur in the dependent clause, in medial position right after the subject. But in one example, "For that, they most certainly deserve […]" (ECC_mail8791571) *that* is mis-tagged as CST. In the third from last example, *certainly* occurs in another sentence altogether from *that*, in medial position in an independent clause. Neither of the latter exemplifies *certainly* in initial or medial position within a *that*-clause. This exemplifies why manual processing is still required to ascertain each MAC's actual clausal position and correct type of clause. But getting

subsets of the full concordance of *certainly* makes the process of manually identifying clausal position more efficient, because lots of easy examples are grouped together to be observed and categorised.

In practical terms, each of the smaller concordances described above is downloaded as a .txt file, and then imported into Microsoft Excel for manual annotation, as shown in Figure 4.8.

Text ID	Category				Context before	Query item
	I/M/F	S/C clause r	S Type	I/D		
gdn_books_	**M**	**S**	**Comp**	**D**	are genuine , in the sense that they are almost	**certainly**
gdn_books_	**M**	**S**	**Comp+Adv**	**D**	stay where they are , said that while he "	**certainly**
gdn_comm	**M**	**S**	**Comp**	**D**	Mr Johnson 's latest promise is that " we can	**certainly**
gdn_comm	**I**	**S**	**Comp**	**D**	over and over and over again that real life most	**certainly**
gdn_comm	**M**	**S**	**R**	**D**	be - but it also means that Johnson , who	**certainly**
gdn_comm	**M**	**C**	**that (DD1)**	**I**	n't need Brexit to tell us that , and we	**certainly**
gdn_educat	**M**	**S**	**_**	**D**	're not making any progress - that those students	**certainly**
gdn_food_2	**I**	**_**		**I**	ptillent' , which translates to 'eyes that sparkle' .	**certainly**

Figure 4.8: Some concordance lines for *certainly* followed by *that* at a distance of -4, as saved in Excel for categorisation.

The set of columns headed *Category* is added to the original textual download to contain the manual analysis, recorded using short labels. Here I/M/F means that the example MAC is in clause initial, medial, or final position respectively. I add S or C to indicate MACs in subordinate or coordinate clauses. For ease of data manipulation, I also added codes for independent and dependent clauses, i.e. I and D, though these are technically redundant given the presence of S/C. Where POS tagging has been used to identify relevant instances (as for *that*), if the marker is mis-tagged, the concordance spreadsheet records that information so that the example may be excluded from the count for that particular MAC. However, it is not discarded altogether. As Figure 4.8 shows, there are examples of *and*, *who* and *which* rather than *that* collocating with *certainly*. These are labelled accordingly and, like mis-tagged examples, excluded from the count for *that* + *certainly*. Later, when other MAC plus clause

marker pairs are queried, e.g. *and + certainly*, this example will automatically emerge there as well. So, instead of *that + certainly*, this example will in the end be counted under *and + certainly*.

Sometimes further classification is needed. For example, in Figure 4.8 the label *R* indicates that the example of *certainly* is actually not in a *that*-clause, but rather a relative clause. Moreover, in the second example shown, *that while he "certainly*, I have annotated the presence of both *that* (complementiser) and *while* (adverbial subordinator). This initial labelling flags such examples for later scrutiny. I will need to revisit such examples in the context of queries about for other clause markers occurring alongside *certainly*, i.e. *while + certainly* and *who + certainly*. Once the example *that while he certainly* is located in the *while + certainly* data, I will exclude it from the *that + certainly* group but retain it in the list of examples of the actually relevant clause marker. Therefore, the final counts for each collocate of each MAC exclude these duplicates.

In fact, for my quantitative analysis, I do not actually need to separate out the sequences of MACs occurring with each subordinate and coordinate clause marker. This is because I will present only consolidated figures for the placement of MACs in independent or dependent clauses. So, I actually only need to group subordinate clause and coordinate clause markers occurrences in two separate sets. In my quantitative analysis, I will not go into detail about the clause markers separately. But I *will* ultimately look at some of these collocates in my qualitative analysis. In my semantic analysis of MACs (see 2.4.2.1), I need to evaluate the modal scope that MACs have over other elements. Existing studies on MACs find that MACs explicitly or implicitly focalise the meaning implied in the clause in which they occur (see Section 2.5.1). For instance, when an HCS MAC occurs in a reason clause after *because*, it explicitly "specifies why the speaker feels entitled to make a claim" (Simon-Vandenbergen & Aijmer, 2007, p. 151). In addition, examination of the concordances thus generated will also

help in assessing other possible functions of the MACs under observation in both languages. Thus, the only crucial thing to be done with examples such as those previously mentioned involving *that/who* and *that/while* is to make sure that they are not double counted, by excluding them from the count of examples alongside *that*.

The number of instances of the MAC alongside each clause marker will then be totalled, as shown in Figure 4.9. Finally, the counts for dependent and independent clauses will be tabulated, as in Figure 4.10. Counts will be finalised after tallying them against the total frequency of the relevant MAC. This tallying of frequencies will ensure all occurrences of MACs in the corpora are accounted for.

	Initial	Coordinate markers					Initial Total	G.Total	Percent
HCS		but	yet	so	and	or			
Certainly		166	0	17	374	29	586	8342	7.0
Definitely		156	0	35	129	10	330	8058	4.1
Obviously		302	3	103	434	6	848	6939	12.2

Figure 4.9: An example of calculations recorded for three HCS MACs in clause initial position after coordinate clause markers

INDEPENDENT	I	M	F	
HCS				
Certainly	1146	3793	86	5025
Definitely	1129	4055	183	5367
Obviously	2427	2127	303	4857

Figure 4.10: An example of consolidated results for occurrences of three HCS MACs in clause initial, medial, and final position in independent clauses

Due to the size of the concordances, the possibility exists of human error in the classification of examples. Therefore, I adopt the procedure of re-checking the classification of each concordance three times. In case the total count of a MAC calculated via collocation

searches does not add up to actual frequency of that MAC occurrence in the data, I will re-run the whole query again for that particular MAC and collocate. For rechecking purpose, I will run a simple query for the two words in proximity, download its concordance, and scrutinise it for differences relative to the initial results. This will ensure accounting for any erroneous omission.

4.3.6 Step 5: Qualitative analytical procedures

Once the quantitative analysis is completed, I will start qualitative analysis. I will use the examples saved in Microsoft Excel to examine the concordances of the MACs in various clause positions to determine what different meanings they convey in English and Urdu. This concordance analysis will provide data for a description in terms of the semantics, indexical stances, and rhetorical-pragmatic functions that are associated with MACs by the existing literature on English MACs (see 2.5 and 2.6). This part of analysis will focus on placement of the MACs, and any effect this may have on the function of the MACs or of clauses that are in a MAC's scope (see 2.4.2.1). Together, the semantic and pragmatic analyses will answer to RQ 3.

The framework for concordance analysis that I adopt is a synthesis based on my literature review of previous researchers' frameworks (e.g. Van der Auwera & Plungian, 1999; Boye, 2012; Hoye, 1997; Simon-Vandenbergen, 2007) (see section 2.2). Prior research establishes that the English MACs are a part of a semantic group of epistemic (un)certainty (Van der Auwera & Plungian, 1998; Simon-Vandenbergen & Aijmer, 2007). Simon-Vandenbergen and Aijmer's (2007, p. 21) study multifunctionality and connections between HCS MACs in their cross-linguistic analysis of HCS MACs. Like Simon-Vandenbergen and Aijmer, I aim to observe shared semantic and pragmatic properties, and the meaning relations

between English and Urdu MACs, on the basis of corpus evidence. Table 2.2 (see Section 2.6) lists the parameters for semantic and pragmatic feature used in the relevant literature. I will address these same semantic and pragmatic features. However, I diverge from previous research of this kind in that I follow Boye's (2012) categorisation of MACs and thus look at MACs from the semantic groups both of *certainty* (Simon-Vandenbergen & Aijmer, 2007) and of *possibility* (Van der Auwera & Plungian, 1998): HCS and PS MACs respectively. A second difference is that I rely on comparable corpora for cross-linguistic data, rather than supplemental examples from elicitation experiment (e.g. Simon-Vandenbergen & Aijmer, 2007) or re-citing examples from grammars and prior studies (e.g. Boye, 2012) to show meaning connections in description of uses and functions of English and Urdu MACs.

The English and Urdu comparable corpora, ECC and LUWC, are composed of a mixture of written articles and online chat (which is often considered a hybrid of written and spoken media). However, my analysis does not incorporate a study of differences according to text medium. In case of online chat, neither the term *speaker* nor the term *writer* is wholly correct. Therefore, to keep things simple, I will use the cover term *addresser* in lieu of *speaker* or *writer* to refer to the producer of a given example in my analyses.

I will select and separate supporting examples for clause initial, medial, and final positions for the various functions I discuss. I will also select examples of MACs in specific environments, e.g. in conditional sentences after clause marker *then*. Moreover, I will choose examples by semantic groups (HCS and HCSNC MACs versus PS and PSNC MACs) rather than separately for each MAC; describing the functions of each individual MAC as opposed to the category as a whole is not among the aims of this project. In sum, to exemplify a description of usage and function, I will choose one example involving either HCS(NC) and HCSNC MACs, and PS and PSNC MACs.

I will create a separate Microsoft Word document to store the selected examples for each part of the analysis, sorted under headings for the parameters of the analysis (see Table 2.2). These separate collections will then be the basis of my discussions of different issues of semantics and pragmatics. In case I need to look at additional context to interpret the function of any particular example of a MAC, I will use the extended context function to access the surrounding text in CQPweb, as illustrated in Figure 4.11.

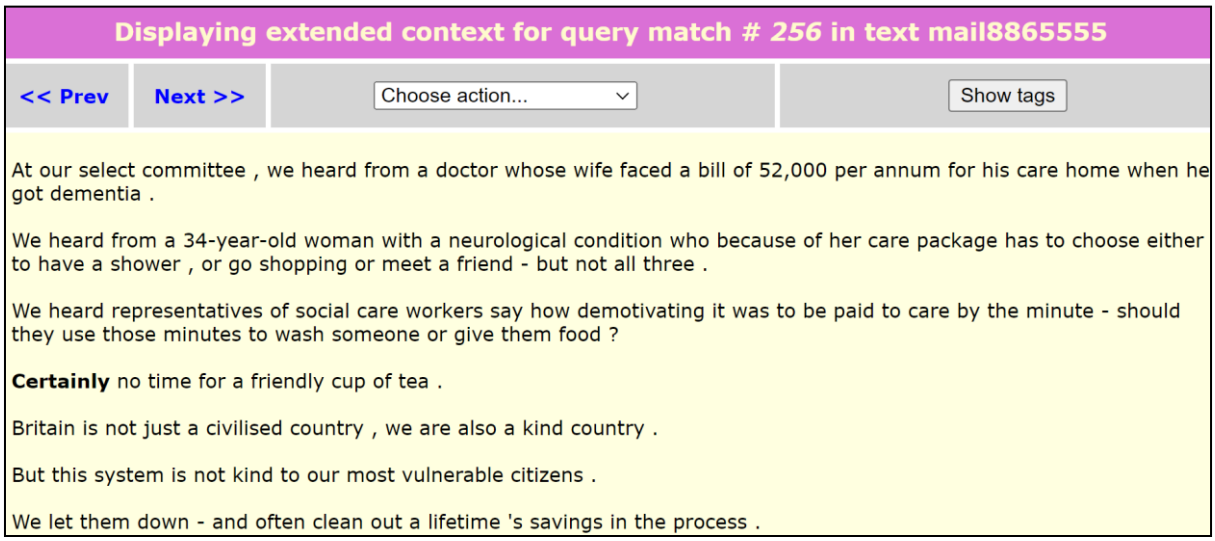


Figure 4.11: Extended context display for one concordance line from the query shown in Figure 4.7

Figure 4.11 exemplifies the use of *certainly* in clause initial position to express referential continuity of a topic discussed in previous clauses. Pragmatically, *certainly* in this example functions as an authority and expectation marker. The addresser, implied by the mention of "our select committee" to be a Member of Parliament, uses *certainly* to emphatically criticise an unreasonable policy for making a "cup of tea" (ECC_mail8865555) definitely impossible as something a care worker might do for a disabled person. Use of *certainly* also expresses the addresser's expectation that this viewpoint is shared by their addressees. The stance of the addresser here is authority, because they express certainty about the knowledge they present.

Concordance analysis will also help me identify similarities and differences in the functions of the English and Urdu MACs, and to determine which MACs are preferentially used for particular functions. As the descriptions of semantic and pragmatic functions of English and Urdu MACs will be primarily informed by the existing literature on these features as studied in English MACs, I expect that differences may well emerge between how MACs are used in English versus Urdu. This is to be expected, if for no other reason, because of differences in word order between the two languages. There may also be some differences due to the relatively lesser importance of modal verbs in Urdu as compared to English (see Section 1.4). This part of analysis will answer RQ 4 in the Discussion chapter.

4.4 Identifying English-Urdu MACs via literature and corpus evidence

I now present the first step of my analysis. This part of the analysis is included in the methodology chapter because finalising which English and Urdu MACs will be examined in later chapters is a matter of method design, and largely a preparatory step for subsequent phases of the investigation.. As section 4.3.2 related, the first step of the whole procedure is to establish which Urdu MACs correspond with each of the English MACs known from the literature.

4.4.1 English and Urdu MACs in the parallel corpora

Table 4.8 lists the MACs (including MAC phrases; see 2.3.4) that occur in the English parallel corpus and their translations as observed in the Urdu parallel corpus. To assemble this list, I searched for all the English MACs as listed in Table 2.1. I present my findings following Boye's (2012; 2016) scale of certainty for a proposition categorisation (see section

2.4.2.3; Figures 2.1 and 2.2). So, for English MACs I started with searches for HCS MACs like *certainly* before moving on to PS MACs like *possibly*. To locate combinations of HCS MACs + NEG (i.e. HCSNC), and PS + NEG (i.e. PSNC), I searched for each MAC's negative construction. Then I subtracted the frequency of the negative constructions from the MAC's overall frequency to calculate, by exclusion, the frequencies of the HCS and PS MACs. For each MAC, I identified the corresponding Urdu element via the corpus alignment. I then counted the occurrences of Urdu MACs, to ascertain whether Urdu MACs, rather than some other element, are *always* used to translate English MACs. The purpose of these calculations is to generate a list of English MACs and their Urdu translation equivalents to be the object of the main analyses. Cross-tallying the English and Urdu frequencies illustrates the extent to which other lexical means are used to convey degree of certainty about a proposition. Table 4.8 shows the results obtained from these searches.

Table 4.8: MACs in English and their translations into Urdu extracted from the parallel corpora

Epistemic support level	English MAC	Frequency	Translation(s) into Urdu MACs	Frequency	Translation(s) into other Urdu modal expressions	Frequency
HCS	*Of course*	8	*bilāśubha* 'undoubtedly'	2		
			beśak 'of course'	1		
			yaqīnī tōr par 'certainly'	1		
			śak nahīṁ 'no doubt'	1		
	Definitely	2	*zarūr* 'definitely'	1		
			yaqīnī tor par 'certainly'	1		
	Certainly	1	*yaqīnī tor par* 'certainly'	1		
	Obviously	1	*bilāśubha* 'undoubtedly'	1		
	Undoubtedly	1	*yaqīnān* 'definitely'	1		
	No doubt	1	*koī śubha nahīṁ* 'no doubt'	1		
PS	*Likely*	54	*śāyad* 'possibly'	2	*imkān* 'likelihood'	32
			hō saktā 'maybe'	8	*mumkin* 'possibility'	2
					mumkinah 'possible'	2
					imkānī 'possible'	1
					umīd 'hope'	1
					tawaqo 'expectation'	1
	Probably	20	*ġālibān* 'probably'	6	*amūmān* 'most likely/usually'	2
			śāyad 'possibly'	4	*mumkin* 'possible'	4
			hō saktā 'maybe'	3	*umīd* 'hope'	1
	Perhaps	18	*śāyad* 'perhaps'	9		
			ġālibān 'perhaps'	5		
	Possibly	5	*śāyad* 'possibly'	1	*imkānī* 'possibility'	1
			hō saktā 'maybe'	2	*mumkinah* 'probable'	1
	Maybe	3	hō saktā 'maybe'	1		
HCSNC	*Definitely not*	1	*yaqīnān nahīṁ* 'definitely not'	1		

The tabulated frequencies in Table 4.8 do not tally, as any English MAC can potentially be translated as a noun, a verb or an adjective in Urdu, and vice versa. In some cases, an English MAC such as *likely* (adverb/adjective) has not been translated by an Urdu MAC but rather by some other modal expression such as *imkān* 'likelihood' (noun). Other English MACs are translated by Urdu MAC phrases. In particular, the root *sak* 'can', which is an MV in Urdu, is used to form these phrases in its nominative masculine singular imperfective participle form marked with *-tā*. The concordances show that *sak* preceded by existential verb *ho* 'be' forms a MAC phrase which overall indicates probability (see Genady, 2005, p. 181).

Yet other instances of English MACs are not translated by any Urdu modal expression at all, as in (60).

60. "Beware of traders who pose as private sellers, *perhaps* at a car boot sale or through a small advertisement" (411_En_EMILLE)

"Aēsē		tājirōṁ=sē		bac	kar	rahīyē	jō
Such.M.PL.OBL		trader.M.PL.OBL=from		save	do	remain.IMP	REL
kār	būṭ	saēl=mēṁ	yā	kisī	chōtē		
car	boot	sale=in	or	some	small.M.PL.OBL		
iśhtēhār=kē			zarīa	sāmān	bēcanē	vālā	
advertisement=GEN.M.PL			through	good	sell.INF.OBL	VALA.M.SG	
hōnē=kā		dhōng	rachātē	haiṁ"			
be.INF=GEN.M.SG		pretend	set.OBL	be.PRS.3.PL (411_Ur_EMILLE).			

All three Urdu MACs conveying possibility (*ġālibān*, *śāyad*, and *hō saktā*) interchangeably translate *possibly*, *probably* and *maybe* in different sentences. This observation confirms that the same linguistic expression may convey partial or neutral support for epistemic stance, depending on context. Similarly, among the HCS MACs, *of course* occurs eight times in the data and has been translated as *bilāśubha* 'undoubtedly',

beśak 'of course', *yaqīnī tor par* 'certainly', and *śak nahīṁ* 'no doubt'. Likewise, *definitely* occurs twice, translated by *żarūr* 'definitely' and *yaqīnī tor par* 'certainly'. Furthermore, as anticipated, English MACs are not always translated to Urdu MACs, but if they are, then an English HCS MAC translates as an Urdu HCS MAC, and an English PS MAC translates as an Urdu PS MAC.

Interestingly, one highly frequent English word which corresponds in the parallel data to Urdu MACs is *likely*. *Likely* is translated either as a MAC (e.g. *śāyad* 'probably') or a noun (e.g. *imkān* 'possibility'). Across fifty-nine instances of *likely*, the most common constructions are *likely* + *to*-INF (37), followed by *more* + *likely* (1), *more* + *likely*+ *to* (10), *less* + *likely* (6), and *most* + *likely* (4). Of the construction *likely* + *to*-INF (37), ten are translated as Urdu MACs: *śāyad* 'probably' (2) and *ho saktā* (8). Huddleston & Pullum (2002, p. 568) say that *likely* is in that category "where there is little difference in meaning" between the adverbial and adjectival forms. The POS of *likely* depends on its sentence position. For instance, when *likely* modifies a noun (e.g. *likely departure),* or precedes conjunction *that,* it functions as an adjective. Manual observation of concordance lines shows that there are only two instances in which *likely* functions as an adverb. Nevertheless, I retain *likely* among the English MACs to be investigated; but before any analysis, I will first manually exclude all constructions from the data where *likely* functions as an adjective by subtracting the frequency of those constructions from the overall frequency of *likely* in the comparable corpus.

Twice, *likely* is translated as Urdu adverb *umūmān,* with the sense of 'probably'; but this Urdu word is commonly used as a frequency adverb (translated 'usually'), not as a MAC. Therefore, I will not address *umūmān* 'usually' in any subsequent analysis. Having excluded *umūmān* the complete lists of English and Urdu MACs, by categories, with parallel corpus frequencies, in Tables 4.8 and 4.9.

Table 4.9: Frequencies of English MACs in the parallel corpora

English MAC	Frequency	Relative frequency per 100,000 words
HCS		
Certainly	1	0.7
Definitely	2	1.5
Obviously	1	0.7
of course	8	5.9
no doubt	1	0.7
Undoubtedly	1	0.7
PS		
Likely	2	1.5
Maybe	3	2.2
Probably	20	14.7
Perhaps	18	13.2
Possibly	5	3.7
HCSNC		
definitely not	1	0.7

Table 4.10: Frequencies of the translated Urdu MACs in the parallel corpora

Urdu MAC	Frequency	Relative frequency per 100,000 words
HCS		
bēśak 'undoubtedly'	3	1.5
bilāśubha 'of course'	2	1.0
śak nahīṁ 'no doubt'	2	1.0
koī śubha nahīṁ 'no doubt'	1	0.5
yaqīnān 'certainly'	1	0.5
yaqīnī tor par 'certainly'	7	3.6
ẓarūr 'definitely'	12	6.1
PS		
śāyad 'perhaps'	48	24
hō saktā 'maybe'	531	269.7
ġālibān 'probably'	11	5.6
HCSNC		
yaqīnān nahīṁ 'definitely not'	1	0.5

Tables 4.9 and 4.10 show that PS MACs (e.g. *probably, śāyad*) occur more frequently than HCS MACS (e.g. *certainly, yaqīnān*). The frequencies of some entries in Table 4.9 and 4.10 are greater than the frequencies of the same MACs in Table 4.8. These differences arise because, like the English MACs, Urdu MACs are sometimes translated by MVs or other modal expressions, due to "translator's choice" (Malmkjær, 2009). There are twelve instances of HCS *żarūr* 'definitely' in Urdu data, but only twice does it occur as a translation for English MAC *definitely*. Similarly, PS *śāyad* 'perhaps' occurs forty-six times in the data but only four times translating *probably* and nine times translating *perhaps.* Table 4.11 presents Urdu MACs in the parallel data and the elements that they translate. The last column of Table 4.11 presents are the number of instances in which the listed Urdu MAC does not directly translate any specific English word in the text.

Table 4.11: Frequencies of Urdu MACs and what they translate in the parallel data

Urdu MACs in parallel data	Frequency	Translated from	Frequency	Frequency translating nothing
HCS				
bēśak ‘undoubtedly’	3	of course	2	1
bilāśubha ‘of course’	2	undoubtedly	2	
śak nahīṁ ‘no doubt’	2	of course	1	1
koī śubha nahīṁ ‘no doubt’	1	no doubt	1	
yaqīnān ‘certainly’	1	undoubtedly	1	
yaqīnī tor par ‘certainly’	7	certainly	1	3
		definitely	1	
		of course	1	
żarūr ‘definitely’	21	definitely	1	15
		do	2	
		ensure	1	
		make sure	2	
PS				
śāyad ‘perhaps’	48	could/could be	7	
		likely to be	2	
		may/may be	19	
		might	6	
		perhaps	9	
		possibly	2	
		probably	3	
hō saktā ‘maybe’	531	likely	8	
		may	452	
		maybe	3	
		might	65	
		possibly	2	
		probably	1	
ġālibān ‘probably’	11	perhaps	1	4
		probably	6	
HCSNC				
yaqīnān nahīṁ ‘definitely not’	1	definitely not	1	

Table 4.11 shows that Urdu MACs not only translate English MACs but also at times English MVs, including *may*, *might*, and *could*, where the sense is epistemic modality as in

example (61). This shows that in the absence of any corresponding Urdu MV, the speakers may use a MAC or some other modal expression to translate an English MV which they think approximately has the same meaning as the English MV.

61. "You *might* want to think about:" (1065_En_EMILLE).

"Āp	*śāyad*	mandarjha.zēl=kē	mut'aliq	sōcnā
You	*perhaps*	following=GEN.M.PL	about	think.INF

cāhiēṁ	gē"	(1065_Ur_EMILLE).
want.SBJV.3.PL	go.FUT.M.PL	

Urdu MAC *żarūr* 'definitely' is used as both epistemic and deontic modal adverb in the parallel data. In addition, it occurs in imperatives, even though the original English of such clauses has no MV or MAC. Similarly, it sometimes translates emphatic *do* as in (62).

62. "Sometimes those strengths can be hard to find, but they *do* exist"

(2379_En_EMILLE).

"Ba'az	aōqāt	in	ṣalāhīyatōṁ=kō		pānā	muśkil
Some	time.M.PL	DEM	strength.M.PL.OBL=ACC		find	difficult

hōtā	hai,	magar	yeh	mōjūd	*żarūr*
be.IPFV.M.SG	be.PRS.2.SG,	but	DEM	exist	*definitely*

hōtī	haiṁ"	(2379_Ur_EMILLE).
be.IPFV.M.SG	be.PRS.3.PL	

My original queries found only one example of a MAC phrase formed with the negator, MAC + NEG: *definitely not* translated as *yaqīnān nahīṁ* (both HSCNC MACs). By searching for each Urdu MAC with *nahīṁ* 'not' (subsequently, or for *ho saktā* 'maybe', between the two words of the phrase), I found one more: *koī śubha nahīṁ* 'no doubt, lit. any doubt not', which was added to the list (see Table 4.11) under category HCS.

It may be noted that when *nahīṁ* functions as a prohibitive or imperative marker, it is used as a strong negative marker (Koul, 2008, p.118). Another strong negative in Urdu is adverb phrase *hargiz nahīṁ* 'absolutely not', occurs twice in the English-Urdu parallel corpus and is translated for 'must not' and 'never'. *Hargiz nahīṁ* is used in Urdu to exclude any possibility of the alternative proposition to be true. The low or zero frequencies of the MAC + NEG phrases are probably due to the small size of the parallel corpora. Therefore, beyond this point, I retain the negative phrases within my set of MACs under consideration if they have zero instance in the parallel data.

4.4.2 English and Urdu MACs in the comparable corpora

The primary analyses of the thesis are based on searching for English and Urdu MACs in the comparable corpora, as previously discussed (see 4.3.2-4.3.7). I need to account for all possible English and Urdu MACs at this stage. The process to extract English and Urdu MACs is as given in section 4.3.3. English and Urdu MACs as identified in parallel corpus will be searched for HCS and PS and their negative phrases categorised as HCSNC and PSNC. I present the English and Urdu MACs found in the comparable corpora in Tables 4.12 and 4.13.

Table 4.12: English MACs in ECC

English MACs	Frequency	Relative frequency per 100,000 words
HCS		
certainly	8,342	8.5
definitely	8,058	8.3
obviously	6,939	7.1
of course	9,961	10.2
no doubt	2,445	2.5
undoubtedly	858	0.9
Total	**36,603**	**37.6**
PS		
likely	6,149	6.3
maybe	12,899	13.3
perhaps	9,426	9.7
possibly	4,896	5.0
probably	20,681	21.3
Total	**54,051**	**55.6**
PSNC		
likely not	167	0.2
maybe not	382	0.4
perhaps not	221	0.2
possibly not	56	0.1
probably not	1,010	1.0
Total	**1,836**	**1.9**
HCSNC		
certainly not	926	1.0
definitely not	699	0.7
obviously not	296	0.3
of course not	150	0.2
no doubt not	-	-
undoubtedly not	1	0.001
Total	**2,072**	**2.1**

Table 4.13: Urdu MACs in LUWC

Urdu MACs	Frequency	Relative frequency per 100,000 words
HCS		
beśak ‘undoubtedly’	1,088	4.5
bilāśubha ‘of course’	1,659	6.9
koī śubha nahīṁ ‘no doubt’	199	0.8
śak nahīṁ ‘lit. any doubt not’	435	1.8
yaqīnān ‘certainly’	2,596	10.8
yaqīnī tor par ‘certainly’	206	0.9
żarūr ‘definitely’	8,220	34.2
Total	**14,403**	**59.9**
PS		
śāyad ‘perhaps’	13,300	55.3
ho saktā ‘maybe’	8,461	35.2
ġālibān ‘probably’	1,538	6.4
Total	**23,299**	**96.9**
PSNC		
śāyad nahīṁ ‘perhaps not’	92	0.4
ho nahīṁ saktā ‘possibly not’	98	0.4
ġālibān nahīṁ ‘probably not’	10	0.0
Total	**200**	**0.8**
HCSNC		
beśak nahīṁ ‘undoubtedly not’	1	0.0
bilāśubha nahīṁ ‘of course not’	1	0.0
yaqīnān nahīṁ ‘definitely not’	55	0.2
yaqīnī tor par nahīṁ ‘certainly not’	4	0.0
żarūr nahīṁ ‘definitely not’	3	0.0
Total	**64**	**0.3**

In the comparable corpora, we find more distinct MACs in the HCS category than in the PS category, but the PS MACs tend to occur more often in aggregate: 54.0 versus 46.0 per 100,000 words in English, a slight difference, and 97.0 versus 61.0 per 100,000 words in Urdu, a considerable difference. The negated MAC phrases are less frequent than non-negated MACs in both languages. Interestingly, the HCSNC type in aggregate is more frequent than PSNC in English but less frequent in Urdu. This is the opposite way around from the HCS and PS categories for English, but for Urdu the probability support categories

(both positive and negative) are both higher than the corresponding high certainty category. I revisit this issue for further discussion in section 8.2.

A notable difference is the overall lower frequency of Urdu HCSNC and PSNC MACs. While the negative content categories are rare in both languages, in Urdu they are even more infrequent than in English

4.4.3 Distribution of English and Urdu MACs

The distribution of English and Urdu MACs across types of text in comparable corpora is presented in Tables 4.14 and 4.15 (visualised in Figure 4.12 and 4.13).

Table 4.14: Distribution of English MACs in ECC according to category of texts as frequency of per hundred thousand words[28]

HCS	BP	OC	E	F	N	Op
Certainly	17.0	17.2	24.3	11.2	6.0	15.9
Definitely	5.8	39.0	2.4	5.8	2.9	2.6
Obviously	7.3	21.8	9.7	6.5	4.2	4.7
of course	19.9	20.4	36.4	11.8	5.7	28.9
no doubt	6.3	1.8	14.6	3.0	2.2	5.7
Undoubtedly	1.0	0.5	4.9	1.2	0.7	2.7
PS	**BP**	**OC**	**E**	**F**	**N**	**Op**
Likely	1.9	10.2	0.0	4.9	6.2	3.4
Maybe	16.0	53.8	12.1	10.7	4.4	11.7
Perhaps	43.6	13.2	34.0	13.4	5.6	30.4
possibly	4.8	9.5	7.3	4.0	4.3	7.9
Probably	25.2	89.3	24.3	14.0	8.8	17.3
PSNC	**BP**	**OC**	**E**	**F**	**N**	**Op**
likely not	0	0.2	0	0.1	0.1	0
maybe not	0	1.5	0	0.3	0.1	0.5
perhaps not	1.5	0.3	0	0.3	0.1	0.5
possibly not	0	0.2	0	0.1	0.0	0.1
probably not	0.5	4.4	0	0.6	0.4	1.1
HCSNC	**BP**	**OC**	**E**	**F**	**N**	**Op**
certainly not	2.9	2.2	0	0.8	0.6	1.5
definitely not	0.5	3.4	0	0.3	0.2	0.4
obviously not	0	1.2	0	0.1	0.2	0.3
of course not	0	0	0	0	0	0
no doubt not	0	0	0	0	0	0
undoubtedly not	0	0	0	0	0	0

[28] BP= Blog Posts; OC= Online chat; E= Editorials; F= Features; N=News; O=Opinion

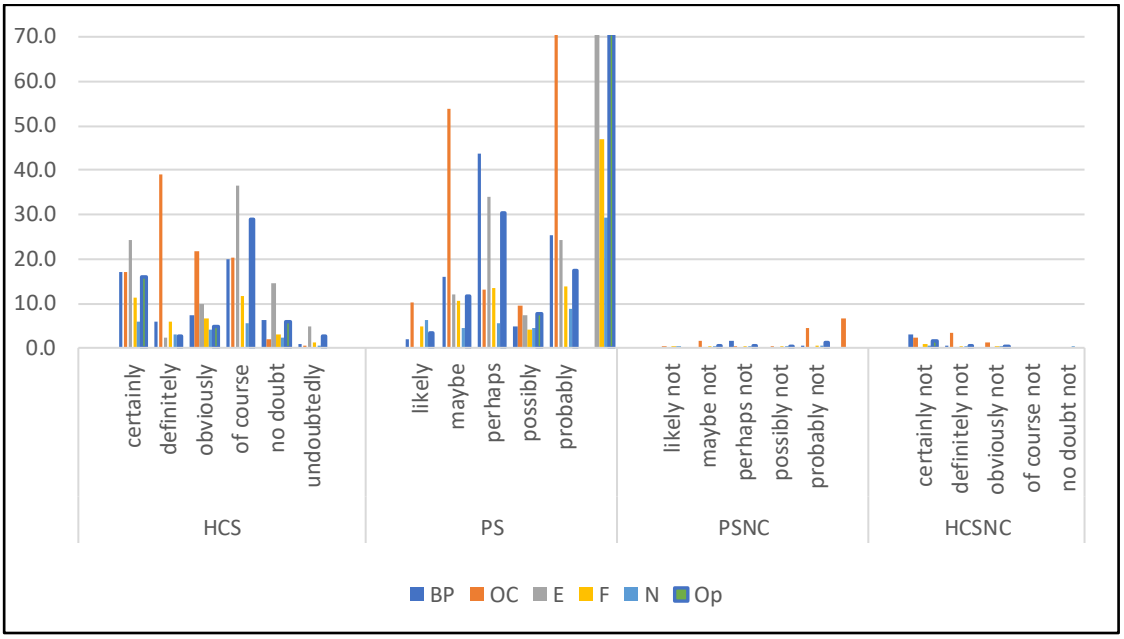


Figure 4.12: Relative frequencies of English MACs across categories in ECC

Table 4.14 and Figure 4.12 show that overall, English MACs are most frequent in online chat, followed by editorials and news texts (which are roughly equal). The frequencies of *probably*, and *perhaps* (PS) and *of course* and *certainly* (HCS) are noticeably higher in editorials, whereas the frequencies of *maybe*, *probably* and *likely* (PS) and *definitely*, *obviously* and *of course* (HCS) are higher in online chat than elsewhere. Among the negative MACs, *definitely not* (HCSNC), *probably not* and *maybe not* (PSNC) are most frequent in online chats and news respectively. The ECC text type in which MACs are least frequent is blog posts.

Table 4.15: Distribution of Urdu MACs in LUWC according to category of texts as frequency of per hundred thousand words

HCS	BP	OC	E	F	N	Op
beśak 'undoubtedly'	0.0	5.7	1.1	3.0	0.0	2.5
bilāśubha 'of course'	5.9	2.2	30.8	5.7	1.3	6.5
koī śubha nahīṁ 'no doubt'	0.0	0.6	9.9	12.3	0.0	2.1
śak nahīṁ 'lit. any doubt not'	4.5	2.2	3.5	3.2	0.2	3.0
yaqīnān 'certainly'	13.4	10.5	11.9	6.4	2.6	12.9
yaqīnī tor par 'certainly'	3.0	0.7	1.5	1.0	0.3	0.7
żarūr 'definitely'	31.2	41.1	12.0	18.0	11.1	23.6
PS	**BP**	**OC**	**E**	**F**	**N**	**Op**
śāyad 'perhaps'	28.2	69.5	10.2	24.5	6.9	34.7
ho saktā 'maybe'	1.5	38.1	32.2	22.0	14.9	24.1
ġālibān 'probably'	3.0	7.8	2.1	5.2	0.5	4.0
PSNC	**BP**	**OC**	**E**	**F**	**N**	**Op**
śāyad nahīṁ 'perhaps not'	1.5	0.5	0.0	0.2	0.3	0.1
ho nahīṁ saktā 'possibly not'	0	0.4	0.3	0.5	0.0	0.4
ġālibān nahīṁ 'probably not'	0	0.1	0	0	0	0
HCSNC	**BP**	**OC**	**E**	**F**	**N**	**Op**
beśak nahīṁ 'undoubtedly not'	0	0	0	0	0	0
bilāśubha nahīṁ 'of course not'	0	0	0	0	0	0
yaqīnān nahīṁ 'definitely not'	0	0.3	0.1	0	0	0
yaqīnī tor par nahīṁ 'certainly not'	1.0	0	0	0	0	0
zarūr nahīṁ 'definitely not'	0	0	0.1	0	0	0

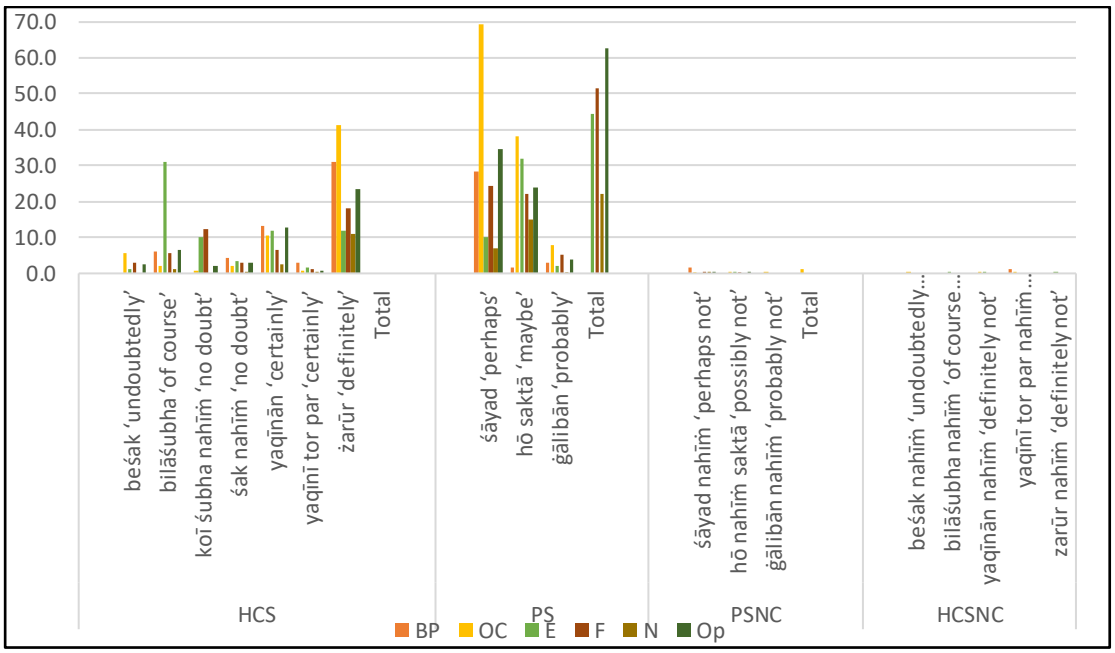


Figure 4.13: Relative frequencies of Urdu MACs across categories in LUWC

Table 4.15 and Figure 4.13 show that overall, Urdu MACs are most frequent in online chat, followed by editorials and opinion articles (which are roughly equal). The frequencies of *śāyad* 'perhaps' (PS) and *zarūr* 'definitely' (HCS) are noticeably higher in online chat than elsewhere. HCSNC MAC *śak nahīṁ* 'no doubt' and PSNC MAC *śāyad nahīṁ* 'perhaps not' are most frequent in opinion and blog posts respectively. The LUWC text types in which MACs are least frequent is news, followed by features.

English and Urdu PSNC, and English HCSNC, MACs occur in all categories except editorial texts; Urdu HCSNC MACs occur in all except news and feature article texts. Both HCS and PS MACs are more frequent in online chat in English than in Urdu, whereas they occur more commonly in editorial texts in Urdu than in English. By contrast, PS MACs are more common in editorials in English than in Urdu, and slightly higher in opinion in English than in Urdu. PSNC MACs are most frequent in online chat in both English and Urdu than any other text type. Within online chat, English PSNC MACs are noticeably higher than Urdu PSNC MACs. In contrast, HCSNC MACs are more common in online chats in English than any other text type. Urdu HCSNC MACs are rare in all text types.

4.4.4 Summarising the results

Overall, the foregoing quantitative results indicate that, all four categories of HCS, HCSNC, PS, and PSNC exist in Urdu as in English. Categories of MACs in both languages occur in all text types, but MACs and their negative constructions in different text types vary in frequency. The results justify my stance that the available parallel corpus is not sufficient to carry out detailed corpus-based descriptive research on MACs, due to their low frequency, although it was appropriate to use parallel data at the initial stage of inquiry to identify which Urdu MACs correspond with English MACs. However, insight into the similarities and

differences in behaviour of the English and Urdu MACs, requires more data for which it is necessary to turn to monolingual comparable corpora. Similarities and differences in occurrence of English and Urdu MACs will be discussed in Chapter 8.

4.5 Chapter summary

This chapter has outlined my methodology and my preliminary quantitative investigation. First, I discussed the data that I have used, and my justification for compiling and using novel comparable corpora of English and Urdu instead of relying on previously available data. I also explained my presentation of examples from the corpora. Subsequently, I explained my analytical procedures step by step, clarifying how they are linked to my research questions. Finally, I presented the quantitative analysis which answers RQ 1(see 3.6). This part of the analysis has been included within the methodology because finding out what MACs actually exist in Urdu was a necessary part of devising procedures (e.g. query terms). The answer to RQ 1 determines what English and Urdu MACs are to be considered in subsequent chapters. The frequency data support my decision to proceed with semantic and pragmatic analyses using concordances extracted from comparable corpora.

5 Placement of English and Urdu MACs within the clause

5.1 Chapter overview

In this chapter, I consider the tendency (or lack thereof) for English and Urdu MACs to occur in particular positions in various types of independent and dependent clauses. For this purpose, I observe patterns of occurrence, at clause level, of English MACs in the English Comparable Corpus (ECC) and Urdu MACs in the Lancaster Urdu Web Corpus (LUWC). This part of the analysis answers my research question 2: whether the placement of English and Urdu MACs in different (independent and dependent) clauses is similar to or different from how it is observed in previous studies on English MACs (see 2.4.1).

I first present the frequencies with which the English and Urdu MACs occur in different clausal positions in independent clauses in sections 5.2 and 5.3 respectively; in sections 5.4 and 5.5, I present the equivalent frequencies in dependent clauses. Cumulative frequency counts folding together the figures for both independent and dependent clauses are given in section 5.6. Finally, I summarise these quantitative findings in section 5.7.

5.2 Clause level position of English MACs in independent clauses

In this section, I present the frequencies with which English MACs occur in initial, medial, and final position in independent clauses, identified according to the methods outlined in section 4.3.5. Table 5.1 presents, for each English MAC, both the frequency with

which it occurs in the three possible positions, and also the percentage of the whole frequency of the MAC in independent clauses that each frequency represents. The MACs are grouped as *high certainty support* (HCS), *probability support* (PS), *probability support for negative content* (PSNC), and *high certainty support for negative content* (HCSNC); see section 2.3.3.3.

Table 5.1: Positions of English MACs occurring in independent clauses in ECC

MACs	Initial position		Medial position		Final position		Total	Overall percent
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
certainly	1,146	22.8%	3,793	75.5%	86	1.7%	5,025	
definitely	1,129	21.0%	4,055	75.6%	183	3.4%	5,367	
obviously	2,427	50.0%	2,127	43.8%	303	6.2%	4,857	
of course	4,272	55.6%	2,433	31.7%	972	12.7%	7,677	
no doubt	288	17.3%	1,269	76.1%	110	6.6%	1,667	
undoubtedly	77	35%	142	64.5%	1	0.5%	220	
Total	**9,339**		**13,819**		**1655**		**24,813**	**40.9%**
Percent	**37.6**		**55.7**		**6.7**		**100.0**	
PS								
likely	169	4.6%	3,501	95.4%	0	0	3,670	
maybe	6,657	63.4%	1,065	10.1%	2,775	26.4%	10,497	
perhaps	4,144	60%	592	8.6%	2,166	31.4%	6,902	
possibly	1,369	39.8%	298	8.7%	1,770	51.5%	3,437	
probably	2,418	28.9%	2,342	28%	3,597	43%	8,357	
Total	**14,757**		**7,798**		**10,308**		**32,863**	**54.2%**
Percent	**44.9**		**23.7**		**31.4**		**100.0**	
PSNC								
likely not	4	3.2%	119	94.4%	3	2.4%	126	
maybe not	170	51.2%	106	31.9%	56	16.9%	332	
perhaps not	145	76.3%	26	13.7%	19	10%	190	
possibly not	30	68.2%	11	25%	3	6.8%	44	
probably not	334	41.5%	409	50.8%	62	7.7%	805	
Total	**683**		**671**		**143**		**1,497**	**2.5%**
Percent	**45.6**		**44.8**		**9.6**		**100.0**	
HCSNC								
certainly not	220	35.9%	370	60.5%	22	3.6%	612	
definitely not	169	34.6%	301	61.7%	18	3.7%	488	
obviously not	70	32.4%	140	64.8%	6	2.8%	216	
of course not	93	66%	23	16.3%	25	17.7%	141	
no doubt not	0	0	0	0	0	0	0	
undoubtedly not	0	0	1	100%	0	0	1	
Total	**552**		**835**		**71**		**1,458**	**2.4%**
Percent	**37.9**		**57.3**		**4.9**		**100.0**	
Grand Total	**25,331**	**42%**	**23,123**	**38%**	**12,177**	**20%**	**60,631**	**100%**

Overall, in independent clauses, the most frequent MACs are those in the PS group, accounting for 54.2% of total examples. The second most frequent group is HCS, accounting

for 40.9%. This difference between English PS and HCS MACs is not very pronounced. The negative MACs are approximately same in frequency: 2.5% examples of PSNC and 2.4% of HCSNC. Both these negative groups are markedly less common than the positive groups in ECC.

5.3 Clause level position of Urdu MACs in independent clauses

In this section, I present the frequencies with which Urdu MACs occur in initial, medial, and final positions in independent clauses, identified following the methods outlined in section 4.3.5. Table 5.2 presents, for each Urdu MAC, both the frequency with which it occurs in the three possible positions, and also the percentage of the whole frequency of the MAC in independent clauses that each frequency represents. Again, the MACs are categorised as HCS, PS, PSNC, and HCSNC.

Table 5.2: Positions of Urdu MACs occurring in independent clauses in LUWC

MACs	Initial position		Medial position		Final position		Total	Overall Percentage
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
beśak 'undoubtedly'	554	57.7%	330	34.4%	76	7.9%	960	
bilāśubha 'of course'	1,098	70.8%	438	28.2%	15	1.0%	1,551	
koī śubha nahīṁ 'no doubt'	8	4.4%	142	78.5%	31	17.1%	181	
śak nahīṁ 'no doubt'	12	3.3%	235	64.2%	119	32.5%	366	
yaqīnān 'certainly'	726	38.9%	1,123	60.1%	19	1.0%	1,868	
yaqīnī tor par 'certainly'	32	21.3%	107	71.3%	11	7.3%	150	
żarūr 'definitely'	1,570	24.1%	4,100	62.9%	853	13.1%	6,523	
Total	**4,000**		**6,475**		**1,124**		**11,599**	**36.5%**
Percentage	**34.5%**		**55.8%**		**9.7%**		**100%**	
PS								
śāyad 'perhaps'	4,841	44.4%	5,588	51.2%	483	4.4%	10,912	
hō saktā 'maybe'	770	9.9%	2,129	27.4%	4,868	62.7%	7767	
ġālibān 'probably'	262	19.9%	1,035	78.7%	18	1.4%	1315	
Total	**5,873**		**8,752**		**5369**		**19,994**	**62.9%**
Percentage	**29.4%**		**43.8%**		**26.9%**		**100%**	
PSNC								
śāyad nahīṁ 'perhaps not'	43	59.7%	22	30.6%	7	9.7%	72	
ho nahīṁ saktā 'possibly not'	1	2.2%	0		45	97.8%	46	
ġālibān nahīṁ 'probably not'	2	20.0%	8	80.0%	0		10	
Total	**46**		**30**		**52**		**128**	**0.4%**
Percentage	**35.9%**		**23.4%**		**40.6%**		**100%**	
HCSNC								
yaqīnān nahīṁ 'definitely not'	16	34.8%	17	37.0%	13	28.3%	46	
yaqīnī tōr par nahīṁ 'certainly not'	1	50.0%	1	50.0%	0		2	
żarūr nahīṁ 'definitely not'	0		2	100.0%	0		2	
Total	**17**		**20**		**14**		**51**	**0.2%**
Percentage	**33.3%**		**39.2%**		**27.5%**		**100%**	
Grand Total	**9,936**	**31%**	**15,277**	**48%**	**6,559**	**21%**	**31,772**	**100%**

In Urdu as in English, in independent clauses, the most frequent MACs are those of the PS group, accounting for 62.9% of total examples. However, unlike the roughly equally frequent English HCS and PS MACs, the Urdu PS MACs are almost twice as frequent as the Urdu HCS MACs, a noticeable difference between the Urdu PS and HCS MACs. As with the

English MACs, the Urdu MACs expressing negative support for a proposition are rare overall, at only 0.4% and 0.2% for PSNC and HCSNC respectively, rates even lower than those of the corresponding groups in English.

5.4 Clause level positions of English MACs in dependent clauses

In this section, I present the frequencies with which English MACs occur in initial, medial, and final positions in dependent clauses, identified according to the methods outlined in section 4.3.5.

Table 5.3: Positions of English MACs occurring in dependent clauses in ECC

MACs	Initial position		Medial position		Final position		Total	Overall percent
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
certainly	109	3.3%	3,204	96.6%	4	0.1%	3,317	
definitely	83	3.1%	2,603	96.7%	5	0.2%	2,691	
obviously	294	14.1%	1,778	85.4%	10	0.5%	2,082	
of course	295	12.9%	1,966	86.1%	23	1.0%	2,284	
no doubt	40	5.1%	735	94.5%	3	0.4%	778	
undoubtedly	599	93.9%	39	6.1%	0	0	638	
Total	**1,420**		**10,325**		**45**		**11,790**	**34.7%**
Percentage	**12%**		**87.6%**		**0.4%**			
PS								
likely	88	3.5%	2,391	96.5%	0	0	2,479	
maybe	329	13.7%	2,066	86.0%	7	0.3%	2,402	
perhaps	317	12.6%	2,203	87.3%	4	0.2%	2,524	
possibly	70	4.8%	1,389	95.2%	0	0	1,459	
probably	342	2.8%	11,977	97.2%	5	0.04%	12,324	
Total	**1,146**		**20,026**		**16**		**21,188**	**62.4%**
Percentage	**5.4%**		**94.5%**		**0.1%**			
PSNC								
likely not	0	0	41	100%	0	0	41	
maybe not	9	18%	40	80%	1	2%	50	
perhaps not	18	58.1%	12	38.7%	1	3.2%	31	
possibly not	3	25%	9	75%	0	0	12	
probably not	5	2.4%	198	96.6%	2	1%	205	
Total	**35**		**300**		**4**		**339**	**1%**
Percentage	**10.3%**		**88.5%**		**1.2%**			
HCSNC								
certainly not	3	1%	309	98.4%	2	0.6%	314	
definitely not	3	1.4%	207	98.1%	1	0.5%	211	
obviously not	7	9.6%	73	91.3%	0	0	80	
of course not	2	28.6%	7	77.8%	0	0	9	
Total	**15**		**596**		**3**		**614**	**1.8%**
Percentage	**2.4%**		**97.1%**		**0.5%**			
Grand Total	**2,616**	**7.7%**	**31,247**	**92.1%**	**68**	**0.2%**	**33,931**	**100%**

Overall, the distribution of MACs across clausal positions in dependent clauses is similar to what was observed for independent clauses. The PS MAC group is the most common (62.4%) followed by HCS MACs (34.7%), in all positions. The difference between English PS and HCS MACs is more pronounced than was observed in independent clauses.

The two negative groups are again infrequent overall. We see, then that the patterns of occurrence of English MACs are more or less similar in both dependent and independent clauses. That is, PS MACs are the most frequent group, followed by HCS MACs; both PSNC and HCSNC MACs are rare.

5.5 Clause level positions of Urdu MACs in dependent clauses

In this section, I present the frequencies with which Urdu MACs occur in initial, medial, and final positions in dependent clauses identified according to the methods outlined in section 4.3.5.

Table 5.4: Positions of Urdu MACs occurring in dependent clauses in LUWC

MACs	Initial position		Medial position		Final position		Total	Overall percent
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
beśak 'undoubtedly'	85	66.4%	42	32.8%	1	0.8%	128	
bilāśubha 'of course'	60	55.6%	48	44.4%	0		108	
koī śubha nahīṁ 'no doubt'	3	16.7%	7	38.9%	8	44.4%	18	
śak nahīṁ 'no doubt'	35	50.7%	18	26.1%	16	23.2%	69	
yaqīnān 'certainly'	469	64.4%	251	34.5%	8	1.1%	728	
yaqīnī tor par 'certainly'	28	50.0%	27	48.2%	1	1.8%	56	
żarūr 'definitely'	724	42.7%	926	54.6%	47	2.8%	1697	
Total	**1,404**		**1,319**		**81**		**2,804**	**45.3%**
Percentage	**50.1%**		**47.0%**		**2.9%**		**165.2%**	
PS								
śāyad 'perhaps'	1,344	56.3%	1,024	42.9%	20	0.8%	2,388	
hō saktā 'maybe'	152	22.0%	81	11.7%	459	66.3%	692	
ġālibān 'probably'	142	63.7%	73	32.7%	8	3.6%	223	
Total	**1,638**		**1,178**		**487**		**3,303**	**53.4%**
Percentage	**49.6%**		**35.7%**		**14.7%**		**100%**	
PSNC								
śāyad nahīṁ 'perhaps not'	13	65.0%	2	10.0%	5	25.0%	20	
hō nahīṁ saktā 'possibly not'	6	11.5%	4	7.7%	42	80.8%	52	
Total	**19**		**6**		**47**		**72**	**1.2%**
Percentage	**26.4%**		**8.3%**		**65.3%**		**100%**	
HCSNC								
beśak nahīṁ 'undoubtedly not'	0		14	100.0%	0	0	14	
yaqīnān nahīṁ 'definitely not'	6	66.7%	2	22.2%	1	11.1%	9	
yaqīnī tōr *par nahīṁ* 'certainly not'	2	100.0%	0		0	0.0%	2	
żarūr nahīṁ 'definitely not'	0		1	100.0%	0	0.0%	1	
Total	**8**		**3**		**1**		**12**	**0.2%**
Percentage	**66.7%**		**25.0%**		**8.3%**		**100%**	
Grand Total	**3,069**	**50%**	**2,506**	**40%**	**616**	**10%**	**6,191**	**100%**

Overall, the distribution across clausal positions in dependent clauses is similar to what was observed for independent clauses. The PS group of MACs is the most common (53.4%), followed by HCS MACs (45.3%). This difference in frequencies is less pronounced than that observed between these groups in independent clauses. Otherwise, the patterns for Urdu and English are more or less the same across MACs occurring in independent and dependent clauses. The two negative groups are rare, just as in independent clauses.

Overall, we see that the patterns of occurrence of both English and Urdu MACs in independent and dependent clauses across clause initial, medial, and final positions are rather similar. In both languages, PS MACs are more frequent than HCS MACs in both independent and dependent clauses. One difference is that, in English, the prominence of PS MACs compared to HCS MACs is pronounced in dependent clauses, but less marked in independent clauses. In contrast, in Urdu PS MACs are relatively more frequent in independent clauses than HCS MACs, whereas there is no marked difference between these groups in dependent clauses. HCSNC and PSNC MACs in both languages have approximately similar frequencies in the respective data (very small in all types of clause). Therefore, subsequent analysis, both quantitative and qualitative, will henceforth primarily be focused on use of English and Urdu MACs in different clause positions, without any distinction being made so as to treat independent and dependent clauses separately.

5.6 Combined data for clausal positions of English and Urdu MACS

Tables 5.5 and 5.6 give combined frequencies for the clausal positions of the English and Urdu MACs respectively, that is, the frequencies observed for each group in initial,

medial and final positions if we ignore the distinction between independent and dependent clauses. Thus, Table 5.5 combines the figures in Tables 5.1 and 5.3 and Table 5.6 combines the figures in Tables 5.2 and 5.4. Alongside the tabulated data, I present figures of the same data in graphical form. This data visualisation is provided for the sake of those readers who may prefer a visual overview of the relative magnitude of the various values to the precise reading of the actual values supplied by the tables.

Table 5.5: Positional distribution of English MACs in ECC

MACs	Initial position		Medial position		Final position		Total	Overall percent
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
certainly	1,255	15%	6,997	84%	90	1%	8,342	
definitely	1,212	15%	6,658	83%	188	2%	8,058	
obviously	2,721	39%	3,905	56%	313	5%	6,939	
of course	4,567	46%	4,399	44%	995	10%	9,961	
no doubt	328	13%	2,004	82%	113	5%	2,445	
undoubtedly	676	79%	181	21%	1	0%	858	
Total	**10,759**		**24,144**		**1,700**		**36,603**	**39%**
Percentage	**29%**		**66%**		**5%**		**100%**	
PS								
likely	257	4%	5,892	96%	0		6,149	
maybe	6,986	54%	3,131	24%	2,782	22%	12,899	
perhaps	4,461	47%	2,795	30%	2,170	23%	9,426	
possibly	1,439	29%	1,687	34%	1,770	36%	4,896	
probably	2,760	13%	14,319	69%	3,602	17%	20,681	
Total	**15,903**		**27,824**		**10,324**		**54,051**	
Percentage	**29%**		**51%**		**19%**		**100%**	**57%**
PSNC								
likely not	4	2%	160	96%	3	2%	167	
maybe not	179	47%	146	38%	57	15%	382	
perhaps not	163	74%	38	17%	20	9%	221	
possibly not	33	59%	20	36%	3	5%	56	
probably not	339	34%	607	60%	64	6%	1,010	
Total	**718**		**971**		**147**		**1,836**	
Percentage	**39%**		**53%**		**8%**		**100%**	**2%**
HCSNC								
certainly not	223	24%	679	73%	24	3%	926	
definitely not	172	25%	508	73%	19	3%	699	
obviously not	77	26%	213	72%	6	2%	296	
of course not	95	63%	30	20%	25	17%	150	
undoubtedly not	0		1	100%	0		1	
Total	**567**		**1,431**		**74**	**2072**		
Percentage	**27%**		**69%**		**4%**	**100%**		**2%**
Grand Total	**27,947**	**30%**	**54,370**	**57%**	**12,245**	**13%**	**94,562**	

Table 5.6: Positional distribution of Urdu MACs in LUWC

MACs	Initial position		Medial position		Final position		Total	Overall percent
	Freq	Percent	Freq	Percent	Freq	Percent		
HCS								
beśak 'undoubtedly'	639	58.7%	372	34.2%	77	7.1%	1,088	
bilāśubha 'of course'	1,158	69.8%	486	29.3%	15	0.9%	1,659	
koī śubha nahīṁ 'no doubt'	11	5.5%	149	74.9%	39	19.6%	199	
śak nahīṁ 'no doubt'	47	10.8%	253	58.2%	135	31.0%	435	
yaqīnān 'certainly'	1,195	46.0%	1,374	52.9%	27	1.0%	2,596	
yaqīnī tor par 'certainly'	60	29.1%	134	65.0%	12	5.8%	206	
żarūr 'definitely'	2,294	27.9%	5,026	61.1%	900	10.9%	8,220	
Total	**5,404**		**7,794**		**1,205**		**14,403**	**37.9%**
Percent	**37.5%**		**54.1%**		**8.4%**			
PS								
śāyad 'perhaps'	6,185	46.5%	6,612	49.7%	503	3.8%	13,300	
hō saktā 'maybe'	922	10.9%	2,210	26.1%	5,327	63.0%	8,459	
ġālibān 'probably'	404	26.3%	1,108	72.0%	26	1.7%	1,538	
Total	**7,511**		**9,930**		**5,856**		**23,297**	**61.4%**
Percent	**32.2%**		**42.6%**		**25.1%**		**100.0%**	
PSNC								
śāyad nahīṁ 'perhaps not'	56	60.9%	24	26.1%	12	13.0%	92	
hō nahīṁ saktā 'possibly not'	7	7.1%	4	4.1%	87	88.8%	98	
ġālibān nahīṁ 'probably not'	2	20.0%	8	80.0%	0		10	
Total	**65**		**36**		**99**		**200**	**0.5%**
Percent	**32.5%**		**18.0%**		**49.5%**		**100.0%**	
HCSNC								
beśak nahīṁ 'undoubtedly not'	0		14	100%	**0**	0%	14	
bilāśubha nahīṁ 'of course not'	0		0		1	100.0%	1	
yaqīnān nahīṁ 'definitely not'	22	40.0%	19	34.5%	14	25.5%	55	
yaqīnī tōr par nahīṁ 'certainly not'	3	75.0%	1	25.0%	0		4	
żarūr nahīṁ 'definitely not'	0		3	1	0		3	
Total	**25**		**23**		**15**		**63**	
Percent	**39.7%**		**36.5%**		**23.8%**			
Grand Total	**13,005**	**34%**	**17,783**	**47%**	**7,175**	**19%**	**37,963**	**100%**

The cumulative data naturally reflects the same tendencies observed separately for independent and dependent clauses. In both English and Urdu, PS MACs are the most common group of MACs regardless of position, followed closely by HCS MACs. The PSNC and HCSNC groups are low in frequency in both languages; this is unsurprising as there are fewer of them, but they are individually rare as well as collectively, especially in Urdu. The frequencies in Tables 5.5 and 5.6 are visualised in figures 5.1 and 5.2 respectively, showing clearly the dominance of a small number of forms, and the infrequency of the negative groups, in both English and Urdu.

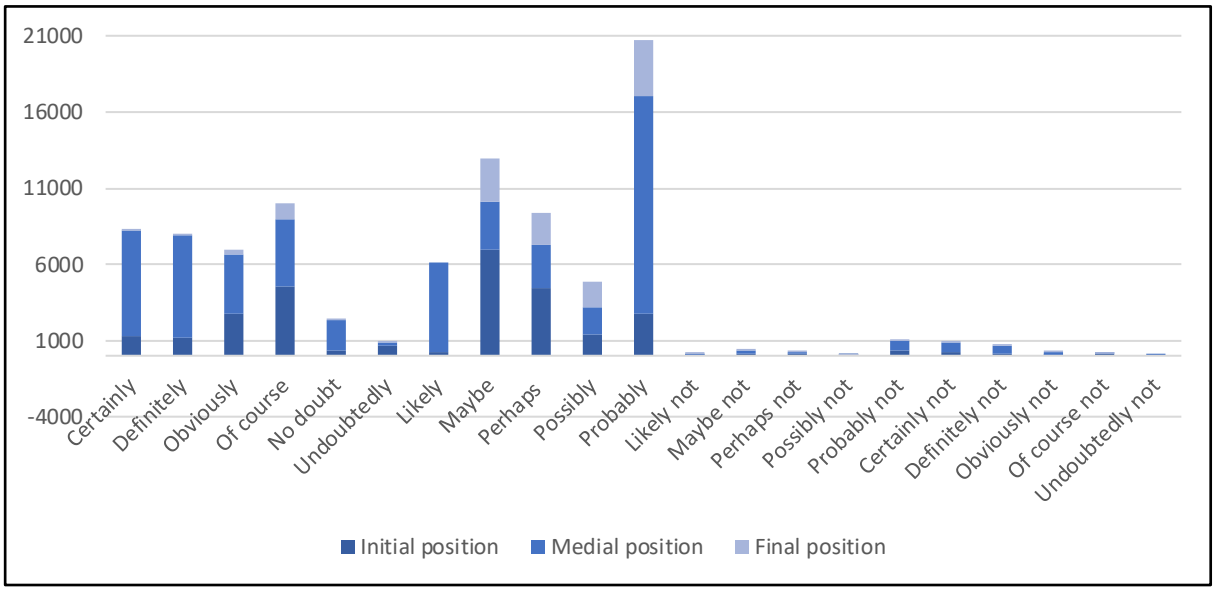


Figure 5.1: Distribution of English MACs across clause initial, medial, and final positions

In English, the PS MAC *probably* is the most frequent MAC followed by another PS MAC, *maybe*. HCS MAC *of course* is the third most frequent, then another PS MAC, *perhaps*. So, of the four most frequent MACs, only one is an HCS MAC. The HCS MACs

certainly, *definitely*, *obviously* and the PS MACs *likely* and *possibly* are also frequent, although less than the first four. PSNC and HCSNC MACs are rarely used.

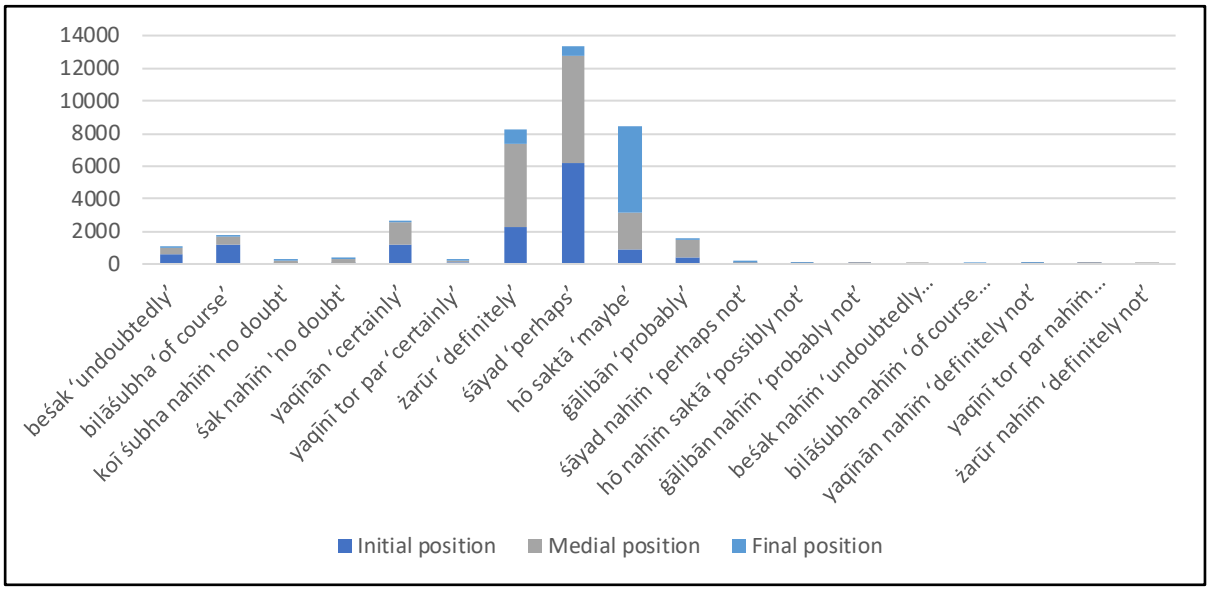


Figure 5.2: Distribution of Urdu MACs across clause initial, medial, and final positions

In Urdu, the PS MAC *śāyad* 'perhaps' is the most frequent; PS *hō saktā* 'maybe' and HCS *żarūr* 'definitely' are about equal in frequency, and thus approximately joint second in rank. Every individual MAC other than these three is either rare (*yaqīnān* 'certainly, *bilāśubha* 'of course', *beśak* 'undoubtedly', and *ġālibān* 'probably') or very rare (all the rest). Especially PSNC and HCSNC MACs are rarely used. This trend in the frequencies is not unexpected because word frequencies in language typically follow a distribution similar to Zipf's Law (1949), that is, a few elements occur very frequently, and many elements occur rarely, leading to a long-tailed distribution with most of the probability density in the tail.

Figure 5.3 shows the percentage rates at which each English MAC occurs in each position.

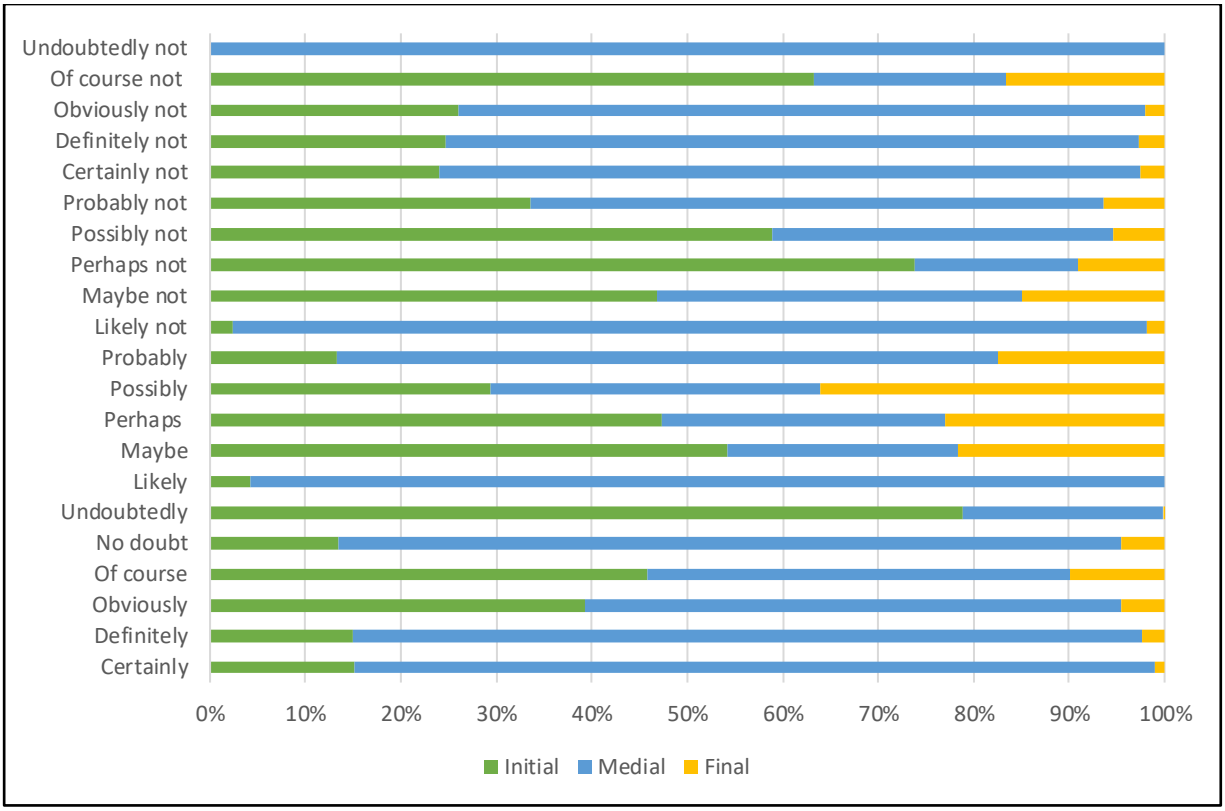


Figure 5.3: Percentage rates at which English MACs occur in each clausal position in ECC

Figure 5.3 shows that, although English MACs can occur at any clausal position, not every MAC is observed in every position or uniformly. For instance, *likely* occurs most frequently (96%) in medial position but never in final position; *undoubtedly* occurs only once in final position and predominantly (79%) in initial position. Conversely, *probably* is most frequent (69%) in medial position and *possibly* is most balanced as it occurs almost uniformly in all positions, final position being the most frequent by only a small margin (36%). The same tendency can be observed for the negative MAC groups. Like their positive counterparts, negative MACs do not occur uniformly in every position. For instance, *likely not* occurs predominantly in medial position (96%), but rarely in initial and final position. All the HCSNC MACs, except *of course not*, occur most commonly in medial position. *Of course not* is most frequent in clause initial position (63%).

Figure 5.4 shows the percentage rates at which each Urdu MAC occurs in each position.

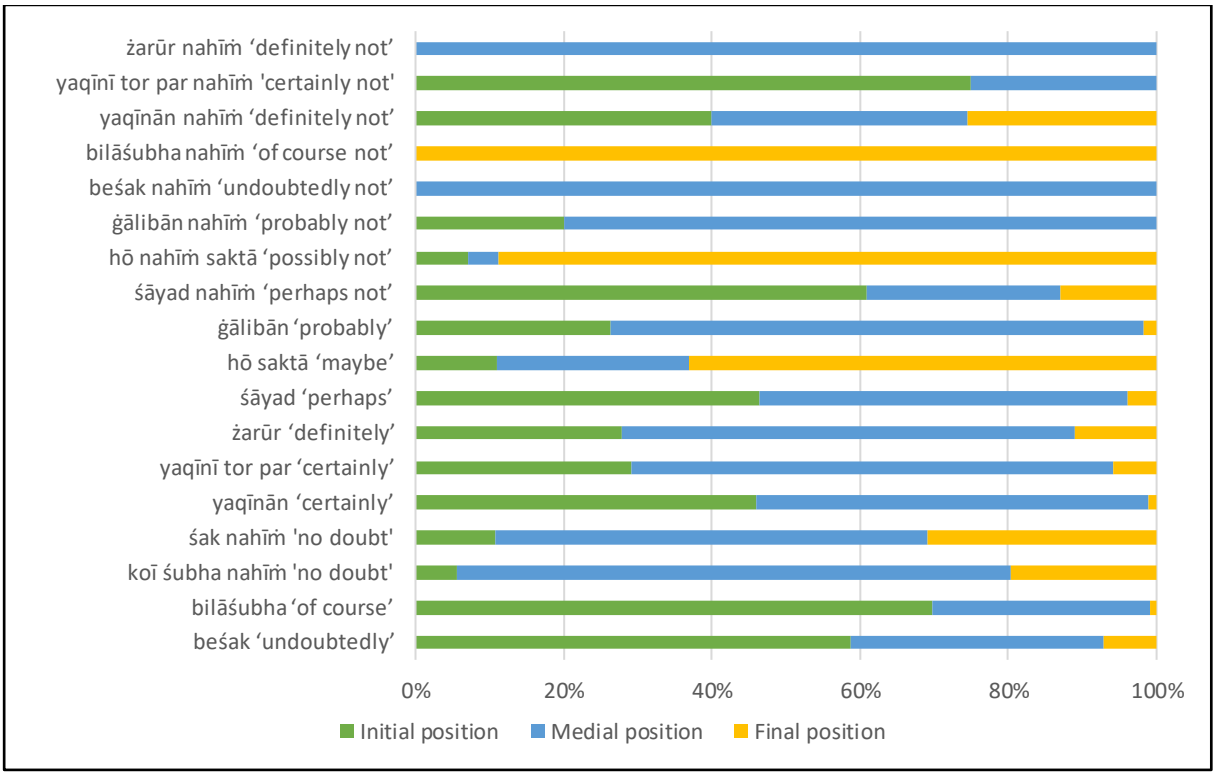


Figure 5.4: Percentage rates at which Urdu MACs occur in each clausal position in LUWC

Figure 5.4 shows that, like the English MACs, Urdu MACs can also occur in any clausal position, but not every MAC is observed in every position uniformly. For instance, *bilāśubha* ‘of course’ preferentially appears in initial (70%) and *yaqīnān* ‘certainly’ preferentially appears in medial position (53%) respectively but both rarely occur in final position (0.9% and 1% respectively). Meanwhile, *śāyad* ‘perhaps’ is also more frequent in initial and medial position (47% and 50% respectively) and are rare in final position (4%).

The negative MACs similarly tend to appear in all clausal positions, but not every negative MAC occurs in all three positions or uniformly. For instance, *hō nahīṁ saktā* ‘possibly not’ is observed in all three positions, but overwhelmingly (89%) in final position;

ġālibān nahīṁ 'probably not' occurs predominantly in medial position (80%) but never in final position. HCSNC MACs are very rare, and the few that do occur have a tendency to occur in a specific position only. For instance, *beśak nahīṁ* 'undoubtedly not' and *żarūr nahīṁ* 'definitely not' occur only in medial position; similarly *yaqīnī tor par nahīṁ* is most often in medial position (75%) and does not occur in final position.

Figures 5.5 and 5.6 present the frequencies of English and Urdu MACs proportionally for each of the three positions: initial, medial, and final positions. The data and the calculations are same but the data in the following figures is presented from the opposite perspective. That is, Figures 5.3.and 5.4 show the preferences of MACs to occur in three clausal positions, whereas Figures 5.5. and 5.6 show the distribution of MAC types at each position.

Figure 5.5: Clause position-wise distribution of English MACs in ECC

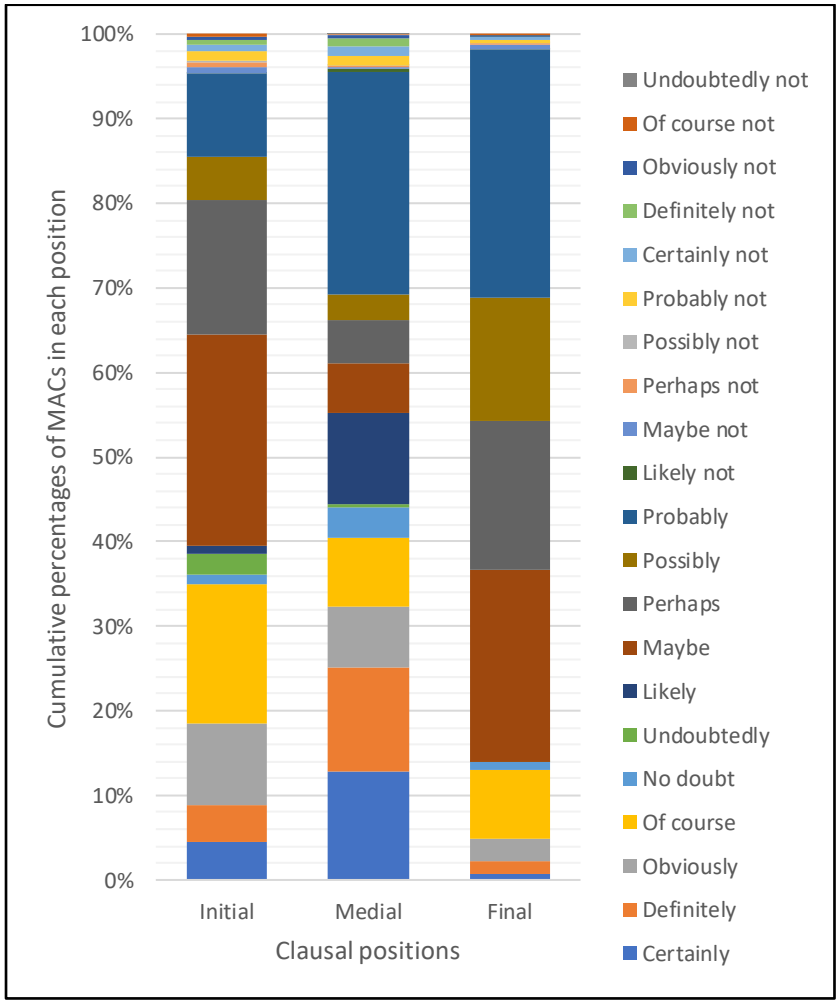


Figure 5.6: Clause position-wise distribution of Urdu MACs in LUWC

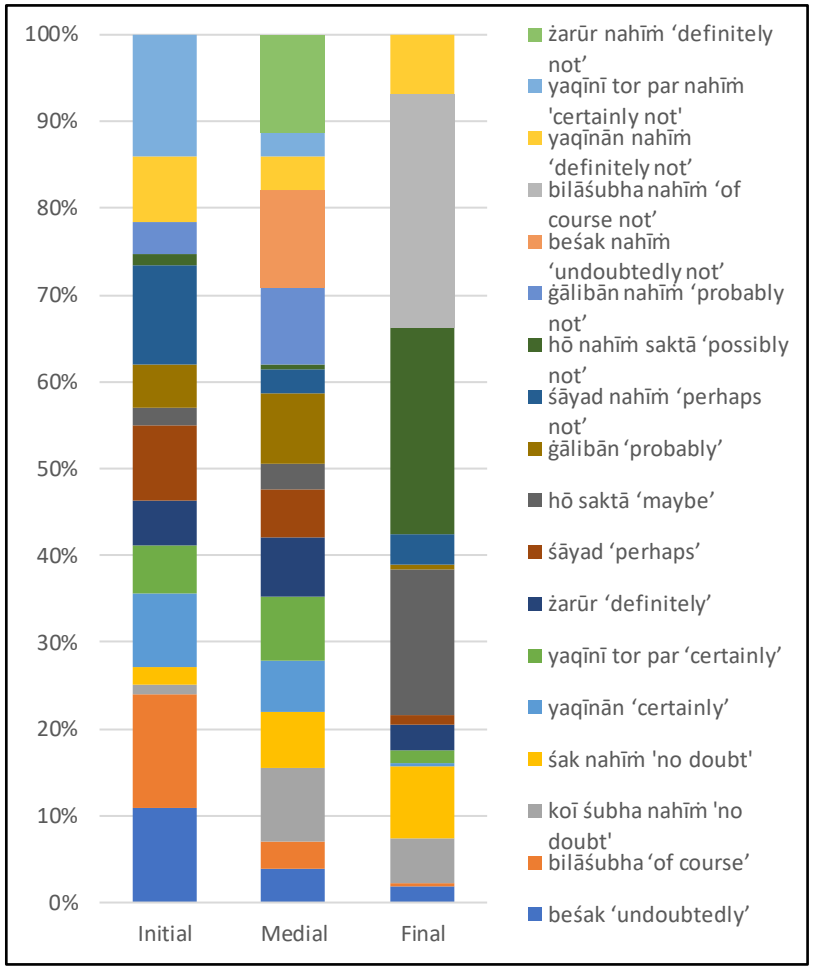


Figure 5.5 shows that in initial position, the most frequent English MACs are *undoubtedly* (79%) followed by *maybe* (54%) and *perhaps* (47%). In medial position, *likely* (96%) is the most common, followed in joint second position by *certainly* (84%) and *definitely* (83%). In final position, *possibly* (36%) is the most frequent, followed in joint second position by *perhaps* (23%) and *maybe* (22%). *Undoubtedly* occurs only once and in medial position. *Likely not* is the most frequent negative MAC in medial position (again 96%), followed jointly by *certainly not* and *definitely not* (73% each). Negative MACs do not occur commonly in the final position. The most frequent at that position is *of course not* (17%) followed by *maybe not* (15%).

Figure 5.6 shows that the most frequent MAC in initial position is *bilāśubha* 'of course' (70%) followed jointly in second position by *śāyad* 'perhaps' (47 %) and *yaqīnān* 'certainly' (46%). In medial position, the most frequent MAC is *koī śubha nahīṁ* (75 %) and the second most frequent MAC is *ġālibān* 'probably' (72 %). In final position, *hō saktā* 'maybe' is most frequent (63 %) and rest of the MACs are rare in this position.

Figures 5.5 and 5.6 also show that, although the semantic types of HCS, PS, PSNC, and HCSNC are not evenly distributed across the three clause positions, each type may be found across all three positions.

5.7 Chapter summary

The most frequently used group of MACs, in both English and Urdu, is PS, followed by HCS. PSNC and HCSNC MACs are substantially less frequent. Interestingly, in both English and Urdu, the set of the three most frequent MACs consists of one HCS MAC and two PS MAC. In each case, HCS *of course* and *żarūr* 'definitely' appear in any of the three

positions without any overly preferred position, a characteristic typical of adverbs in general. However, none of the MACs in English and Urdu appear in all three positions uniformly.

Therefore, on the whole, the analysis in this chapter has shown that most MACs are used in all three clausal positions, with few exceptions. Interestingly, all semantic types of English MACs occur readily in clause medial position. Similarly, most of the Urdu HCS and PS MACs preferably occur in clause medial position. On the other hand, separately, the three Urdu PSNC MACs have different preferred clause positions. However, of the three PSNC MACs, *hō nahīṁ saktā* 'possibly not' has the highest value in the data. As *hō nahīṁ saktā* 'possibly not' occurs most commonly in clause final position, it has driven higher the combined value of PSNC MACs in the clause final position. In fact, the data indicates that both *hō saktā* 'maybe' and *hō nahīṁ saktā* 'maybe not' occur more readily in clause final position than in clause initial and medial positions. This preference for clause final position can be accounted for by the default final positioning of both *hō saktā* 'maybe' (63%) and *hō nahīṁ saktā* 'maybe not' (89%) consisting, as they do of finite verbs. The further discussion on *hō saktā* 'maybe' and *hō nahīṁ saktā* 'maybe not' preference for final position is given in chapter 6. The HCSNC MACs are the least frequent of all in the Urdu data; the few that were found occur mostly in one specific position, that is, in initial, medial or final position.

Therefore in both languages, the difference in occurrence within clause is mainly in preferential positioning of MACs. Across the next chapters, the semantic and pragmatic features of MACs, which are the most likely drivers of such differences, are discussed in detail.

6 Modal semantics of the English and Urdu MACs

6.1 Chapter overview

This chapter focuses on the first two parts of RQ 3, namely, how do the clausal positions of modal adverbs of certainty (MACs) influence the meaning of an element or a clause, and what modal scope do MACs have because of their placement in a clause? A semantic *scope* is a range of influence that an element (here a MACs) has over other elements in a clause due to their meaning (see 2.4.2.1). The previous chapter established that MACs do occur in various positions in the clause, though they are not evenly distributed across the three clause positions (see 5.6). The semantic categories analysed follow the framework given in section 2.6 and based on my review of the literature: HCS (high certainty support), PS (probability support), PSNC (probability support for negative content), and HCSNC (high certainty support for negative content). The existing literature (see 2.3.2.2 and 2.3.2.3) establishes that the core function of HCS MACs is to express the addresser's *certainty*, *conviction*, and *confidence* in a proposition; the core function of PS MACs is to express the addresser's *low-certainty* (probability), *doubt* and *low-confidence* in a proposition. The following analysis is an attempt to discuss English and Urdu MACs' semantic function as certainty and probability markers, and other supplementary semantic functions, in three clausal positions and in certain environments. The analysis will help in understanding the cross-linguistic similarities and differences in English and Urdu MACs' usage and functions.

To this end, in sections 6.2 to 6.4, I discuss clause placement-related functions of MACs according to the semantic categories that I have worked with already. Then, in section 6.5, I discuss the scope of MACs over negation specifically. I then address the interaction of MACs with four kinds of element that both the existing literature, and corpus evidence,

shows them to frequently cooccur with. In section 6.6, I discuss the interaction of MACs with modal verbs (MVs). In section 6.7, I discuss the meaning conveyed by the interactions between different MACs occurring nearby to one another. In sections 6.8 and 6.9, I discuss the influence of MACs in interrogative and conditional sentences respectively. Finally, in section 6.10, I summarise this part of the analysis.

6.2 The semantic scope of HCS and PS MACs occurring in clause initial position

The concordance analysis shows that, when English and Urdu MACs are in clause initial position, they have scope over not only the immediate element in the clause but also the whole proposition. In this position, MACs are used in three main functions, two being the core functions of *certainty* and *probability markers*. These functions are discussed one by one in this part.

6.2.1 Certainty marker

The first function that MACs have in clause initial position is as *certainty markers* (see 2.5.4). As a certainty marker, in clause initial position an HCS MAC acts as an emphasiser or intensifier to strengthen the proposition (i.e. to reinforce the truth value of the statement). For instance, when an HCS MAC occurs in initial position in a dependent clause (e.g. *because*) that gives a reason as an explanation for what is said in the independent clause. The HCS MAC may express the degree of certainty the addresser has in the truth of the proposition presented in the dependent clause. Example (63) illustrates this.

63. I don't want to say anything now *because of course* it is too early to say stuff like that" (ECC_ind_b909659).

Similarly, example (64) illustrates that when an HCS MAC occurs in initial position in a dependent clause (e.g. relative clause) that describes a consequence of circumstances set forth in the independent clause, that HCS MAC may express the degree of certainty the addresser has in the truth of the proposition presented in the dependent clause.

64. Bilāśubha Pānāma līks=par jārī.o.sārī tūfān kissī
No.doubt Panama leak.PL=on ongoing storm any

had tak hakūmat=kī tuwajha mēgā taraqīātī
limit till government=GEN.F.SG attention mega development

mansūbōṁ=sē hatā rahā hai *jō* *ke*
plan.PL.OBL=from remove PROG.PFV.M.SG be.PRS.3.SG *REL* *that*

yaqīnān mulk=kē līē sūdmund nahīṁ hai.
definitely country=GEN.M.PL for beneficial NEG be.PRS.3.SG.

"No doubt the ongoing storm over the Panama leaks has more or less diverted government attention from the mega development plans, *which* is *definitely* not beneficial for the country" (LUWC_j0099687).

In some instances, a MAC is added in clause initial position after some particular clause marker to expresses the addresser's confidence in the truth value of the proposition. For instance, when an HCS MAC occurs in initial position after a contrastive coordinating clause-marker (i.e. English *but*, Urdu *lēkin* 'but'), it intensifies the contrast between low certainty, or negation, in the clause before and high certainty in the clause following the clause marker (i.e. *not/perhaps/probably X but of course/definitely/undoubtedly Y*), as in (65) and (66).

65. "But if you go back through the current prime minister's history, he's often said quite striking things. And he never apologises. So, Boris might have done this anyway, *but certainly*, having watched Trump in action, he wouldn't have been put off" (ECC_ind_b404416).

66. Mūmkīn hai aisā hūā hō … *lēkin*
Probable be.PRS.3.SG such.M.SG be.PFV.M.SG be … *but*

yaqīnān yeh fitnabāz haīṁ.
definitely DEM malevolent be.PRS.3.PL.

"It's probable such a thing happened … *but* they *definitely* are malevolent"

(LUWC_f199p0068715).

6.2.2 Probability marker

The second function in clause initial position is expressing *probability* in the truth of a proposition by adding a PS MAC. For instance, adding a PS MAC in clause initial position of a dependent clause expresses the addresser's lower degree of certainty in the circumstances mentioned in the dependent clause, as in examples (67) and (68).

67. "During this period there was a recognised failure to reduce the level of nitrogen dioxide to within the limits set by EU and domestic law, *which possibly* contributed to her death" (ECC_mail9063237).

68. Ab ākhirkār asembalī=nē maśrūṭ pāvar
Now finally assembly=ERG conditional power

dēnē kā ‘andīyah dīyā hai *jō* *ke*
give.INF.OBL GEN.M.SG opinion give.PFV.M.SG be.PRS.3.SG *REL* *that*

śāyad rēnjarz=kē liē qābil-e-qabūl nahīṁ
perhaps rangers=GEN.M.PL.OBL for acceptable NEG

hō gā.
be FUT.M.SG.

"Now the Assembly has finally given an indication of giving conditional power [to the Rangers] *that will perhaps* not be acceptable to the Rangers" (LUWC_j0023054).

When a PS MAC occurs in initial position after a clause marker, it lowers the certainty in contrast to a higher degree of certainty in the other clause (see Simon-Vandenbergen & Aijmer, 2007, p.31). For example, if a PS MAC occurs after *but*, it does not intensify the element occurring immediately after it, rather it downtones the possibility of the proposition being true, in contrast to the preceding statement. In example (69) the following clause is additionally marked by an *or*-extender (a structure defined by Aijmer, 2013, p. 139-140 as an extension of a sentence with a clause beginning with *or*, that may convey vagueness, function as a hedge, or qualify something by conditions or exceptions). In (69), the addresser describes many people's initial assessment that Covid would end soon as having proven foolish. Then, in the following statement, the addresser hedges by introducing another possibility, albeit one downtoned by the PS MAC *perhaps*.

69. "This was mainly because everyone thought the pandemic would be all over in a matter of weeks, and that life would soon return to normal. It would be like one of those silly bird-flu scares, over before your first mini-bottle of hand sanitizer had run out. How foolish that seems now, *but perhaps* Covid has made fools of us all - or at least made us understand what is and is not important" (ECC_mail8821689).

Similarly, in Urdu a PS MAC following *lēkin* 'but' non-assertively presents the clause's proposition in contrast to the relative assertiveness of the prior clause.

In (70), like the English example (69), the contrasting assertive statement is followed by an *or*-extender.

70. *Hō* *saktā* *hai* ke kisī ēk āītum=kī
Be *can.IPFV.M.SG* *be.PRS.3.SG* that some one item=GEN.F.SG

qīmat hō *yā* mārkit mēkānizum=kī
price be.SBJV.3.SG *or* market mechanism=GEN.F.SG

vajah=sē kuch aśiyā=kī qīmatēṁ
reason=by some commodity.PL=GEN.F.SG price.PL

kum huē hōṁ.
reduce be.PFV.F.SG be.SBJV.3.PL.

"*Maybe* it is the price of some single item *or* due to market mechanisms there is a reduction in prices of some commodities" (LUWC_e0378576).

The use of a PS MAC to mark a non-assertive contrast to an assertive statement is also observed in both languages with clauses other than those followed by an *or*-extender. Moreover, the clause with the PS MAC may precede or follow the clause with the assertively claimed proposition.

6.2.3 Short response

The third function of MACs in clause initial position is as a short response to a statement by a previous addresser (see Table 2.2). Typically, an English HCS MAC of this sort is used to convey a positive response to what the prior addresser said, as shown in (71).

71. A: "My family come from Shrewsbury, and my gran was always very vocal that it had to Shroo, never Shrow."

B: "Yes, *of course*" (ECC_AskUK201909).

Conversely, an English HCSNC MAC is typically used to convey a negative response to what a previous addresser said, as in (72).

72. A: "Are people without degrees now left behind?"

B: "*Certainly not*. The trader (plumbers, electrician etc.) can all get a good income, as well as other jobs like TT and Management often not having formal qualification requirements" (ECC_AskUK202001).

On the other hand, LUWC shows that in Urdu, HCS MACs are typically used as a positive response, as in (73). Those that are in short negative responses are most commonly addressers answering their own rhetorical question (see further section 6.5.3).

73. A: Ṭarīqa-e-kār badalnē=mēṁ thōṛī buhat tabdīlī

tabdīlī.method change.INF.OBL=in less more change

tō āē gē. Agar voh mofīd
EMPH come.PFV.M.PL FUT.M.PL. But DEM beneficial

sābit hōtī hai tō buhat acchī
prove be.IPFV.F.SG be.PRS.3.SG then very good.F

bāt hai.
thing be.PRS.3.SG.

B: Jī *bēśak*.
Yes *no.doubt*.

A: "A little bit of change will come through changing the method. But if that proves beneficial then it will be a very good thing"

B: "Yes *no doubt*" (LUWC_f059p0057162).

Use of a PS MAC as a short response reflects the addresser's lack of certainty. In English, a PS MAC as a short response is usually followed by a possible alternative assertion, as in (74).

74. A: "Better to be safe than sorry imo, but as I said OPs landlord is probably just saying this for insurance reasons.

B: "*Perhaps*. I'd assume you'd need some sort of identification or utility bills etc." (ECC_AskUK202001).

In Urdu too PS MACs are used as short responses. Typically, the second addresser replies with a PS MAC to convey that they think it a possibility that the first addresser's assertion is true, as in (75).

75. A:

Agar	āp	lōg	mērā	munh
If	you.PL	people	1.SG.POSS.M.SG	face

dēkh	saktē	tō	mujhē	batānē=kī
see	can.IPFV.M.PL	then	1.SG.ACC	tell.INF.OBL=GEN.F.SG

zarūrat	kabhī	paēś	nā	ātī
need	ever	requirement	NEG	come.IPFV.F.SG

B:

Hāṁ	*hō*	*saktā*	*hai*
Yes	*be*	*can.IPFV.M.SG*	*be.PRS.3.SG.*

A: "If you people could see my face then I wouldn't have needed to tell [you about it]".

B: "Yes *maybe*" (LUWC_f029p0017829).

An HCS MAC as a short response may also be used to confirm with certainty something the previous addresser asked about, as in (76) and (77).

76. A: "Are you able to think for yourself and make independent decisions?"

B: "Yes, *of course*" (ECC_AskUK201305_44).

77. A: Bīvī bhī kyā hukam hākim=mēṁ ātī
Wife INC Q order royal=in come.IPFV.F.SG

hai?
be.PRS.3.SG?

B: *Yaqīnān*.
Definitely.

A: "Does a wife count as royalty?

B: "*Definitely*" (LUWC_f014p0009317).

On the other hand, in both languages, a PS MAC can be used to express a noncommittal response, as in (78) and (79).

78. A: "On land it's meant to be flag rather than jack yes".

B: "Not according to the admiralty or the government. Our national flag is the union jack"

A: "*Perhaps*" (ECC_AskUK202005).

79. A: *Hō* *saktā* *hai* unhēṁ hamārē
Be *can.IPFV.M.SG* *be.PRES.3.SG* DEM.PL.ACC.OBL 1.PL.POSS.OBL

khānē āchchē na lagtē hōṁ.
food.M.OBL.PL good. M.OBL.PL NEG feel.IPFV.M.PL be.SBJV.3.PL.

Ab dekhēṁ na bāz log salāis=per
Now see.SBJV.2.PL NEG some people slice=on

machlī=kē andē lagā kar buhat
fish=GEN.M.PL.OBL egg.PL put.PFV.M.SG do much

śōq=sē khātē haiṁ.
delight=with eat.IPFV.M.PL be.PRS.3.PL.

B: Hāṁ *śāyad*
Yes *perhaps*

A: "*Probably* they don't like our dishes. Now listen, some people enjoy eating fish eggs on toast."

B: "Yes *perhaps*" (LUWC_f066p0011526).

6.3 The semantic scope of HCS and PS MACs occurring in clause medial position

Medial is the most frequent position for HCS and PS MACs in my corpora (see 5.6). The concordance analysis shows that the primary function of both English and Urdu MACs in medial position is to express an unambiguous degree of emphasis. When a MAC occurs in medial position, it intensifies, focuses, and emphasises modal values (Hoye, 1997, p. 150). Therefore, in this position MACs are used to emphasise the strength of the addresser's commitment to the proposition (Hoye, 1997, p. 121). Simon-Vandenbergen & Aijmer (2007, p. 86-87) say that, with few exceptions, MACs in medial position have the whole proposition in scope. That is, by use of an HCS MAC in medial position, the addresser *emphasises* the truth value of the content of the proposition as a certainty; by using a PS MAC in medial

position the addresser *downtones* the truth value of the content of the proposition as a probability.

By default, then, in clause medial position HCS MACs function as certainty markers or *emphasisers* expressing confidence in, and PS MACs as probability markers or *downtoners* expressing doubt about, the truth of the proposition. Both emphasising and downtoning functions are discussed in this section.

6.3.1 Certainty marker

When an HCS MAC in medial position immediately follows the first obligatory element of the clause but precedes the last obligatory element of the clause, then that MAC has scope over the whole clause and expresses the (high) level of certainty ascribed by the addresser to the proposition. Examples of English HCS MACs in medial position immediately after a noun phrase, like (80), illustrate their function as certainty markers in this context. This use of the HCS MAC serves to underline the addresser's stance on the "leading scientist's" sexist remarks quoted in the preceding sentence.

80. "Just last week, a leading scientist presented a talk claiming that "physics was invented and built by men". He *obviously* chose to ignore contributions from Marie Curie, Lise Meitner and Chien-Shiung Wu" (ECC_ind_a8557416).

In (81), an HCS MAC in medial position occurs immediately after a postposition phrase; this example illustrates the function of an Urdu HCS MAC as an emphasiser in this context.

81. Is fēslē=sē *yaqīnān* awām=kē ġālab
DEM.SG.OBL decision.OBL=from *certainly* public=GEN.M.PL.OBL majority

hissē=kō rēlīf millē gā tāham yeh
part=ACC relief receive.SBJV.3.PL FUT.M.SG although DEM

rēlīf us waqt hī zyāda fāedamand hō
relief DEM.SG.OBL time EXC more benefit be.SBJV.3.SG

gā jab trānspōrṭ=kē karāyōṁ=mēṁ
FUT.M.SG when transport=GEN.M.PL.OBL fare.M.PL.OBL=in

munāsib kamī hō gī.
suitable decrease be SBJV.3.SG FUT.F.SG.

"This decision will *certainly* come as a relief to the majority of the public although this relief will be more beneficial once there is a suitable decrease in transport fares" (LUWC_e0299785).

Typically, the natural location of a MAC intended to cover the whole clause is right before the main verb, because this verb supplies the core element of the proposition, the state or action. For instance, in example (82) an HCS MAC occurs right before *delivered* to focus on it and to emphasise the addresser's certainty in the truth of the proposition in the clause.

82. "He *certainly delivered* on his first campaign promise to 'deconstruct the administrative state'" (ECC_mail8919703).

Similarly, in (83), the addresser adds an HCS MAC to focus on the conjunct verb (defined by Koul, 2008, p.101, as a verbal construction that consists of a noun or adjective and a verb) *nasb kīyē* to express confidence in their proposition that the action described will take place.

83. Sī sī ṭī vī kaimrē *zarūr* *nasb* *kīyē*
C C T V camera.PL *definitely* *installation* *do.PFV.PL.OBL*

jaīṁ gē lēkin kaimrē vōṭ kāsṭ karnē
go.SBJV.3.PL FUT.M.PL but camera.PL vote cast do.INF.OBL

vālē maqam=par nahīṁ hōṁ gē.
VALA.M.SG.OBL place=on NEG be.SBJV.3.PL FUT.M.PL.

"CCTV cameras will *definitely* be *installed* but cameras will not be in the voting area" (LUWC_e0348917).

6.3.2 Probability marker

When PS MACs occur in medial position right after the first element of the clause, they have scope over the whole clause and function as downtoners (i.e. probability markers) of the content of the proposition. In example (84) the addresser adds a PS MAC right after the subject *answer* to downtone the certainty of the inference expressed in their proposition.

84. "His answer *probably* came as a revelation to many" (ECC_gdn_commentisfree_2020_dec_03).

In (85), the PS MAC occurs *within* the subject of the clause (between two pre-modifying postposition phrases) and has scope over *sub sē ēham aur būnyādī kām*. This addresser's expressed lower certainty applies to the modification of the noun *kām* 'work': the point made to be tentative is that this work was Dr Qazi's most important and fundamental contribution, not that Dr Qazi's work was analysis and dissemination of Muslim philosophy (the latter being the content of the clause predicate). The addresser's hedge allows for the possibility that of all the works that Dr Qazi did, his greatest contribution was something other than this; it does not allow for the possibility that he did not do this.

85. *Ḍākṭar* *Qāzī* *Jāvēd=kā* *śāyad* *sub=sē*
Doctor *Qazi* *Javed=GEN.M.SG* *possibly* *all=from*

ēham *aur* *būnyādī* *kām* bar-e-āzīm=mēṁ mūslim
important *and* *fundamental* *work* subcontinent=in Muslim

fikr=kā tajzīā aur us=kī
philosophy=GEN.M.SG analysis and DEM.SG.OBL=GEN.F.SG

taśrīh thā.
dissemination be.PST.3.SG

"*Perhaps Doctor Qazi Javed's most important and fundamental work of all* was the analysis and dissemination of Muslim philosophy in the subcontinent" (LUWC_j0852024).

When English or Urdu PS MACs occur immediately before the verb, they have scope over the clause due to focusing on the core element of the proposition, that is the main verb. In such instances, use of a PS MAC lowers the certainty of the action or state that the verb represents. In example (86), the use of PS MACs *perhaps* and *likely* marks a contrast between the main clause assertion and the dependent clause assertions. The placement of a dependent clause at the beginning of the sentence puts focus on the topic of the sentence, *people getting mild infections*, and excludes other possibilities (i.e. people getting non-mild infections). Then the addresser adds PS MAC *perhaps* in a location where it has scope over the second of the two coordinated restrictive relative clauses, and *likely* in the main clause, to express the lower possibility of the main clause proposition.

86. "For those who would only ever get a *mild infection, or perhaps experience no symptoms at all* - such as the young - it's *likely* that they *may not get infected* at all" (ECC_mail8910979).

In Urdu example (87), the PS MAC placed immediately before the main verb *dēkh* 'see' lowers the certainty of the whole proposition.

87. Āp=mēṁ=sē kuch dōstōṁ=nē *śāyad* *dēkh*
You=in=from some friend.PL.OBL=ERG *perhaps* *see*

līyā *hō* ke śūmārīyāt=kē safē=par
take.PFV.M.SG *be.SBJV.3.SG* that statistics=GEN.M.SG.OBL page=on

cand aur ādāt-o-śūmār mōjūd haiṁ.
few more digit-and-addition present be.PRS.3.PL.

"Some friends among you *have perhaps seen* that on the statistics page there are a few more calculations" (LUWC_f020p0000647).

6.4 The semantic scope of HCS and PS MACs occurring in clause final position

Clause final position is the least typical for both English and Urdu MACs (see 5.6). Again, the corpus shows that in clause final position, the core function of MACs is to express an addresser's degree of confidence, by either emphasising (using an HCS MAC) or downtoning (using a PS MAC) the degree of certainty in the proposition being true. The concordance shows that in both languages, MACs in final position typically have scope over the whole clause. However, in some instances, a MAC in final position may have scope only over the element immediately preceding it instead of the whole clause. Furthermore, regardless of whether they have scope over the whole clause or over some specific element, MACs are sometimes also used for *tagging* (see Table 2.2). These will be discussed in turn.

6.4.1 Certainty marker

In clause final position, HCS MACs foreground certainty in the contents of the proposition and have scope over the whole clause. By using an HCS MAC, the addresser expresses certainty in the proposition and links it to previous information, as in (88) and (89).

In example (88), the addresser restricts their certainty to three specific types of meats and thereby asserts authority on that particular information only.

88. "Tesco meat has way, way more water in it than Sainsbury's. True for mince, chicken and bacon, *certainly*" (ECC_AskUK202005).

In (89), the addresser infers that another chat group member must be drinking tea whilst checking social media (in reference to two topics which happen to have coincided in the preceding discourse). The final *żarūr* 'definitely' expresses certainty in that inference.

89.

Caē	pītē	hūē	sōśal	mīdīā	dekh
Tea	drink.IPFV.F.PL	be.PFV.M.PL.OBL	social	media	watch

rahī	hōṁ	gī	āpā	*żarūr*.
PROG.PFV.F.SG	be.SBJV.3.SG	FUT.F.SG	elder.sister	*definitely.*

"While drinking tea, elder sister will be looking at social media, *definitely*" (LUWC_f041p0067580).

Other instances of HCS MACs in final position have scope over an immediately preceding clause element instead of the whole clause. For instance, in (90), the HCS MAC has scope over adjective *fictional*, and emphasises that the addresser is talking about the death of a fictional character and not a real-life person. Although *fictional of course* is formally part of the question, it reads as an afterthought (since *fictional* would normally premodify but here follows its head noun) and not an integral part of the proposition that is being questioned.

90. "Do you have any other candidates for unexpected violent death, *fictional of course*?" (ECC_mail8684709).

Similarly, in example (91), the scope of the Urdu HCS MAC is on the immediately preceding elements *mēhngā hai* 'is expensive', the main adjective plus copula predicate, and therefore over the clause. The scope does not extend to the earlier clause in which the addresser suggests a means by which the addressee can read an expensive book.

91.

Mēhngā	tō	hai	lēkin	merē		
Expensive	EMPH	be.PRS.3.SG	but	1.SG.POSS.OBL		
jaisā	ġarīb	ādmī	is=kō			paṛh
like.M.SG	poor	person	DEM.SG.OBL=ACC			read
saktā	hai	lā'abrerī=sē	lē	kar,		*mehngā*
can.IPFV.3.SG	be.PRS.3.SG	library=from	take	do,		*expensive*
hai	*bilāśubha.*					
be.PRS.3.SG	*no.doubt.*					

"It is expensive but a poor person like me can read it by getting it from the library, it *is expensive no doubt*" (LUWC_f071p0031116).

6.4.2 Probability marker

In both languages, PS MACs in final position express a lower degree of certainty in the clause proposition; that is, as with the HCS MACs, clause final PS MACs typically have the whole clause in scope. Both this probability function and the scope of clause final PS MACs are illustrated in examples (92) and (93).

92. "The kids will eat only the squishy vitamins, *perhaps*"

(ECC_gdn_lifestyle_2020_oct_30_i_have_never_met_a_pharmist).

93.

Yeh	pālisī=kē	khilāf	nahīṁ	hai	*ġālibān.*
DEM	policy=GEN.M.PL	against	NEG	be.PRS.3.SG	*probably.*

"This isn't against the policy *probably*" (LUWC_f23p0084857).

Again as with the HCS MACs, there are instances of PS MACs in final position with scope over the immediately preceding element instead of the whole clause, as in (94) and (94).

94. "Joan I think he he's some kind of pirate or fortune-teller *maybe*".

(ECC_gdn_tv_and_radio_202_oct_24)

95.

rezalṭ=kā	din	thā	…	tīsarī	jamā'at=kā
Result=GEN.M.SG	day	be.PST.M.SG	…	third	grade=GEN.M.SG

ġālibān.
probably.

"It was a result day … grade three *probably*" (LUWC_f034p0089884).

6.4.3 Tagging

According to Simon-Vandenbergen and Aijmer (2007, p.138), sometimes an English HCS MAC is placed in clause final position to prompt confirmation from an addressee. This function of clause final HCS (and, as we will see, PS) MACs is similar to that of tag questions and therefore this function is called *tagging* (see Table, 2.2). The corpus examples show that in both languages, final position HCS MACs do indeed sometimes function to invite confirmation from an addressee regarding the addresser's proposition, as in (96) and (97). In (96), while interviewing a football player the addresser mentions a certain football practice activity and adds an HCS MAC after a noun phrase (whose existence is the proposition elliptically asserted) as a tag, signalling their own understanding of this subject matter, as well as giving an opening to the addressee to elaborate further.

96. A: "Heading practice too, *no doubt*?"

B: "Absolutely" (ECC_mail8955437).

In (97) the addresser shares an expectation about a potential action by a doctor. To prompt a response from the addressee in confirmation of their assertion that there will be a celebratory meal, they add a final position HCS MAC. The use of the HCS MAC indicates, moreover, that the addresser's proposition is based on a strong inference from general knowledge (here, that it is customary for a person who has had success in some venture to treat others to a meal in celebration).

97. A: Nēyā kalīnik khōlnē=kī khuśī=mēṁ
New.M.SG clinic open.INF.OBL=GEN.F.SG happiness=in

ḍāktar=kī jānib=sē d'avat hō gī
doctor=GEN.F.SG behalf=from meal be.SBJV.3.SG FUT.F.SG

yaqīnān.
definitely.

B: Jī rasmalai=kī palaitēṁ hōṁ gī.
Yes rasmalai=GEN.F.SG plate.PL be.SBJV.3.SG FUT.F.SG.

A: "To celebrate the opening of the new clinic, there will be a meal on behalf of the doctor, *definitely*"

B: "Yes there will be plates of rasmalai" (LUWC_ f041p0067580).

As with the HCS MACs, in both languages there are instances of clause final PS MACs being used to prompt the addressee to respond by confirming some inference the addresser has made, as in (98).

98. A: "I expect you're right – he probably is American. Just didn't know England didn't use 'high school'.

B: What do you use instead? 'Secondary school', *perhaps*?"

(ECC_AskUK201402).

In comparable Urdu examples too, the addresser may add a PS MAC in clause final position to prompt the addressee to confirm or deny their inference, as in (99).

99.	A:	Yeh	kām	hāl		hī	mēṁ	śōrū	kīyā
		DEM	work	recently		EXC	in	start	do.PFV.M.SG
		hai		*śāyad*?					
		be.PRS.3.SG		*perhaps*?					
	B:	Nahim		kaī	sāl		hō	gaē.	
		NEG		many	year.PL		be	go.PFV.M.PL	

A: "This work has started recently *perhaps*?"

B: "No, it has been many years" (LUWC_f029p000091149).

6.5 Semantic scope of MACs over negation

Sequences of MACs and negator, MAC+NEG, are counterparts of positive HCS and PS MACs, forming MACs categorised as *high certainty support for negative content* (HCSNC) and *probability support for negative content* (PSNC) (see 2.4.2.4). HCSNC and PSNC MACs have may have in scope an immediate element, an entire clause, or even content beyond the single clause (see 2.4.2.4). The core function of an HCSNC MAC is as a certainty marker applied to the negative content in a proposition; that of a PSNC MAC is as a probability marker applied to the negative content of a proposition. The concordances show that in both languages, addressers also use HCSNC and PSNC MACs as negative responses to rhetorical questions. PSNC MACs are moreover used as short responses to a previous addresser, a function already discussed in section 6.2.4.

In my data, HCSNC and PSNC MACs occur in main clauses; after dependent clause markers (e.g. *because*); and after coordinate clause markers (e.g. *but*).

6.5.1 Certainty marker

As mentioned, English HCSNC MACs may express certainty about the negative content of a proposition, as in (100).

100. "I've been out since 6am and there were no long lines. It's *definitely not* the same like years prior" (ECC_ind_b1762813).

Urdu HCSNC MACs likewise express the addresser's certainty in the negative content of a proposition, as in (101).

101. Gūziśtā rāt=kā humlā karnē vālē
Previous night=GEN.M.SG attack do.INF.OBL VALA.PL.OBL

Caman=sē tō *yaqīnān* *nahīṁ* āiē hōṁ
Chaman=from EMPH *certainly* *NEG* come.PFV.M.PL be.SBJV.3.PL

gē unhōṁ=nē *ġālibān* Ṭōrkham=kā bārḍar
FUT.M.PL DEM.PL.OBL=ERG *probably* Torkham=GEN.M.SG border

hī istemāl kīyā hō gā.
EXC use do.PFV.M.SG be.SBJV.3.SG FUT.M.SG.

"Last night's attackers *certainly* will *not* have come by Chaman, they *probably* will have used the Torkham border" (LUWC_e0455194).

Both in English and Urdu, HCSNC MACs occur in coordinate clauses for emphasis on the negative content. For instance, an HCSNC MAC after *but* (or Urdu equivalent) intensifies the negation in the *but*-clause in contrast to the assertion in the prior clause, as in (102) and (103).

102. "I'm lucky and happy to be where I'm now *but* it's *definitely not* where I thought I'd be 11 years ago" (ECC_ind_b1766549)

103. Śāyad āp=kē savāl=kā kisī had tak javāb
Perhaps you=GEN.M.PL question=GEN.M.SG some extent till answer

hō *magar* mukamal aur tafṣīlān *yaqīnān*
be.SBJV.3.SG *but* complete and detailed *definitely*

nahim hō gā
NEG be.SBJV.3.SG FUT.M.SG.

"Perhaps your question has an answer to some extent, *but* it *definitely* will *not* be complete and detailed" (LUWC_f049p0106090).

English HCSNC MACs are also used in dependent clauses, e.g. after *because,* to emphasise the addresser's certainty in a negative assertion given in the dependent clause, as in (104).

104. "Bikes aren't licensed *because* it would almost *certainly not* be cost-effective" (ECC_AskUK202005).

English HCSNC MACs sometimes occur initially in the *then*-clause of a conditional sentence (see 6.9). In these cases, addressers express their certainty in the truth of the negative assertion about the consequences of the condition given in the *if*-clause, as in (105).

105. "If I do something else, *then* I am *definitely NOT* listening to the meeting, and I will definitely miss something important" (ECC_AskUK202005).

I did not find examples of English or Urdu HCSNC MACs in *if*-clauses. Also, I did not find any example of Urdu HCSNCs MAC in *then*-clauses of conditional sentences.

6.5.2 Probability marker

The concordances show that English and Urdu PSNC MACs occur in main, dependent and coordinate clauses to express lower certainty in the truth of the negative content of the proposition, as (106) and (107).

106. "The title Mr Epstein has earned here is *perhaps not* fit for mixed company" (ECC_ind_b1773788).

107.

Āp=nē	śāyad	sirf	kitab	paṛhī	hai
You=ERG	perhaps	only	book	read.PFV.F.SG	be.PRS.3.SG

tārīkh	*śāyad*	*nahīṁ*	paṛhī.
history	*perhaps*	*not*	read.PFV.F.SG.

"You have perhaps only read a book, [you have] *perhaps not* read history" (LUWC_f199p0051316).

Like HCSNC MACs, PSNC MACS are found in coordinate and dependent clauses. When they occur after *but*, they lower the certainty in the truth of the negative content of the clause (which follows a concessive assertion), as in (108) and (109).

108. "It might pay your electricity bills, *but probably not* your rent" (ECC_AskUK201509).

109.

Billī	tō	dūdh	pī	jātī	hai
Cat	EMPH	milk	drink	go.IPFV.F.SG	be.PRS.3.SG

magar	bakrī	*śāyad*	*nahīṁ*	pītī.
but	goat	*perhaps*	*NEG*	drink.IPFV.F.SG.

"A cat drinks milk *but* a goat *perhaps* doesn't" (LUWC_f121p0075130).

PSNC MACs also occur in dependent clauses, as in examples (110) and (111). In both, the PSNC MAC follows the dependent clause marker, and expresses the addresser's

tentative stance towards the negative content of that clause as a possible reason for the main clause proposition to be true.

110. "I'd be very careful of ordering anything that goes on or in your body, *because* it's *probably not* been tested and may not be safe" (ECC_AskUK201909).

Example (111) additionally illustrates an interaction between a PS MAC and *saktā* functioning as an MV. They express the addresser's low certainty in a proposition potentially interpretable as disparaging of the poet under discussion (and thus are hedges possibly motivated by politeness, see 7.7 and 7.8, to avoid offence to any fan of that poet).

111.

Amām	Dīn	Gujrātī	bunyādī	tōr=par	Punjābī=kē
Imam	Din	Gujrati	basic	manner=on	Punjabi=GEN.M.PL

śāir	thē	yā	Urdū	Punjābī=kē	alfāz
poet	be.PST.3.PL	or	Urdu	Punjabi=GEN.M.PL	word.PL

milā	kar	aśār	kahtē	thē	*lēhāzā*
join.PFV.M.SG	do	verse.PL	say.IPFV.M.SG	be.PST.3.SG	*therefore*

ītnī	śustā	Urdū=mēṁ	śāir	un=kā
this.much.F	cultured	Urdu=in	verse	DEM.PL.OBL=GEN.M.SG

tō	*śāyad*	*hō*	*nahīṁ*	*saktā*.
EMPH	*perhaps*	*be*	*NEG*	*can.IPFV.M.SG*

"Imam Din Gujrati was basically a Punjabi poet, or spoke verses mixing words of Urdu and Punjabi; *therefore* such cultured verses in Urdu *perhaps cannot be* his" (LUWC_f059p0011414).

English PSNC MACs can occur in the *then*-clause of a conditional sentence to express the addresser's stance towards the probability of that clause's negative content, as in (112). However, Urdu PSNC MACs do not appear with equivalent function. I did not find any PSNC MAC in either English or Urdu within the *if*-clause of a conditional sentence.

112. "If they treat you differently than everyone else, *then* you're *probably not* their favourite person" (ECC_ind_a8303256).

6.5.3 Response to rhetorical questions

A rhetorical question is defined by Leech (2006, p. 103) as a question that "does not seek information, but rather implies that the answer is self-evident". In both English and Urdu, an HCSNC or PS MAC may be used as an addresser's reply to their own rhetorical question, having the effect of foregrounding authoritatively their point of view on the issue raised in the question (see Table 2.2). The use of an HCSNC MAC conveys certainty in a negative answer to the positive claim implicitly proposed by the rhetorical question, as in (113) and (114).

113. "But for those who come here today from Eritrea, Syria, Sudan and elsewhere, is the choice really only the UK or death? And should Priti Patel be demonised as a heartless and cruel racist for trying to control their numbers? *Of course not*, but the propaganda war steams on unchecked" (ECC_mail8696255).

114.

Pūrē	mulk=kō	dō	hissōṁ=mēṁ		taqsīm	kar	
Whole.OBL	country=DAT	two	part.PL.OBL=in		divide	do	
dīyā	jāīē.	Ēk	hissa	mardāna	hō		
give.PFV.M.SG	go.SBJV.3.SG.	One	part	masculine	be.SBJV.3.SG		
jis=mēṁ	galīyōṁ,	bāzārōṁ	aur	har	jaġa	sirf	
which=in	street.PL.OBL	bazar.PL.OBL	and	every	place	only	
mard	pāē	jātē	hōṁ	aur			
man.PL	find.SBJV.3.PL	go.IPFV.M.PL	be.SBJV.3.PL	and			
auratōṁ=kā		dākhla	mamnūn	hō.			
woman.PL.OBL=GEN.M.SG		entry	forbidden	be.SBJV.3.SG.			
Aur	dūsrā	hissa	zanāna	jahāṁ	mardōṁ=kē		
And	second	part	feminine	where	man.PL.OBL=GEN.M.PL		

līē	nō	entarī	hō	kyā	aisā	amlī	tōr=par
for	no	entry	be	Q	such.M.SG	practical	manner=on

mūmkīn	hai?	*Yaqīnān*	*nahīṁ*	tō	phir
possible	be.PRS.3.SG?	*Certainly*	*not*	EMPH	then

kyā	yeh	zarūrī	nahīṁ	hō	jātā	ke	laṛkē
Q	DEM	necessary	NEG	be	go.IPFV.M.SG	that	boy.PL.OBL

laṛkīyōṁ=kō	ibtedā	hī	sē	ēk	dūsrē=kē	
girl.PL.OBL=ACC	beginning	EXC	from	one	other.PL.OBL=GEN.M.PL	

acchē	tōr	tarīkē,	rakh-rakhāō	aur
good.M.PL.OBL	manner	way.PL.OBL,	manner-manner	and

tahzīb-o-tamadan=kē	sāth	rehnē=kī
civility-and-courtesy=GEN.M.PL	together	live.INF.OBL=GEN.F.SG

ādat	ḍālnē	kē	līē	makhlūṭ	ṭarīqa	
habit	put.INF.OBL	GEN.M.PL	for	mixed	method	

talīm=kō	farōġ	dīyā	jāē.
education=ACC	propagate	give.PFV.M.SG	go.SBJV.3.SG.

"The whole country should be divided into two parts. One part should be male, where in the streets and markets and all places only men can be found, and women should be forbidden entry. And the second part female, where it's no entry for men. Is this practically possible? *Definitely not,* then doesn't it become necessary that boys and girls, from the very beginning, be given a mixed education to get them into the habit of living together in a harmonious, courteous, and civil manner with each other." (LUWC_j0021727).

Conversely, use of a PSNC MAC after the rhetorical question expresses the addresser not being convinced that the implied proposition in their rhetorical question can be refuted outrightly, as in (115) and (116).

115. "But can we track down the super-spreaders before they cause too much damage? *Perhaps not*" (ECC_mail8850511).

116. Hum āj jis dōr=sē ġūzar rahē haiṁ
1.PL today which era=from pass PROG.PFV.M.PL be.PRS.1.PL

yeh hawāēṁ jō tabdīlī=kā pēġām lā
DEM wind.PL REL change=GEN.M.SG message bring

rahī haiṁ kyā yeh voh tabdīlī hai
PROG.PFV.F.SG be.PRS.3.PL Q DEM DEM change be.PRS.3.SG

jis=sē humārē hālāt badlēṁ
which=from 1.PL.POSS.OBL condition.PL change.SBJV.2.PL

gē? *Śāyad* *nahīṁ*!!
FUT.M.PL? *Perhaps* *NEG*!!

"The era which today we are passing through, the winds that are bringing the message of change, is this the change by which our conditions will change? *Perhaps not*!!" (LUWC_f032p0006074).

6.6 Interaction of MACs and MVs

As the parallel data shows, in English there is pervasive interaction between MACs and MVs (see 4.4.1). On the other hand, Urdu has only two MVs, but there are other means, e.g. MACs and subjunctive verbs, to convey modality (see Table 4.12). In this section, I will discuss some of the common MAC + MV patterns in the English and Urdu data. As discussed in section 4.4.1, the parallel corpus illustrates how, due to the lack of Urdu MVs that correspond directly to most English MVs, the translators have used Urdu MACs in lieu of *may, might, could* and *must* to convey epistemic modality. English MVs are polysemous and therefore their meaning as epistemic, deontic or dynamic in a sentence is context dependent. For instance, *can* may convey all three modal meanings in the appropriate context, as examples (117) to (119) show.

117. He *can* come in now. [deontic]

118. He *can* write well. [dynamic]

119. It *can* be a slow process. [epistemic]

In this section, I present MACs cooccurring with MVs in contexts where the MV has epistemic meaning, in contexts where the MV has deontic meaning, and in negative constructions.

6.6.1 Interaction of MACs with MVs in epistemic context

Hoye (1997, p. 158) says that in English, when a MAC cooccurs with an MV, it focalises that MV, whereas Quirk et al. (1985, p. 584) say that a MAC cooccurring with an MV does not necessarily bring that MV into focus. A pattern MAC + *will* is ambiguous regarding whether the focus is on just the MV or the clause as a whole (Quirk et al., 1985, p.584).

The data shows that in both languages, HCS MACs intensify the addresser's argument when they cooccur with an MV in epistemic context, as in (120) and (121). In such instances, the MAC + MV combination expresses a higher level of epistemic probability than the MV would alone. In (120), the English HCS MAC has scope over an MV conveying epistemic modality. The MAC is high-certainty, but the MV is low-certainty. Therefore, on one level the sentence assesses the act of inference under discussion as possible, on the other level it assesses that act of assessment as high confidence.

120. "We also acknowledge that a person *certainly might* reasonably – and justifiably – infer that different actions by the officers could have saved Mr Easter's life" (ECC_mail8795959).

When an Urdu HCS MAC cooccurs with an MV conveying epistemic modality, as in (121), it expresses a higher level of epistemic probability than the MV would alone.

121. Tāhum yeh aisī bāt hai jissē żabt
Although DEM such.F.SG matter be.PRS.3.SG REL.OBL capture

taḥrīr=mēṁ tō *yaqīnān* lāyā jā *saktā*
writing=in EMPH *certainly* bring.PFV.M.SG go *can.IPFV.M.SG*

hai magar is=par ‘amal darāmad
be.PRS.3.SG but DEM.SG.OBL=on practical implementation

nahīṁ karāyā jā saktā.
NEG do.CAUS.M.SG go can.IPFV.M.SG.

“Although it is such a matter that it *certainly can* be captured in writing, but it may not be implemented upon” (LUWC_e1897462).

As we have seen earlier in this chapter, both English and Urdu addressers use PS MACs to lessen the strength of a proposition. But this effect is more pronounced when such MACs cooccur with MVs that also convey low certainty. The concordance shows that when MVs such as *may*, *might*, *can*, and *could* cooccur with PS MACs, the epistemic meaning weakens and tentativeness is enhanced, as in examples (122) and (123).

122. “I’m sure it can’t hurt to ask and he *may perhaps* offer an explanation as to the weird times they ring” (ECC_AskUK202005).

123. Yeh śaēr pehlē bhī kahīṁ paṛh rakhā
DEM verse before.OBL INC somewhere read keep.PFV.M.SG

hai. Lēkin hērat hai ke gūgal
be. PRS.3.SG. But amaze be.PRS.3.SG that google

karnē=par milā hī nahīṁ. *Ġālibān*
do.INF.OBL=on find.PFV.M.SG EXC NEG. *Probably*

kisī ġērmārūf śāīr=kā *hō* *gā*.
some unknown poet=GEN.M.SG *be.SBJV.3.SG* *FUT.M.SG*.

"I have read this verse somewhere before. But it's amazing that it isn't found by googling. *Probably* it *will be* by some unknown poet" (LUWC_f059p0091070).

Example (123), has the future construction (*gā*), used to express strong inference like English *will* because there is no equivalent Urdu MV. Therefore, I am counting such constructions as cooccurrence with MV for practical reasons because of future function of *gā*, even though it is not really an MV.

6.6.2 Interaction of MACs with MVs in deontic context

In English, MACs may cooccur with MVs (e.g. *should*) that are primarily used to express obligation and necessity (i.e. deontic modality), to shift the meaning from deontic to epistemic modality and thus convey a tentative assertion (see Hoye, 1997, p.95; Simon-Vandenbergen & Aijmer, 2007, p. 20). Hoye (1997, p. 111) says that in English, when a MAC cooccurs with an MV with deontic sense, there remains a sense of indeterminacy regarding whether the reading is obligatory, epistemic, or imperative. However, Simon-Vandenbergen and Aijmer (2007, p.286) say that when certain HCS MACs (*certainly* and *definitely*) occur in such constructions, they do not express "epistemic necessity", rather they strengthen the force of "deontic elements" and thus can modify the force of a "speech act".

My data shows that when an English HCS MAC cooccurs with an MV expressing a deontic sense, it expresses high confidence in the obligation. For instance, in (124) the HCS MAC has scope over the MV, and the MAC expresses high confidence in the assessment of the obligation.

124. "I think there's enough evidence now on the benefits of mask-wearing - *certainly* secondary school students *should* be told to wear masks and primary school children should be encouraged to wear masks ."
(ECC_gdn_world_2020_dec_14_what_know).

In Urdu too, in some instances, the addition of an HCS MAC strengthens the deontic meaning conveyed by the MV, as in (125).

125.

Jōharī	kānfarans=kē	'almī	modabarīn=kō		
Nuclear	conference=GEN.M.PL	international	thinker.PL=ACC		
Ūkrāin=kē		mustaqbil	aur	ḥalīa	boḥrān=kō
Ukraine=GEN.M.SG.OBL		future	and	recent	crisis=ACC
żarūr	mad-ē-nazar	rakhnā	*cāhīē*.		
definitely	consideration	keep.INF	*should*.		

"The international thinkers of the nuclear conference *should definitely* consider Ukraine's future and the present crisis" (LUWC_e0239170).

In some cases in Urdu, addition of an HCS MAC to an MV expressing deontic sense shifts the deontic meaning towards epistemic meaning. For instance, on its own in (126), the reading of *cāhīē* is of obligation; given that the situation now in his favour, Imran Khan has the duty to be content with his lot. But by adding *yaqīnān,* the addresser shifts the meaning to express their assessment of the likelihood of Imran Khan being happy in the present circumstances; epistemic meaning is dominant.

126. Ē pī sī=kē bād Imrān Khān sahib=kō
A P C=GEN.M.PL.OBL after Imran Khan mister=ACC

yaqīnān khuś hōnā *cāhīē* ke ab "voh darvāza"
definitely happy be.INF.M.SG *should* that now "DEM door"

bhī bund hūā. Navāz Śarīf=nē bilkul
INC close be.PFV.M.SG. Nawaz Shareef=ERG absolutely

vāhzē alfāz=mēṁ ēlān kar dīyā hai ke
explicit.OBL word=in announce do give.PFV.M.SG be.PRS.3.SG that

un=kī "larāī" Imrān Khān sahib=sē nahīṁ.
DEM.SG.OBL= GEN.F.SG "fight" Imran Khan mister=with NEG.

"Mr. Imran Khan *should definitely be* happy after APC that now "that door" has also been closed for good. Nawaz Shareef has announced in absolutely explicit terms that his "fight" is not with Mr. Imran Khan" (LUWC_n1222491).

When an English PS MAC cooccurs with an MV that typically expresses deontic modality, the PS MAC lessens the degree of the imposition of an obligation. The use of a PS MAC with *should* or *must* pragmatically reduces the force of the deontic meaning to tentative suggestion, as in examples (127) and (128). It may be noted that these examples can be interpreted as less-than-full confidence in the truth of the obligation expressed.

127. "Booked for a late challenge in the second half, and *perhaps should* have been sent off after a series of fouls, but sealed the deal with the assist for Cavani's effort" (ECC_ind_b1671747).

128. Mutāsirah hissē=par *śāyad* lōhā ragaṛnā *cāhīē*.
Affected part=on *perhaps* iron rub.INF *should*.

"*Perhaps* [they] *should* rub an iron on the affected part" (LUWC_ f041p0067580).

6.6.3 Interaction of MACs with negated MVs

An HCS MAC cooccurring with a negated MV (e.g. *may not*, *might not*) strengthens the confidence with which the truth of the negative content of the proposition is asserted, as in (129) and (130). In (129), an (epistemic) MV has scope over the negative (i.e. *might not*). The negation thus comes within the scope of the HCS MAC before the MV, and the addresser expresses their level of certainty – higher than 'X might not Y' but lower than 'X certainly doesn't Y' – in the proposition being untrue.

129. "It's just what they value – their values *certainly might not* coincide with yours" (ECC_AskUK201810).

My corpus shows that Urdu MACs may also have scope over negated MVs. In (130), an HCS MAC has scope over the negated MV, and thereby over the whole clause, conveying higher confidence in the negative content of the proposition than negated *sak* 'can' would express unaided.

130.

Bijlī=kā	buhrān	itnā	śadīd
Electricity=GEN.M.SG	crisis	this.much.M.SG	severe

hai	issē	*yaqīnān*	rātōṁ	rāt	khatam
be.PRS.3.SG	DEM.OBL	*certainly*	night.PL.OBL	night	end

nahīṁ	kīyā	jā	*saktā.*
NEG	do.PFV.M.SG	go	*can. IPFV.F.SG.*

"The electricity crisis is so severe that it *certainly cannot* be ended overnight" (LUWC_e0306694).

Examples (131) and (132) illustrate that PS MACs alongside negated MVs similarly express the addresser's lower degree of confidence in the truth of the negative content of the proposition, just as PS MACs do when they cooccur with a negator but no MV. In both (131)

and (132) addressers express an even lower confidence level in the content of the proposition than the MV alone would convey.

131. "Personally, I think my favourite (which *perhaps might not* be the right word here) namesake isn't of two famous people, but of one infamous person, and then a relative or acquaintance" (ECC_AskUK202001).

132.

Ēk	rekard	qaim	kīyā	hai	jō
One	record	set	do.PFV.M.SG	be.PRS.3.SG	REL

śāyad	kabhī	ṭūṭ	*nahīṁ*	*saktā*.
perhaps	ever	break	NEG	*can*.IPFV.M.SG.

"A record has been set that *perhaps cannot* ever be broken" (LUWC_F034P0098707).

6.7 Interaction of MACs with MACs

In both languages, some but not all MACs may cooccur with another MAC. There are instances where a PS MAC cooccurs with an HCS MAC, either immediately next to each other or within a distance of three word-tokens. Other MACs are not observed to cooccur with another MAC, for example, *no doubt* and *bilāśūbha* 'no doubt'.

In my data, the observed sequences of HCS + PS MAC in English are *certainly probably*, *obviously perhaps*, and *definitely maybe*. There is only one instance of a PS + HCS MAC sequence in the English corpus, *possibly of course*. The Urdu HCS + PS MAC sequences observed are *yaqīnān hō saktā* 'definitely maybe', *beśak hō saktā* 'no doubt maybe', and *żarūr hō saktā* 'certainly maybe'. Interestingly, these sequences show that the Urdu HCS MACs tend to occur with the fixed MAC phrase *hō saktā* 'maybe', formally a clause, so syntactically, it is not strictly a MAC + MAC sequence. Moreover, the meaning of *saktā*, like that of English MVs, is context dependent (see 2.3.1).

When an HCS MAC precedes a PS MAC, it raises the degree of certainty expressed by the PS MAC. In example (133), the combination of an HCS and a PS MAC gives a multilevel confidence reading. The PS MAC lowers the addresser's confidence in the proposition, but the HCS MAC increases the addresser's certainty; so overall this sequence expresses higher addresser confidence in the proposition than a PS MAC alone, but lower confidence than would be expressed by an HCS MAC alone.

133. "He *certainly probably* could use one, as investigations in New York continue to heat up" (ECC_ind_b1778807).

There are no instances of a PS MAC preceding a HCS MAC in the English data. But Urdu PS+HCS MAC sequences are observed: *ġālibān yaqīnān* 'probably definitely' and *śāyad żarūr* perhaps certainly'. A PS MAC before an HCS MAC lowers the expressed confidence in the proposition, as in (134).

134.	*Śāyad*	*żarūr*	kisī	miśon=pē	nikl
	Perhaps	*certainly*	some	mission=on	leave
	gaē	hōṁ	gē.		
	go.PFV.M.PL	be.SBJV.3.PL	FUT.M.PL.		

"He will have *perhaps certainly* left on some mission" (LUWC_f041p0067580).

There are some instances of the Urdu PS + HCS MAC sequences *śāyad yaqīnān* 'perhaps definitely' and *yaqīnān śāyad* 'definitely perhaps', but they are separated by *nahīṁ* 'not' (see 6.5 and 7.2) or occur in a separate clause marked by contrastive marker *lēkin* 'but' or *balkeh* 'rather' (e.g. *śāyad balkeh yaqīnān* 'perhaps, or rather definitely').

In the Urdu corpus, the PS + PS MAC sequences are *śāyad hō saktā* 'perhaps maybe', *ġālibān śāyad* 'probably perhaps', *śāyad ġālibān* 'perhaps probably'. Of the three sequences, the first two also occur in reverse order. In these cases, the PS MAC cooccurring with another

PS MAC further lowers the degree of certainty that is conveyed beyond the effect of a single PS MAC, as in (135).

135.	*Ġālibān*	*śāyad*	hī	aisā	koī	karē
	Probably	*perhaps*	EXC	such	any	do.SBJV.3.SG

"*Probably, perhaps* hardly anyone would do so" (LUWC_f034p0080799).

In ECC, the PS + PS MAC sequences observed are *probably possibly* and *possibly probably*. These seem to express tentativeness and weak probability, as in (136), which is a part of a conversation quoted in a newspaper. But in reality, this and similar examples are not plausible sequences of PS + PS MACs in English. Instead, they either result from a speaker saying one word and then replacing it when they think of a better substitute, as in (136), or else occur in two separate clauses so as to be adjacent but not actually linked together.

136. "There's *probably, possibly* drugs involved, that's what I hear" (ECC_ind_b421830).

ECC has only one instance each of the HCS +HCS MAC sequences *obviously of course* and *of course obviously*. These intensify the degree of certainty expressed. In LUWC, no sequences of HCS +HCS MAC were observed. However, Urdu HCS MACs do cooccur with other certainty markers, such as in *yaqīnān haqīqat mēṁ* 'definitely in reality', where the postposition phrase *haqīqat mēṁ* could in fact potentially be counted as a MAC, although it was not on my initial list of MACs selected for analysis.

6.8 MACs and interrogative sentences

The existing literature (Boye, 2012, p. 36; Simon-Vandenbergen & Aijmer, 2007, p. 89) establishes that, in English, HCS MACs normally do not occur in interrogative clauses

because MACs cannot simultaneously assess the truth of the proposition and question the truth of that proposition. On the other hand, Boye (2012, p. 37) and Suzuki (2015, p. 1370) point out that the English PS MAC *perhaps* can occur in interrogative constructions (see 2.3.2.2).

ECC contains a few instances of *perhaps* in interrogative clauses, as in (137).

137. "*Do* you *perhaps* think that Ibiza in particular was a really expensive and exclusive destination?" (ECC_AskUK201810).

But such examples are rare in my corpora (see 8.4).

For Urdu, I found no MACs in interrogative clauses. While there were examples of *hō saktā* within interrogatives, in all cases it was not an adverbial but rather part of the clause verb group. For instance, in (138), *hō saktā* is a part of a verb group *śāmil hō saktā hai* in which *hai* (the inflected auxiliary) governs *saktā* which governs *hō* which governs *śāmil*. In such cases, *saktā* functions as an MV instead of part of a MAC phrase.

138.	Kyā	voh	dōbāra	śāmil	hō	saktā	hai	yahāṁ?
	Q	DEM	again	join	be	can.IPFV.M.SG	be.PRS.3.SG	here?

"Can he join again?" (f022p0007250).

Therefore, Urdu PS MACs do not occur in interrogative clauses.

6.9 MACs and conditional sentences

A conditional sentence is an aggregate of two clauses: the antecedent (i.e. *if*-clause or protasis) and the consequent (i.e. *then*-clause or apodosis) (Sharma, 2010, p.107). Adding MACs to conditional sentences, whether in the *if*-clause or the *then*-clause, alters the meaning. Simon-Vandenbergen and Aijmer note, for instance, that when the English HCS

MAC *indeed* is added to the *if*-clause, a high degree of "confirmation is made subject to conditionality" (Simon-Vandenbergen & Aijmer, 2007, p.111) (though they do not observe any other MACs in the *if*-clause of conditionals). Instances of MACs within *epistemic conditionals*, defined by Dancygier and Sweetser (2005, p. 17) as those conditionals that express an addresser's epistemic stance, are good illustrations of MACs expressing the addresser's confidence in their knowledge about some precondition (see 2.4.2). A typical epistemic conditional sentence asserts that *if* some identified state of affairs (in the *if*-clause) turns out to be true *then* we can believe with confidence what is being predicted (in the *then*-clause) (Dancygier & Sweetser, 2005, p. 81). In (139), an example fabricated to illustrate epistemic conditionals, the presence of the car is a precondition to the logical deduction of John being at home. We will see some of the examples of MACs within conditionals of this sort in this section.

139. If the car is in the porch, John is home from his office.

When an HCS MAC occurs in the *then*-clause, it expresses the addresser's strong belief in the truth of the predicted state of affairs as illustrated in (140) and (141).

140. "If they got a diagnosis of anxiety *then, of course*, it is recognised"

(ECC_AskUK201906).

141.

Tārē	zamīn=par		agar	vālīdēṁ	dekhēṁ	
star.PL	earth.M.SG=on		if	parent.PL	watch.SBJV.2.PL	
tō	*yaqīnān*	voh	apnē		baccōṁ=kō	
then	*certainly*	DEM	REFL.POSS.PL.OBL		child.PL.OBL=ACC	
ẕyāda	behtar	ṭarīqē=sē		t'ālīm	dē	saktē
more	better	way=with		education	give	can.IPFV.M.PL

haiṁ.
PRS.3.PL.

"If parents watched 'Stars on Earth' *then certainly* they would be able to educate their own children in a much better way" (LUWC_f045p0011277).

On the other hand, a PS MAC in the *then*-clause expresses the addresser's low certainty in the truth of the prediction, as illustrated in (142) and (143).

142. "I tend not to look at the regrets because *if* I had stayed at Arsenal *then* I *probably* wouldn't be where I am now." (ECC_gdn_football_2020_dec_03).

143.

Agar	voh	har	gaē	*tō*	*śāyad*	voh	doṛ=sē
If	DEM	lose	PFV.M.PL	*then*	*perhaps*	DEM	race=from

nikl	jāīṁ	kīūṁke	apnī	hī	rēyāsat=mēṁ	hārnā
leave	go.SBJV.3.PL	because	own	EXC	state=in	lose.INF.M.SG

un=kī	umīdvārī=par	kārī	zarb	hō	gī.
DEM.PL.OBL=GEN.F.SG	candidature=on	hard	strike	be.SBJV.3.SG	FUT.F.SG.

"*If* he loses *then perhaps* he will leave the race because losing in his own state will be a major blow to his candidature" (LUWC_f032p0010590)

In my search for epistemic conditionals, I found that in both languages, MACs are typically used in the *then*-clause rather than the *if*-clause. The reason might possibly be that addressers tend to use MACs to express confidence in inferences about possible outcomes, and so add MACs in the *then*-clause where they have scope over the prediction that is expressed.

6.10 Chapter summary

The foregoing corpus investigation demonstrates, first, that English and Urdu MACs are similar in the range of different scopes over other elements they may have in the various

clausal positions. This similarity in scope is also reflected, with minor differences, in the similarity of the semantic functions they perform in those different positions.

The core function of all MACs is to express the addresser's higher or lower confidence in the truth of a proposition. The corpus investigation shows that MACs have other, supplementary functions depending on their clausal positions. In clause initial position, MACs have scope over the whole clause and have two core functions (as certainty and probability markers) and one supplementary function (short response). However, in English, short responses can be formed with both positive and negative MACs, whereas in Urdu they are only formed with positive MACs. In clause medial position, MACs in both languages similarly focalise the immediately subsequent element but have scope over the meaning of the whole clause. In clause final position, MACs in both languages typically have scope over the whole clause although in some instances they have scope over only the directly preceding element. In addition to the core function of expressing certainty or probability of the truth of the proposition, MACs in clause final position sometimes function as *tagging*, to prompt a response from an addressee.

Urdu, with few modal verbs, regularly uses MACs where English would use an MV. In English, on the other hand, use of MACs *alongside* MVs to emphasise or downtone the meaning of the MVs is pervasive. In both languages, MACs interact with other MACs in specific sequences such as HCS + PS MAC. Similarly, in both languages, HCS MACs do not occur in interrogative clauses, though PS MACs do in English but not Urdu. In both languages, MACs in epistemic conditional sentences influence the expressed degree of certainty in the inference expressed in the *then*-clause in a similar manner.

7 Pragmatic functions of English and Urdu MACs

7.1 Chapter overview

In this chapter, I answer the third part of RQ 3, namely, in what pragmatic functions are English and Urdu MACs employed? For this part of the analysis, I describe the rhetorical-pragmatic functions (see 2.4.3) performed by English and Urdu MACs. To explore the pragmatic functions of English MACs, I apply to the data the existing literature on this topic. As there is no literature to date on the pragmatic functions of Urdu MACs, I work from similarities exhibited by Urdu MACs to the pragmatic functions already established for English MACs, as per my review in section 2.4.2. I then enhance this beginning by incorporating differences I observe in the data (see 4.3.6). Across sections 7.2 to 7.8 the following rhetorical-pragmatic functions are discussed: authority (see 2.4.2.1), emphasis (see 2.4.2.2), solidarity (see 2.4.2.3), expectation (see 2.4.2.4), counter-expectation (see 2.4.2.7), hedging (see 2.4.2.6), and politeness (see 2.4.2.8). In section 7.9, I give a summary of this chapter.

7.2 Authority

High certainty support (HCS) MACs frequently function to convey an addresser's assertion of authority over some topic under discussion (see 2.4.2.1). The data shows that English and Urdu HCS MACs are indeed used in various constructions to convey an addresser's claim to superior knowledge.

The literature on English reports that one common means for an addresser to convey their authority over the subject matter is to use *I think* before presenting a point of view, to

overtly express that authority. The use of *I think* encodes the addresser's conceptual information. Addition of an HCS MAC to a clause which follows *I think* then strengthens the addresser's claim to authority. Example (144) illustrates use of an HCS MAC in a clause following *I think* to strengthen the addresser's authoritative assertion of a proposition. Use of *definitely* after *I think* expresses the conviction of the addresser that the proposition is true.

144. "They only really work in certain locations, but *I think* they are *definitely* a good way of more effectively managing traffic than just leaving people to figure stuff out for themselves and they are significantly cheaper than adding extra lanes to the carriageway" (ECC_mail9027299).

Prior research shows that, in Urdu too, phrases such as *merā khayal hai ke* 'my opinion is that' are used overtly by addressers to express a claim to authority regarding the subsequent proposition (see 2.4.3.2; see also Genady, 2005, p. 110). As in English, adding an HCS MAC to such a construction strengthens this claim. Example (145) illustrates how in some cases, the addresser further reinforces their claim to authoritative knowledge by presenting reasons in support of the assertion that contains the HCS MAC.

145.

Merā	*khēyāl*	*hai*	ke	ilēkṣun	rēfārumz
1.SG.POSS.M.SG.NOM	*opinion*	*be.PRS.3.SG*	that	election	reform.PL

żarūr	hōṁ	gī.	Yeh	in	sub
definitely	be.SBJV.3.PL	FUT.F.SG.	DEM	DEM.PL.OBL	all

jamā'atōṁ=kī	majbūrī	hai
party.PL.OBL=GEN.F.SG	necessity	be.PRS.3.SG

"*My opinion is* that there will *definitely* be election reforms. This is a necessity for all the [political] parties" (LUWC_ f032p0083373).

In some cases in Urdu, an addresser strengthens their assertion of authoritative knowledge by placing an HCS MAC near a word such as *tāsūr* 'impression', *mālūm* 'known',

or *mālumāt* ‘knowledge, information’. Using an HCS MAC in this way is a strategy used by the addresser to imply that their argument is based on objective knowledge of the circumstances. Thus, the source of knowledge is marked as high certainty to underline that it should be believed, as shown in (146).

146.

Sūbāī	vazīr=kī		prēs	kānfarēns=kē		bād
Provincial	minister=GEN.F.SG		press	conference=GEN.M.PL.OBL		after
bādī-ūl-nazar=mēṁ	yeh	tāsūr	*żarūr*	paēdā		hūā
appearance=in	DEM	impression	*definitely*	born		be.PFV.M.SG
hai	jaēsē	chōtē	sūbōṁ=kō	tārgut	kīyā	
be.PRS.3.SG	as.if	small.M.PL.OBL	province.PL.OBL=ACC	target	do.PFV.M.SG	
jā	rahā	hai	aur	un=kē	khilāf	
go	PROG.PFV.M.SG	be.PRS.3.SG	and	DEM.PL.OBL=GEN.M.PL	against	
koī	inteqāmī	kāravāī	hō	rahī	hai.	
some	retaliatory	action	be	PROG.PFV.F.SG	be.PRS.3.SG.	

“After the provincial minister’s press conference, apparently the impression has *definitely* been made that the small provinces are being targeted, and retaliatory action is being taken against them” (LUWC_e0245401).

The existing literature (see 2.4.3.1) shows that another way an addresser can express authority in English is to use an HCS MAC alongside the first person plural pronoun *we* and a cognition verb (e.g. *understand*). This strategy may be used when an addresser does not want to come across as claiming to be a direct personal authority; instead, the proposition is overtly framed as shared knowledge. Addressers may well be motivated to use *we* in public discourse in particular in order to avoid coming across as condescending towards their addressee(s) and thus causing offence. Example (147) illustrates how an addresser’s use of HCS MAC *of course* with *we* expresses their acknowledgement that the addressees are equally aware of the stated proposition.

147. “The fact that the new vaccine appears to give initial protection in 90 per cent of cases is better than even the greatest optimist could have predicted. *Of course, we* have to understand that the crisis isn’t over yet” (ECC_mail8932615).

Combinations of HCS MACs and *ham* ‘we’ (or *ham lōg* ‘we people’) also occur in Urdu to express collective or societal responsibility. In example (148), the addresser wishes to argue for a collective wish among their national group for a necessary catharsis. By using *ham baḥaṣīyat millat-ō-qōm* ‘we as a nation and community’ and repeating the reference as *hamēṁ* ‘us’ directly before the MAC, the addresser avoids distancing themself from the addressees. By adding *yaqīnān* the addresser adds emphasis to *ham* ‘we’ as well as to *kathārsis kī bhī żarūrat* ‘need for catharsis’.

148.	Ham	baḥaṣīyat		millat-ō-qōm		jis
	1.PL	in.capacity.of		nation-and-community		RELSG.OBL
	fēz=mēṁ	haiṁ	vahāṁ	infarādī	aur	ijtemā‘aī
	phase=at	be.PRS.3.PL	there	individual	and	collective
	saṭaḥ=par	*hamēṁ*	*yaqīnān*	kathārsis=kī		bhī
	level=on	*1.PL.ACC*	*certainly*	catharsis=GEN.F.SG		INC
	żarūrat	hai.				
	requirement	be.PRS.3.SG.				

“The phase we are at as a nation and community, *we certainly* need catharsis too at individual and collective level” (LUWC_f022p0001157).

In addition, in Urdu addressers also use PS + NEG + HCS sequences (e.g. *śāyad nahīṁ yaqīnān* ‘not perhaps but certainly’) to endorse a previous addresser’s lower-confidence assessment of the truth of a proposition. By negating the previous addresser’s probability assessment, the addresser shows confidence in their authority on the subject, as in example (149).

149. A: Issē qahva kahtē haiṁ *śāyad*
DEM.OBL tisane call.IPFV.M.PL be.PRS.1.PL *perhaps*.

B: *Śāyad* *nahīṁ* *yaqīnān* qahva hī
Perhaps *NEG* *certainly* tisane EXC

kahtē haiṁ.
call.IPFV.M.PL be.PRS.3.PL.

A: "It is called tisane *perhaps*".

B: "*Not perhaps, certainly* it is called tisane" (LUWC_f119p0054442).

Such sequences of PS + NEG + HCS are not observed in the English Comparative Corpus (ECC), where instead I found such instances as example (150). Though not exactly the same pattern as in Urdu, (150) is equivalent to the Urdu examples in that it involves one addresser authoritatively negating the probability assessment made by the other.

150. A: "*Perhaps* :) Does it truly matter at the end of the day I wonder? :)"

B: "*Not perhaps*, its fact" (ECC_AskUK202005).

7.3 Emphasis

Certain HCS and HCSNC MACs, such as *of course,* may function in a clause to emphasise an addresser's conviction that the hearer should accept the truth of their proposition (Simon-Vandenbergen & Aijmer,2007, p. 220). In examples (151) and (152), to express high certainty in one key facet of their assertion, the addressers use an HCS MAC immediately before what they wish to emphasise.

151. "I don't know what I would call them *but certainly not* a restaurant".
(ECC_AskUK201906).

152. Pānc pānc sau=kī ṭikaṭ lē kar lōg
Five five hundred=GEN.F.SG ticket get do people

kyā dēkhanē jātē haiṁ, *yaqīnān* itnā
what watch.OBL.INF go.IPFV.M.PL be.PRS.3.PL, *certainly*
this.much.M.SG

mehngā mizāḥ tō nahīṁ hōtā.
expensive humour EXC NEG be.IPFV.M.SG.

"What do people go to watch by getting a five hundred [rupee] ticket? *Certainly* humour is *not* so expensive" (LUWC_ f029p0007626).

In some instances, an addresser emphasises their viewpoint using a MAC to raise an expectation but then counters the raised expectation (see 7.6) in a following concessive clause, as in examples (153) and (154).

153. "We could *definitely* do with a bit more culture, *but* I think that's something that has to happen organically over time. It's not something that can be bought and implemented IMO" (ECC_AskUK202005).

154. *Bēṣak* vafāq Karācī sarkūlar raēlwē=par
No.doubt Federal Karachi Circular Railway=on

da'avah karē *par* yeh mehnat Sindh hakūmat=kī
claim do.IMP.M.SG *but* DEM hard.work Sindh government=GEN.F.SG

hai aur Murād Alī Ṣāh=kī hai.
be.PRS.3.SG and Murad Ali Shah=GEN.F.SG be.PRS.3.SG.

"*No doubt* the Federal Government may claim the Karachi Circular Railway, *but* the hard work is down to the Sindh Government and Murad Ali Shah" (LUWC_j0822931).

In English sometimes, degree adverbs (e.g. *most*) and MACs are used together to intensify the addresser's high certainty in the truth of their proposition.

155. “I do think you get better quality from local suppliers and *most certainly avoid* frozen meat in supermarket, one of my bugbears is the amount of water they add to chicken to get the weight up” (ECC_AskUK202005).

In Urdu, degree adverbs do not occur before MACs for emphasis in this way. Instead, sometimes an Urdu HCS MAC is repeated for emphasis. Earlier I showed, in the analysis presented in Table 4.12, that English dummy auxiliary *do*, when used for emphasis, is sometimes translated into Urdu as the HCS MAC *żarūr* ‘definitely’ (see Biber et al., 1999, p.908-9 on use of *do* as emphasiser). The LUWC data shows that reduplicated HCS MACs in Urdu function to reinforce what a person is arguing about, to confirm a previous addresser’s proposition, or to agree emphatically with someone’s invitation. Example (156) illustrates reduplication of *żarūr* ‘definitely’ as *żarūr żarūr*, to express a strong positive response to a request. It can be translated as ‘yes, of course/definitely/certainly’ or ‘most definitely’.

156. A: Sēyārah” ‘Arabī=mēṁ kār=kō kahtē
“Siyara” Arabic=in car=ACC call.IPFV.M.PL.OBL

haiṁ. Aur “sētārah” ‘Arabī=mēṁ khiṙkī,
be.PRS.1PL. And “sitara” Arabic=in window,

darvāẕē=kē pardē=kō kahtē haiṁ
door.OBL=GEN.M.PL curtain=ACC call.IPFV.M.PL be.PRS.3.PL.

B: *Żarūr* *żarūr* kīyōṁ nahīṁ.
Certainly *certainly* why NEG.

A: “In Arabic, a car is called ‘siyara’. And in Arabic a window or door curtain is called ‘sitara’.”

B: “*Certainly, certainly,* why not” (LUWC_f041p0009317).

Another form of *żarūr* reduplication is *żarūr-bilżarūr*, approximately ‘most definitely, most certainly’. *Żarūr-bilżarūr* is used to emphasise a recommended best course of action, as in example (157).

157. Sāins ēk hī mōżū‘a=kē dō mutażād
Science one EXC topic=GEN.M.PL.OBL two opposite

naẓaryāt=kō kēsē taslīm kar saktī hai.
concept.PL=ACC how accept do can.IPFV.F.SG be.PRS.3.SG.

Kōī ēk dūsrē=kō *żarūr-bilżarūr* rad
Someone one other.OBL=ACC *most.definitely* reject

kartā hai aur phir rāij hōtā hai.
do.IPFV.M.SG be.PRS.3.SG and then establish be.IPFV.M.SG be.PRS.3.SG.

“How can science accept two opposing concepts on one single topic? *Most definitely* someone rejects one or the other, and then establishes it” (LUWC_f049p0106090).

Surprisingly given the otherwise limited role of reduplication in English, English-speaking addressers also reduplicate HCS MACs for emphasis, as shown in (158).

158. “Although the atmosphere was very flat overall, at the end of the match and as I was starting to turn it around, I could at least look up and see some faces in different points of the court to give me a little bit of encouragement, which *definitely, definitely* helped” (ECC_ind_a9700196)

Another instance of English HCS MACs performing the function of emphasis is when an addresser uses a MAC with determiner *such* or adverb *so* as a non-correlative (see 2.4.2.2). The term *non-correlative* is used by Huddleston and Pullum (2002, p.1276) for coordination markers that are not paired with other coordination markers. For instance, in ‘*either* on Monday *or* Tuesday’, *either* and *or* are correlatives, but in ‘Monday *or* Tuesday’, *or* is non-correlative. *Such* is non-correlative in (159).

159. “*Certainly such* state-sponsored savagery is sickening” (ECC_mail8801969).

Similarly, in Urdu, an HCS MAC near to non-correlative *aisā* 'such' or *itnā* 'so' emphasises an argument, as in (160).

160.

Aisī	rēyāsat=kō	*yaqīnān*	ēk	nākām	rēyāsat
Such.F	state=ACC	*definitely*	one	failed	state

kahā	jā	saktā	hai.
call.PFV.M.SG	go	can.IPFV.M.SG	be.PRS.3.SG.

"*Such* a state can *definitely* be called a failed state" (LUWC_f032p0033953).

In English, HCS MACs also cooccur with restrictive adverbials (e.g. *only*) to emphasise that a proposition is true by excluding any other possibility (see 2.4.2.2). Example (161) illustrates this use of *certainly* alongside restrictive adverb *especially*.

161. "Mr Biden was *certainly* more calm, *especially* when compared to Trump" (ECC_ind_b707273).

Similarly in Urdu, HCS MACs cooccur with restrictive adverbials such as *sirf* 'only', *khasūsān* 'especially/in particular', *khasūsī tōr par* 'especially/in particular', and *sirf/mehz/faqaṭ* 'just'. The data shows that HCS MACs cooccurring with these restrictive adverbials also reinforce the emphasis. In (162), the addresser asserts their proposition using *yaqīnān* and *sirf*; this excludes any other possibility being true and thereby increases the addresser's commitment to, and certainty in, the proposition.

162. Agar vēbsā‘aiṭ ēk kārpōrēṭ yā inṭarpriz nōi‘at=kī
If website one corporate or enterprise type=GEN.F.SG

vēbsā‘aiṭ hai tō *yaqīnān* yeh kām *sirf* ēk māhir
website be.PRS.3.SG then *certainly* DEM job *only* one expert

vēb dēvēlōpar hī anjām dē saktā hai.
web developer EXC accomplish give can.IPFV.M.SG be.PRS.3.SG.

“If a website is a corporate or enterprise type of website, then *certainly* this job can *only* be accomplished by an expert web developer” (LUWC_f067p0077250).

7.4 Solidarity

In English, addressers sometimes include certain HCS MACs (i.e. *certainly, obviously, of course*) alongside *you know* (not the discourse marker, but a full clause with *know* as main verb) to avoid sounding authoritative and to interactionally convey *solidarity* (see 2.4.2.4). They thus balance their power relation with their addressees, and avoid giving the impression of believing themselves superior due to their knowledge (Simon-Vandenbergen & Aijmer, 2007, p.205). In other words, an HCS MAC together with *you know* is a marker of solidarity, used to pre-emptively avoid irritating the addressee(s). This strategy helps to save face (see 7.8) for both addresser (who avoids appearing naïve in judging the addressee unaware of some information) and addressee (who avoids being labelled as unknowledgeable), as in (163).

163. “If a pre-laid trail were being followed properly, surely the hunt would never veer off into land they would be trespassing on, or, indeed, put the hounds and horses in danger by making the trail crossroads. Yet it happens all the time. But *of course you know* all this” (ECC_AskUK202005).

Similarly, in Urdu addressers sometimes utilise an HCS MAC within phrases such as *āp yaqīnān jāntē hōṁ gē* 'you of course will know' or *żarūr jāntē hōṁ gē* 'certainly you will know'. Sometimes, emphatic marker *tō* is added (Genady, 2005, p. 165). If an HCS MAC occurs within such a construction, then that MAC (together with *tō*, if present) puts stress on the addresser's assumption regarding the knowledge of the addressee. These MAC + *jāntē hōṁ gē* phrases in Urdu are hence used to imply that the addresser has the same viewpoint as the addressee and is not criticising them. MAC + *āp jāntē hōṁ gē* phrases are also added to sentences to covertly rebuke the addressee to implicitly challenge them to revise their viewpoint, as in (164). This construction can be paraphrased as it is obvious to both of us that *you certainly know where Gabol Town station* is and earlier lawlessness under its jurisdiction.

164.

Ḥaqā'aiq	yeh	haiṁ	ke	pichlē	tīn	
Fact.PL	DEM	be.PRS.3.PL	that	last.OBL	three	

sālōṁ=sē	Karācī=kē	har	thānē=kī
year.PL.OBL=from	Karachi=GEN.M.PL.OBL	every	police.station.OBL=GEN.F.SG

pōlīs	Lēyārī=kē	gaēg,	isṭarīṭ	karāimz	aur	dīgar
police	Leyari=GEN.M.PL.OBL	gang.PL,	street	crime.PL	and	other

jarā'aim=kē	khilāf	zabardast	kārvāīyāṁ	kar	
crime.PL=GEN.M.PL.OBL	against	fantastic	action.PL	do	

rahī	haiṁ.	Jis=kē	natījē=mēṁ	piclē	cand
PROG.PFV.F.SG	be.PRS.3.PL.	REL=GEN.M.PL	result.OBL=in	last.OBL	few

mahīnōṁ=sē	jarāim=kī	ṣa'arah=mēṁ	hērat-angēz
month.PL.OBL=from	crime.PL=GEN.F.SG	rate=in	incredible

kami	huī	hai.	*Āp*	*yaqīnān*	*jāntē*
decrease	be.PFV.F.SG	be.PRS.3.SG.	*You*	*certainly*	*know.IPFV.M.PL*

hōṁ	*gē*	ke	Gabūl	Ṭāun	thāna	kahān
be.SBJV.3.PL	*FUT.M.PL*	that	Gabool	Town	police.station	where

lagtā	hai	aur	agar	āp	vāqaī	Karācī=kē
attach.IPFV.M.SG	be.PRS.3.SG	and	if	you	really	Karachi=GEN.M.PL.OBL

hālāt=sē	zara	bhī	bākhabar	haiṁ	tō	
condition.PL=from	bit	INC	informed	be.PRS.3.PL	then	

haqāiq=sē	ānkhēṁ	na	cūrāīyē	gā	aur
reality.PL=from	eye.PL	NEG	steal.IMP.3.PL	FUT.M.SG	and

khud	hī	khālī	jaghēṁ	bhar	lēṁ.
self	EXC	empty	space.PL	fill	take.SBJV.3.PL.

"The facts are this, that for the past three years, the police of every station in Karachi have been doing a fantastic job against Leyari gangs, street crimes and other crimes. As a result of which, over the past few months there has been an incredible decline in the crime rate. *You will certainly know* where Gabool Town police station is located, and if you really are a bit informed of the conditions in Karachi, then don't act oblivious to reality and fill in the blanks yourself" (LUWC_f032p0001005).

An addresser sometimes adds an HCS or HCSNC MAC to a *you know* phrase as a positive politeness (see 7.8) and solidarity device. In (165), the addresser uses *of course you know* to acknowledge that both interlocutors in the discourse know *what it is*.

165. "Please focus your eyes on the + sign on the left, and try to identify the letter on the right of it (*of course you know already what it is*, but pretend for the moment that you do not): Is visualising the 'A' a tricky task for you?" (ECC_ind_b1776205).

In Urdu, addressers may use an HCS or HCSNC MAC to express solidarity with addressees without using phases like *āp jāntē hōṁ gē* 'you know', by making references to shared attitudes and knowledge about a common world which the MAC then emphasises. In (166), the addresser first recounts tragic events involving celebratory gunfire (lit. 'air firing'), laying the ground for the argument to come. Then, they use the HCS MAC *yaqīnān* 'certainly' with *ġērdānišmāna* 'unwise' to criticise that celebratory gunfire, implying that their viewpoint on this activity is an attitude shared with their addressees.

166.

Mulk	bhar=kē		chotē		baṛē		ṣehrōṁ=mēṁ					
Country	whole=GEN.M.PL.OBL		small.OBL		big.OBL		city.PL.OBL=in					
logōṁ=nē		Pākistān=kī		jīt=kī			khuṣī=mēṁ					
people.PL.OBL=ERG		Pakistan=GEN.F.SG		win=GEN.F.SG			joy=in					
ātaṣbāzī	kī		jubke		kuch		maqāmāt=par		havāī			
firework.PL	do.PFV.F.SG		whereas		some		place.PL=on		air			
fāēring	kar	kē		khūṣī		izhār		kīyā				
firing	do	do.CNT		happiness		express		do.PFV.M.SG				
gēyā.	Is		mōq‘a=par		kuch	afsōsnāk		vāqēyāt		bhī		
go.PFV.M.SG.	DEM.SG.OBL		occasion=on		some	saddening		incident.PL		INC		
pēṣ	āē		aur	is		havāī	fāēring=kē					
happen	come.PFV.M.PL		and	DEM.SG.OBL		air	firing=GEN.F.PL.OBL					
natījē=mēṁ	2	afrād		jāṁbaḷaq		jubke		Haidarābād,				
result=in	2	people		kill		whereas		Hyderabad,				
Piṣāvar,	Karācī		samaēt		mukhtalif		ṣēhrōṁ=mēṁ					
Peshawar,	Karachi		including		different		city.OBL.PL=in					
darjanōṁ	lōg		zakhmī		hō	gaē.						
dozen.PL.OBL	people		injured		be	go.PFV.M.PL.						
Khūṣī=kē			mōq‘a=par		is		ṭarah=kā					
Happiness=GEN.M.PL.OBL			occasion=on		DEM.SG.OBL		manner=GEN.M.SG					
muẓāhira	*yaqīnān*		ġērdānišmāna			hai.						
show	*certainly*		unwise			be.PRS.3.PL.						

“People in small and large cities across the country set off fireworks in joy at Pakistan’s win, whereas in some places, happiness was expressed through celebratory gunfire. On this occasion some sad incidents also came about, and in consequence of this celebratory gunfire two people died, while in different cities including Hyderabad, Peshawar, and Karachi dozens of people were injured. A show of this kind on an occasion of happiness is *certainly* unwise” (LUWC_e0232511).

7.5 Expectation

Some English HCS MACs, such as *of course* and *indeed*, function as expectation markers: in constructions in which an addresser expresses certainty about some state of affairs, use of an HCS MAC signals that the proposition about the future event can be taken for granted because it is expected to happen as asserted (Simon-Vandenbergen & Aijmer, 2007, p. 156). In (167), adding *of course* to *people will understand* expresses the addresser's expectation that their addressees understand that the world is changing due to technology. The addresser justifies that expectation in the next sentence by using *indeed* to introduce the proposition that *we* are already are using tech services as an example of the scenario in the previous sentence.

167. "People will *of course* understand that new tech-enabled services are transforming the way we buy, sell and manage our money. *Indeed*, many of these proved crucial as we adapted to the pandemic" (ECC_ind_b435733).

Similarly, in Urdu an HCS or HCSNC MAC, such as *żarūr* 'definitely' or *yaqīnān* 'of course', may act as an expectation marker in the same functional context. In (168), the addresser could have used the present tense *yaqīnān nahīṁ haiṁ* 'certainly (they) do not exist' to convey absolute certainty in their proposition. Instead, the future tense construction is used: *yaqīnān hōṁ gē nahīṁ* 'certainly (they) will not exist'. The use of *yaqīnān* strengthens the addresser's expectation, expressed as a prediction, that the addressee has no enemies – an inference from the addresser's knowledge about the addressee. Finally, the exclusive particle *hī* being added to the sequence firmly excludes any other possibility than the predicted scenario.

168. Darēṁ āp=kē duśman jō ke
Fear.SBJV.3.PL you=GEN.M.PL.OBL enemy.PL REL that

yaqīnān hōṁ gē hī nahīṁ.
certainly be.SBJV.2.PL FUT.M.PL EXC NEG.

"May your enemies, which you *certainly* won't have, be afraid"

(LUWC_f029p0029646).

Some background here is that in Urdu, it is not uncommon to express a desire for misfortune to afflict an addressee's enemies as a politeness strategy, based on the superstition that what is said out loud may come true. Therefore, misfortune is urged upon non-specific and/or imaginary enemies (e.g. *rōēṁ āp kē duśman* 'may your enemies cry'; *sunā hai ke āp kē duśmanōṁ kī tabīyat nāsāz̲ hai* 'it is heard that your enemies are unwell'). The addresser in (168) goes further by directly asserting that the enemies in question don't exist (by implication in the broader context, because the addressee is such a nice person).

There are instances in the English data where the addresser states some circumstance and then includes an HCS MAC in the subsequent expression of the expected consequences, as in (169).

169. "There's not an endless supply of top actors so *of course* they will be in multiple shows" (ECC_AskUK201810).

Similarly, in the Urdu data, there are instances where the addresser uses an HCS MAC to express that an expectation is based on a strong inference from a previously asserted state of affairs, as in (170).

170. Khairātī mistarī=kā kuch paisā phasā hai
Khairati mason=GEN.M.SG some money.M.SG stick.PFV.M.SG be.PRS.3.SG

hamārē pās is līē voh *żarūr*
1.PL.POSS.OBL possess DEM.SG.OBL for DEM *definitely*

āyē gā.
come.SBJV.3.SG FUT.M.SG.

"We have some money owing to mason Khairati, therefore he will *definitely* come" (LUWC_ f029p0018102).

In English, the addresser may use an HCS MAC to express confidence in an inference drawn from "the predictability of a state of affairs" based on their "past experience" about some situation or logical assumptions (Simon-Vandenbergenan & Aijmer, 2007, p. 291). In example (171), based on the past experience of in the early days of the Covid-19 crisis, the addresser uses *no doubt* and *certainly* to express the inevitability of the expected consequence (facing a *long road*).

171. "As the Covid-19 pandemic continues to reverberate globally, there is no doubt that we must be ready to face a long road ahead, *certainly* beyond the end of this year. (gdn_commentisfree_2020_sep_06_lets_get_real_no_vaccine_will_work_as_if_by_magic_returning_us_to_normal).

The same can be observed for Urdu, as in (172), where an addresser's previous experience with a product gives them confidence regarding an expected outcome.

172. Apdēṭs=kē ṭarīqakār=kō māzī=kē pērāē=mēṁ
Update.PL=GEN.M.PL.OBL process=ACC past=GEN.M.PL.OBL context=in

dēkhā jāē tō *yaqīnān* yeh kahā jā
see.PFV.M.SG go.SBJV.3.SG then *definitely* DEM say.PFV.M.SG go

saktā hai ke āīfōn sārēfīn=kō bhī kuch zyāda dēr
can.IPFV.M.SG be.PRS.3.SG that iPhone user.PL=ACC INC some more late

intezār	nahīṁ	karnā	paṛē		gā	aur	juld	hī
wait	NEG	do.INF.M.SG	obliged.SBJV.3.SG		FUT.M.SG	and	soon	EXC
voh	bhī	in	na‘aē	fīcarz=sē		mustafīd	hō	
DEM	INC	DEM.PL.OBL	new	feature.PL=from		benefit	be	
sakēṁ		gē.						
can.SBJV.3.PL		FUT.M.PL.						

"If the update process is seen in the context of the past, then *definitely* it can be said that iPhone users will not have to wait for much longer and soon they too will be able to benefit from these new features" (LUWC_f067p0086599).

In Urdu, the combination of an (oblique case) infinitive and *vālā* is used to indicate an "imminent action or event" (Schmidt, 1999, p. 139), that is, something near-future; equivalent phrases are the English semi-modal (*be*) *going to* and/or (*be*) *about to*. When an addresser adds an HCS or PS MAC to this infinitive + *vālā* formulation, it expresses the addresser's (higher or lower) degree of expectation of the actuality of the predicted event or action. Similarly in English, an HCS or PS MAC which cooccurs with *will, would,* or *going to* conveys the addresser's degree of confidence in their prediction of the future event (see 6.6.1). The English data shows that, in fact, all the HCS and PS MACs can cooccur with *will, would* and/or *going to*. Meanwhile, the Urdu data shows that MACs *śāyad* 'perhaps' and *yaqīnān* 'definitely' are preferentially used with the infinitive + *vālā* construction (in the sense of *going to*), relative to other MACs. Examples (173) and (174) illustrate this use of a PS MAC alongside *going to* and infinitive + *vālā* respectively, to indicate only tentative confidence in the predicted near-future event.

173. "If that's not happening, it's *possibly going to* make restaurants shut down" (ECC_mail18721465).

174. Mīyāṁ Sāhib=kē bā‘az khērkhavāhōṁ=kā
Mian Sahib=GEN.M.PL.OBL some wellwisher.PL.OBL=GEN.M.SG

kahnā yeh hai ke is haftē *śāyad*
say.INF.M.SG DEM be.PRS.3.SG that DEM.SG.OBL week.OBL *possibly*

ʿadālat-ē-‘aūzmā=kā fēslā *ānē* *vālā* *hai*.
court-of-supreme=GEN.M.SG decision *come.INF.OBL* *VALA.M.SG* *be.PRS.3.SG*

“Some well-wishers of Mr Mian say that *possibly* the supreme court’s decision is *going to* come out this week” (LUWC_j0252716).

On the other hand, (175) illustrates that adding an HCS MAC to *(be) going to* may express an addresser’s strong expectation. In this case that expectation arises from the logical assumption that when government *balances the books*, a rise in taxation of some commodities is to be predicted.

175. “If they are also serious about what Mr Sunak has described as ‘the sacred duty to balance the books’, there are *certainly going to* have to be tax rises somewhere down the line, Mr Zaranko said” (ECC_ind_b175970).

In example (176), the addresser uses *yaqīnān* with infinitive + *vālā* to express their high confidence in the expected outcome of the earlier prediction.

176. Āp *yaqīnān* tanqīd=par ūksānē *vālē* hōṁ gē.
You *definitely* criticism=on prompt.INF.OBL *VALA.M.SG* be.SBJV.3.PL FUT.M.PL.

“You are *definitely going to* prompt criticism” (LUWC_f256p0077079).

7.6 Counter-expectation

Addressers may use MACs to convey certainty that the nature of a proposition is counter to what is generally expected in a given context (see 2.4.2.6). Two strategies, *concession* and *adversativity*, are commonly used to communicate unexpected or new

information that results in a contrast between the addresser's proposition and an expected normative viewpoint (Mortier & Degand, 2009, p. 305).

A *concession* is defined in grammatical terms by Leech (2006, p. 24) as "an adverbial clause which expresses a contrast of meaning or implication 'unexpectedness' in its relation to the matrix clause". On the pragmatic level, concession need not involve this specific structure with an adverbial dependent clause (as the examples to be discussed illustrate). Rather, concession is a communicative structure in which an addresser acknowledges one proposition and then asserts another proposition which is unexpected given the conceded proposition. Verhagen (2000, p. 367) says that the addresser first asserts a proposition that leads the addressee to draw a causal inference 'X therefore Y', based on their past experience of similar circumstances. However, counter to expectations, the addresser presents the second proposition that denies 'Y' as a valid causal inference (Verhagen, 2000, p. 367).

In (177), the addresser uses first person plural *we* to advance their position as a member of the group being addressed, and, therefore, sharing that group's sentiments. This addresser, discussing the improved situation in Italy since COVID 19 began, concedes that correct actions have been taken, using the MAC *certainly*. This sets up the expectation that the addresser thinks "our" situation is improved and "we" are much safer than in March 2020 (a period of the COVID pandemic during which Italy experienced very high death tolls). Then, however, the addresser uses *but* to introduce a contrast, and counters the assumption of safety.

177. "We are now in a much better situation than in March, so *we* have *certainly* done something right, *but* the infection numbers are rising again" (ECC_ gdn_commentisfree_2020_sep_28_italy_covid_19_response_sweden_coronavirus).

In the Urdu example (178), the addresser begins by stating that people around the world are spending a lot of money to be safe from the virus (i.e. coronavirus), then says that the disease is not ending, and people are dying. The latter proposition is contrary to an expectation set up by the former, that expectation being that spending a lot of money would have an effect. By adding an HCS MAC, the addresser emphasises the contrast between ineffectual efforts to end the disease and their certainty that people are dying.

178.	Is	waqt	tamām		dunyā	is	vāiras=sē	baccnē=kē	
	DEM	time	all		world	DEM	virus=from	safe.OBL=GEN.M.PL.OBL	
	līē	paisā	tō	kharac		kar	rahī	hai,	is=sē
	for	money	EMPH	spend		do	PROG.PFV.F.SG	be.PRS.3.SG,	DEM=from
	bēmārī		tō	khatam		nahīṁ	hō	rahī	*magar*
	disease		EMPH	end		NEG	be	PROG.PFV.F.SG	*but*
	lōg		*żarūr*		mar	rahē		haiṁ.	
	people		*definitely*		die	PROG.PFV.M.PL		be.PRS.3.PL.	

"At present, the whole world is spending money to be safe from this virus, this is not ending the disease *but* people are *definitely* dying" (LUWC_j0757555).

A second counter-expectation strategy is *adversativity*, often considered a subclass of concession (Malchukov, 2004). Martin and White (2005, p. 120) say that formulations labelled as *adversative* are in fact pragmatically used to "invoke a contrary position which is then said not to hold". This contrary proposition is in direct "contradistinction" to the one that generated the addressees' expectation. That is, the addresser gives two contrasting propositions to force their addressees to interpret pragmatically the (assumed) incompatibility between the two. By adding a MAC in the clause expressing the second proposition, the addresser adds strength to their assertion that their contradictory proposition is true, countering the addressees' expectation. In (179), the addresser's use of *might* in the first

clause shows that they are treating that clause's proposition as a possibility, and thus aligning their viewpoint with those of the addressees. But in the subsequent clause, after *but,* they use *undoubtedly* to express certainty in this latter clause's (contrasting) proposition. Martin and White (2005, p. 124) say that it is typical for the addresser first to concur (concede) "as a precursor to a countering"; thus this move is considered a concession strategy. In (179), the addresser first acknowledges the previous addresser's negative assessment of a former prime minister, then dismisses that proposition by adding *undoubtedly,* a relatively high degree of commitment to the countered proposition, in their contrasting positive assessment. In this concede + counter pairing, *might* has the conceding function and *undoubtedly* has the countering function.

179. "She *might* not have been a saint, *but* she was *undoubtedly* one of our better prime ministers" (ECC_AskUK201906).

Example (180) illustrates a contrast in two propositions linked by a contrastive clause marker. Medically, a person develops antibodies to a pathogen upon exposure – that is, after becoming infected or being vaccinated. Here, the addresser commits to two propositions expected to be incompatible: *is not a covid patient* and *has antibodies*. The addresser is discussing the case of a newborn child who had antibodies at birth, which according to doctors his mother transferred to him when infected by COVID-19 during pregnancy. The addresser adds *żarūr* in the *but*-clause to emphasise the contrast, thus strengthening the rejection of the addressees' assumed expectation that only those who contract the virus can develop antibodies.

180. *Agarca* yeh bacca kōvaiḍ19=kā marīz nahīṁ lēkin
Although DEM child covid19=GEN.M.SG patient NEG but

is=mēṁ aintī bāḍīz żarūr mōjūd haiṁ
DEM=in antibody.PL definitely present be.PRS.3.PL

"Although this child isn't a covid19 patient, antibodies are *definitely* present in him" (LUWC_ e2110974).

7.7 Hedges

In English, addressers use PS or PSNC MACs to *hedge*, that is, to make the expression of a proposition more tentative. One reason they do this is to make their assertion acceptable to addressees who, it is assumed, would be unamenable to agreeing with a more directly expressed equivalent statement. Addressers also use hedges to avoid commitment to a proposition of which they are not certain (see 2.4.3.7). Moreover, if the topic under discussion is sensitive, an addresser may use a hedge involving a MAC to mitigate the force of their proposition so as to avoid taking a face-threatening tone (see Coates, 2015, p.90; and see further 7.8). In example (181), the addresser argues in defence of teachers making racist comments in the classroom, but hedges by describing such comments as *perhaps mistaken*. The background context is that there is high intolerance for overtly racist attitudes in the UK especially in educational institutions. Therefore, instead of saying outright that the teachers should be allowed to say racist things, the addresser frames the argument differently. By saying that UK law should not ban teachers from saying *controversial* and *mistaken things*, the addresser hopes to make the proposition more acceptable for the addressees. Framed thus, the addresser concedes that racist speech is *controversial* and *mistaken* to distance themself from racist sentiment. However, the addresser adds *perhaps* before *mistaken things*, because they are not willing to fully concede, or permit the interpretation, that racism is *mistaken* (because *perhaps mistaken* = *probably not mistaken*). Thereby the addresser hedges the

distancing apparently achieved by *controversial* and *mistaken*, to reembrace a pro-racist stance.

181. "Put simply, this means there are now laws in this country preventing teachers from saying certain things- not incitement to violence or stupid rabble-rousing bigotry, just controversial and *perhaps* mistaken things that the dominant elite in our society have decided are offensive" (ECC_mail9021775_debate_article-9021775_PETER).

In Urdu too, there are instances in the data where addressers purposely use a PS MAC as a hedge when a sensitive topic is under discussion. Example (182) comes from a discussion of two well-known Pakistani singers, one highly regarded for her classical singing, the other known for singing vulgar songs in stage shows. In addition to using a PS MAC, the addresser minimises their commitment to the truth of the proposition by starting with *lōg kahtē haiṁ* 'people say', a common phrase in Urdu when an addresser wants to distance themselves from overtly claiming a statement as their own. In this example, the addresser does not wish to present it as their own opinion that stage dramas are immoral and indecent. Then, the addresser uses a MAC *darasal* 'in fact' to impress upon the addressee that what they are saying is the truth, and adds *śāyad* 'perhaps' before *zūmānī jūgtēṁ* 'double entendre humour' in an attempt to mitigate the force of their claim of immoral material on stage. By using *śāyad* 'perhaps', the addresser also tries to non-committedly propose a possible explanation for the negative reactions *of others* towards stage plays. The addresser adds *śāyad* 'perhaps' again at the end of the clause, and then begins the next clause with another PS MAC, *ġālibān* 'possibly'. This hedging strategy helps the addresser to save their own face as well as that of their addressee (see 2.4.2.6), and leave the addressee with room for argument.

182. Āp=kō muāśrē=mēṁ buhat=sē lōg yeh kahtē bhī
You=ACC society.OBL=in many=from people DEM say.IPFV.M.PL INC

milēṁ gē ke *darasal* in darāōṁ=mēṁ
meet.SBJV.3.PL FUT.M.PL that *in.fact* DEM.PL.OB drama.PL.OBL=in

śāyad zūmānī jūgtēṁ hōtī haiṁ aur
perhaps double.entendre joke.PL be.IPFV.F.SG be.PRS.3.PL and

in=mēṁ laṛkīyāṁ raks kartī haiṁ *śāyad*
DEM.PL.OBL=in girl.PL dance do.IPFV.F.SG be.PRS.3.PL *perhaps*

is līē *ġālibān* dalīl yeh hai ke Ṭālibān
DEM.SG.OBL for *possibly* reason DEM be.PRS.3.SG that Taliban

Iqbāl Bānō marhūm=kō tō bardāśt kar liēṁ gē
Iqbal Bano deceased=ACC EMPH bear do take.SBJV.3.SG FUT.M.PL

lēkin *āp* *khud* *hī* *batāēṁ* kyā Nasībō Lāl=kē
but *you* *self* *EXC* *tell*.IMP.2.PL Q Naseebo Lal=GEN.M.PL.OBL

gānē āp faimilī=kē sāth sūn saktē haiṁ?
song.PL.OBL you family=GEN.M.PL.OBL with listen can.IPFV.M.PL be.PRS.3.PL?

"You will meet many people in society saying that *in fact* in these dramas there are *perhaps* risqué jokes and girls dance in them *perhaps* that's why… *Possibly* the reason is that the Taliban will tolerate the late Iqbal Bano, but *you tell me*, can you listen to Naseebo Lal's songs with your family?" (LUWC_f032p0021819).

In English, an addresser may also use a PS or PSNC MAC to mitigate criticism of the addressee. Insertion of a PS MAC in this way proposes that ignorance is the reason for the interlocutor holding an incorrect view, while the addresser, who is more knowledgeable than interlocutor, has the correct view, as in example (183). It is interesting to note that typically in English *understand* would be non-progressive here (*don't understand*). Putting *understand* in the progressive is also a mitigation strategy to focus on addressee's *not understanding* as a temporary state at the current moment; it supports the mitigating function of *perhaps* here.

183. "*Perhaps* you're not understanding that a balanced diet will give you *some* carbs" (ECC_AskUK202005).

In Urdu a PS or PSNC MAC is sometimes used to downtone the harshness of negative evaluations in a proposition, as in (184). The addresser starts by criticising the whole staff of the foreign affairs department, then adds *ġālibān* 'possibly' before the negative word *kōtāhī* 'negligence' to soften the proposition by implying that this may only be one among many reasons.

184. Hamārē sifāratī hukām=kā ravaīya apnē
1.PL.POSS.OBL diplomatic official.PL=GEN.M.SG attitude POSS.PL.OBL

hamvatnōṁ=kē sāth munāsib nahīṁ hōtā balkeh
countryman.PL.OBL=GEN.M.PL.OBL with appropriate NEG be.IPFV.M.SG rather

bērūn-ē-mulk qāim hamarē sifārat-khānōṁ=kā
external-of-country established 1.PL.POSS.OBL embassy.place.PL.OBL=GEN.M.SG

andāz Pākistānī thānōṁ=kē mumāsil hōtā
way Pakistani jail.PL.OBL=GEN.M.PL.OBL similar be.IPFV.M.SG

hai. Is=mēṁ *ġālibān* hamarē ālā
be.PRS.3.SG. DEM.SG.OBL=in *possibly* 1.PL.POSS.OBL elite

hukām=kī kōtāhī=kā bhī 'amal dakhal
high.official.PL=GEN.F.SG negligence=GEN.M.SG INC action authority

hai jō dēyān nahīṁ raktē ke un
be.PRS.3.SG REL attention NEG keep.OBL that DEM.PL.OBL

mumālik=mēṁ jahāṁ Pākistānī mehnatkaśōṁ=kī tadād
country.PL=in where Pakistani labourer.PL.OBL=GEN.F.SG number

lākhōṁ=mēṁ hōtī hai vahāṁ=kē sifārat-khānē
million.PL.OBL=in be.IPFV.F.SG be.PRS.3.SG there=GEN.M.PL.OBL embassy-place.OBL

'amlē=kī qillat=kā śikār rahtē haiṁ.
staff.OBL=GEN.F.SG shortage=GEN.F.SG victim stay.IPFV.M.PL be.PRS.3.PL.

"Our diplomatic officials' attitude to their countrymen is not appropriate; rather the behaviour of our embassies in foreign countries is like Pakistani jails. *Possibly* the

negligence of our elite officials also has a role in this, who don't pay attention: in those countries where Pakistani labourers number in their millions, embassies there are continuing to be the victim of staff shortages" (LUWC_e0593120).

In both English and Urdu, an addresser's PS MAC may also function as a hedge associated with lack of confidence in the correctness of their claim. In English, one typical use is when an addresser questions the addressee on whether their interpretation of the addressee's earlier discourse is correct, as in example (185, repeated from 98).

185. A: "I expect you're right – he probably is American. Just didn't know England didn't use 'high school'."

B: "What do you use instead? 'Secondary school' *perhaps*?" (ECC_AskUK201402).

In the Urdu data, a common hedge is *śāyad mumkin na hō* 'perhaps it is not possible'. In (186), *śāyad mumkin na hō* expresses the addresser's uncertainty about the future of Afghanistan. The addresser justifies their stance from circumstantial evidence. At no point in the argument does the addresser state that argument to be their personal opinion. They also do not mention any specific source, using passive *yeh kahā jā rahā hai* 'it is being said', which is used in Urdu by addressers to distance themselves from a proposition without directly attributing it to anyone else.

186.	Agar	voh	Afġān		hakūmat=mēṁ		apnī	ḥēs̱īyat		aur	
	If	DEM	Afghan		government=in		REFL.POSS.F.SG	status		and	
	ṭāqat=sē		baṛh		kar	zyāda	śarākat=kē			ṭalubgār	
	power=from		increase		do	more	partnership=GEN.M.PL.OBL			seeker	
	haiṁ		tō	aisā	mūmkin		nahīṁ.	*Yeh*	*kahā*		
	be.PRS.3.PL		then	such	possible		NEG.	*DEM*	*say*.PFV.M.SG		

jā	*rahā*	*hai*	ke	agar	Amrīkā	Afġānistān=kō
go	*PROG.PFV.M.SG*	*be.PRS.3.SG*	that	if	America	Afghanistan=ACC

khālī	kar	dētā	hai	tō	aisā	karnā
vacate	do	give.IPFV.M.SG	be.PRS.3.SG	then	such.M.SG	do.INF.M.SG

us=kē			*śāyad*	*mumkin*	*na*	*hō*
DEM.SG.OBL=GEN.M.PL.OBL		for	*perhaps*	*possible*	*NEG*	*be.SBJV.3.SG*

kīūṁke	aisī	sūrat=mēṁ	ʻalamī	saṭaḥ=par
because	such.F.SG	situation=in	international	level=on

us=kī	sūbkī	hō	gī.
DEM.SG.OBL=GEN.F.SG	embarrassment	be.SBJV.3.SG	FUT.F.SG.

"If they [the Taliban] are seeking a partnership in the Afghan government to further increase their own status and power, then *that's not possible*. *It's being said that* if America vacates Afghanistan, then *perhaps it would not be possible* for America to do so [empower the Taliban] because in such a situation it would be an embarrassment for them [America] on the international level" (LUWC_e0488867).

7.8 Politeness

Politeness refers to the set of social norms that prescribe, or at least positively evaluate, certain behaviours among interlocutors. Politeness is considered a key "interpersonal interactional phenomenon", due to the fact that it helps people to build up and maintain interpersonal relationships (Kádár, 2017, p. 1). A positive evaluation of people's behaviour arises when they behave in a manner that is perceived to be in alignment with the (assumed) shared values of a society. That is, when an interaction among interlocutors is congruent with the norms, it is a polite interaction. An interaction receives a negative evaluation, and therefore is impolite, when the communicative behaviour of some or all participants in that interaction is contrary to the established norms (Fraser, 1990, p. 220).

A related term, *face*, is defined by Brown and Levinson (1978, p. 66) as the "public self-image that every member [of a social group] wants to claim for himself". Face can be *positive*, in which case it references the desire of the interactants to be approved and appreciated by others, or *negative*, referring to the desire to be free of impositions from other people (Sifianou, 2010, p. 42). In some situations the addressee might feel embarrassed and consider the interaction with addresser as a threat to their positive face. To pre-empt any such face threat, addressers sometimes use redressive strategies to maintain addressees' face (Brown & Levinson,1987, p. 65). On the other hand, the addressee may feel threat to their negative face, that is, they might feel that the addresser is trying to interfere with their freedom of action. In that case, the addresser mitigates the threat to addressee's face by use of hedges (see 7.7). Existing research has shown that HCS and PS MACs are such politeness devices that are used to reduce any damage to both positive and negative face (see 2.4.2.6 and 2.4.2.8).

The politeness function overlaps with certain other pragmatic functions, such as solidarity, expectation and hedging. Therefore, politeness has been mentioned already in some previous parts of this analysis chapter (see 7.4, 7.5, 7.7). In this section specifically concerned with politeness, I expand in this section primarily on a discussion of politeness strategies, with special focus on the role of *face* in politeness strategies.

Addressers use HCS MACs (e.g. *of course*) as a positive politeness strategy to downplay the addresser's superiority as a possessor of information. In example (187, repeated from 163), the addresser uses *of course you know* to present the preceding information as shared knowledge, rather than a challenge to the addressee's knowledgeability. This positive politeness strategy preserves the addressee's dignity.

187. “If a pre-laid trail were being followed properly, surely the hunt would never veer off into land they would be trespassing on, or, indeed, put the hounds and horses in danger by making the trail crossroads. Yet it happens all the time. But *of course you know* all this” (ECC_AskUK202005).

Urdu HCS MACs are also used pragmatically for politeness. In (188), as a preamble to a turn in an ongoing argument, the addresser first acknowledges that the addressee is widely read and well-informed, and then states *khamśī bhetar samajhtā hūṁ* ‘I consider it better to stay silent’. But they do not in fact stop there. They continue with *bus sirf yeh kahōṁ gā* ‘I will say only this’. Clearly, this addresser does not actually want to refrain from giving their opinion. But to respect the addressee’s face, they express, sincerely or otherwise, a strong belief in the addressee’s superior knowledge, and minimise the extent and importance of their own view (disclaiming imposition). The use of *yaqīnān* ‘certainly’ in their opening praise of the addressee underlines that belief and thus strengthens the mitigation of the threat to the interlocutor’s face.

188.

Kahnē=kō	tō	bōhat	kuch	hai	aur	*yaqīnān*	āp	
Say.INF.OBL=ACC	EMPH	very	much	be.PRS.3.SG	and	*certainly*	you	
jaisē	vasīh	mutāli‘ah		vālī	aur	saḥib-e-baṣīrat		
like.OBL	extensive	reading		VALA.F.SG	and	person-of-vision		
śakhsiyat=kē		pās		tō	kahīṁ	zyāda		
personality=GEN.M.PL.OBL		possession		EMPH	somewhere	more		
hō	gā	magar	khamśī		bhetar	samajhtā		
be.SBJV.3.SG	FUT.M.SG	but	silence		better	consider.IPFV.M.SG		
huṁ.	Bus	sirf	yeh	kahōṁ	gā		ke	
be.PRS.1.SG.	Enough	only	DEM	say.SBJV.1.SG	FUT.M.SG		that	
hilāl,	khajūr=kā		darakhat		aur	ūnth		
crescent,	date=GEN.M.SG		tree		and	camel		
musalmānōṁ=kē		taśakhus=kī			akāsī	kartē	haiṁ.	
Muslim.PL.OBL=GEN.M.PL.OBL		identity=GEN.F.SG			reflect	do.IPFV.M.OBL	be.PRS.3.PL.	

Aur	sub	hī	ṭēṛēh	haiṁ.	Kisī	ġairmunāsib
And	all	EXC	unstraight	be.PRS.3.PL.	Some	inappropriate

bāt=kē	liē	ma‘azratkhavah	hūṁ.
talk= GEN.M.PL.OBL	for	sorry	be.PRS.1.SG.

“There is much to be said, and *certainly* an extensively read and visionary personality like you will have more to say, but I consider silence better. I will say only this, that the crescent, date tree, and camel reflect Muslim identity. And all are curved. I am sorry for any inappropriate talk” (LUWC_f202p0076982).

Sometimes an HCS MAC may be used in conjunction with *you know* as a negative politeness strategy to pre-emptively mitigate a face-threatening act (see 2.4.3.7; see also Holmes, 1988; Sifianou, 2010) that might otherwise imply that the addressee is ignorant, as in (189).

189. “As far as I know - Waffle Houses aren’t even nationwide here. Despite that - there is is this : > [Waffle House Index] (https : //en.wikipedia.org/wiki/WaffleHouseIndex) > > The Waffle House Index is an informal metric used by the Federal Emergency Management Agency (FEMA) to determine the impact of a storm and the likely scale of assistance required for disaster recovery. *Of course, you might know that already*” (ECC_AskUK201509).

Similarly in Urdu, an HCS or HCSNC MAC may be used to mitigate a threat to the addressee’s face, as in (190), where the addresser wants to reduce the impact of appearing as a superior authority by using an HCS MAC with phrases such as *patā hō gā* ‘will know’ or *jāntē hōṁ gē* ‘will know’ to convey that they are restating information to an addressee despite that addressee being, they impute, already familiar with such information.

190. Agar āp maraqabē=sē vāqif haiṁ tō nīndh=kō
If you meditation=about know be.PRS.3.PL then sleep=ACC

bhetar banānē=sē lē kar tanāō=kō kam karnē
improve make.INF.OBL=from take do stress=ACC reduce do.INF.OBL

yanī zehnī tandrūstī=kē havālē=sē is=kē
meaning mental health=GEN.M.PL.OBL reference=from DEM.SG.OBL=GEN.M.PL

fōvaed=kē bārē=mēṁ āp=kō *yaqīnān* patā
benefit.PL=GEN.M.PL.OBL about=in you=ACC *certainly* know

hō gā.
be.SBJV.3.SG FUT.M.SG.

"If you know about meditation then *you will certainly know* that it [meditation] has many benefits from improving sleep to reducing stress, that is, from the perspective of mental health" (LUWC_j0824109).

Likewise, an Urdu PS or PSNC MAC may *also* reinforce the addressee's face by disclaiming an implication that the addressee is ignorant of something. In effect, these examples of the form *possibly you don't know X* are variants of the *certainly you know X* examples just discussed. Both are ways of being polite about saying X, when the fact of X needing to be said implies that the addressee does not know X, which is a threat to positive face. For instance, in (191), the addresser avoids presenting themselves as authority on a subject by adding *ġālibān* 'probably' to the sequence *āp kō 'alm nahīṁ* 'you don't have the knowledge' to convey that they are not asserting that the addressee is clueless, but rather that they might lack this particular information.

191. Āp=kō *ġālibān* *'alm* *nahīṁ* ke Sīālkōṭ=kē
You=ACC *probably* *knowledge* *NEG* that Sialkot=GEN.M.PL.OBL

tājirōṁ=nē Sīālkōṭ=mēṁ apnī madad āp=kē
trader.PL.OBL=ERG Sialkot=in REFL.POSS.F.SG help you=GEN.M.PL.OBL

tēḥat bēn-al-aqvāmī havāī mustaqar tāmīr karvāyā
basis international air accommodation build do.CAUS.INF

hai	aur	is=par	bōhat	kum	lāgat	āī
be.PRS.3.SG	and	DEM.SG.OBL=on	very	few	cost	come.PFV.F.SG

hai
be.PRS.3.SG

"*Probably you don't know* that Sialkot's businessmen have had an international air travellers' accommodation built on a self-help basis in Sialkot, and it has cost very little" (LUWC_f202p0059936).

I did not find equivalent use of a PS or PSNC MAC in this context in the English data. An idiomatic equivalent in English would be *you may/might know that* (i.e. with an MV, not a MAC).

Politeness-related strategy can be observed in both languages when PS MACs are used to politely state a difference of opinion. In such cases, the MAC's main, epistemic function shifts instead to express the face-attending politeness function. Examples (192) and (193) each illustrate an addresser politely countering another's opinion in such a way as to maintain the addressee's positive face. In example (192, repeated from 78), A avoids directly stating a conflict of opinion and therefore simply replies with *perhaps*, rather than a contradiction, which could be a face threat.

192. A: "On land it's meant to be flag rather than jack yes".

B: "Not according to the admiralty or the government. Our national flag is the union jack"

A: "*Perhaps*" (ECC_AskUK202005).

In (193), B's *śāyad darūst hō* 'possibly be correct' expresses that what A said might be true for some but not everybody. Thus, B concedes partially, without committing to A's proposition, rather than disagreeing openly and outright (and thus challenging A's face). In

this way B leaves open an avenue for themself and others to disagree with A's assertion without being impolite.

193.	A:	Sab=sē	acchē		lamḥāt	yehī	tō	hōtē
		All=from	good.PL		moment.PL	DEM	EMPH	be.IPFV.M.PL
		haiṁ	baccpan=kē,			jab	sab	każanz
		be.PRS.3.PL	childhood=GEN.M.PL.OBL,			when	all	cousin.PL
		mil	kar	khūb	dīġarā	macātē		haiṁ
		together	do	much	cacophony	make.IPFV.M.PL		be.PRS.3.PL.
	B:	*Śāyad*	*darūst*		*hō*	yeh	bāt	lēkin	mērā
		Possibly	correct		*be.SBJV.3.SG*	DEM	talk	but	1.SG.POSS
		tajruba	nahīṁ		hai.				
		experience	NEG		be.PRS.3.SG.				

A: "Childhood is the best time of all, when all cousins get together to create a cacophony".

B: "This talk is *possibly correct*, but I don't have any experience" (LUWC_f041p0067580).

Example (194) illustrates the use of a PS and a PSNC MAC together to give an explicitly non-committal response. A PS and a PSNC MAC together function as a neutral response marker in instances when a simple yes or no response may not be considered polite; the combination allows a firm commitment, which might threaten face, to be avoided.

194. A: "Look for ones that don't contain petroleum if that matters to you."

B: "I second Burt's Bees and the ones from Lush."

A: "*Maybe, maybe not*" (ECC_AskUK201701).

In Urdu, a typical non-committal response of the same kind is the words for *yes* and *no*, each modified by a PS MAC (which, in the latter case, creates a PSNC MAC phrase). In (195), B effectively claims that A's stated desired outcome has an equal chance of being true

or false. Urdu speakers typically use this construction when they do not want to endorse another's stated prediction but wish to avoid directly gainsaying it for the sake of politeness.

195.	A:	Umīd	hai	is=mēṁ	Pakistānī	nizam
		Hope	be.PRS.3.SG	DEM.SG.OBL	Pakistani	system
		ʻadal=par	sehatmund	baḥas	hō	gī
		justice=on	healthy	discussion	be.SBJV.3.SG	PFV.F.SG
	B:	*Śāyad*	*hāṁ* yā	*śāyad*	*nahīṁ*	
		Perhaps	*yes* or	*perhaps*	*NEG*	

A: "Hopefully there'll be a healthy discussion on the Pakistani justice system here."

B: "*Perhaps yes, perhaps no*" (LUWC_f029p0104798).

Another neutral response in Urdu is the PS MAC *hō saktā* 'maybe' modified by insertion of the inclusive emphatic particle *bhī* . The result, *hō bhī saktā* 'can also be', stresses the addresser's commitment to truth of the proposition . Sometimes, *hō bhī saktā* is followed by *aur nahīṁ bhī* 'and also not'. This combination can be paraphrased as *it is a possibility and also not a possibility*; like *śāyad hāṁ yā śāyad nahīṁ*, it can be used to give a non-committal response and avoid straightforward refusal. In (196), A does not want to talk about some matter, but also does not want to reply with a point-blank *no*. When A's first attempt to demur succeeds only in putting off B's interest till tomorrow, A responds with *hō bhī saktā hai aur nahīṁ bhī*.

196. A: Bāt abhī batānē vālī nahīṁ hai.
Matter now tell.INF.OBL VALA.F.SG NEG be.PRS.3.SG.

B: Yanī ke āp kal batēṁ gē?
Meaning that you tomorrow tell.SBJV.M.PL FUT.M.PL?

A: *Hō bhī saktā hai* aur *nahīṁ bhī*.
Be INC can.IPFV.M.SG be.PRS.3.SG and *NEG INC*.

A: "The matter is not for telling now."

B: "That means you will tell tomorrow?"

A: "*Perhaps, perhaps not*" (LUWC_f041p0001132).

7.9 Chapter summary

In this chapter, I have analysed the rhetorical-pragmatic functions of English and Urdu MACs. I have established that, although there are minor differences, in the main the pragmatic usage of MACs is similar across English and Urdu. My findings on rhetorical-pragmatic functions have been informed both by the corpus and by reference to prior analyses of certainty markers (Simon-Vandenbergen & Aijmer, 2007) and possibility markers (Van der Auwera & Plungian, 1998; Suzuki & Fujiwara, 2017). My findings show that, in both English and Urdu, MACs are pragmatically polyfunctional. All types of MACs function as emphasis, expectation, counter expectation, and politeness devices. Adding an HCS MAC intensifies the degree of emphasis on the truth value of the proposition, whereas adding a PS MAC downtones that degree of emphasis. Moreover, HCS and HCSNC MACs are used with authority and solidarity functions. PS and PSNC MACs may be used as hedges. PS and PSNC MACs may also function as politeness devices, that is, a means by which to mitigate an anticipated threat to the addressee's face. There is some overlap in these pragmatic functions and therefore for a single instance of a MAC there may be simultaneous valid readings of expectation, hedge, politeness and solidarity.

There are, however, some specific differences between English and Urdu in the pragmatics of MACs. For instance, the Urdu expression *śāyad nahīṁ yaqīnān* 'not perhaps, certainly' is used by addressers when they are confident in their authoritative knowledge of some issue. No equivalent sequence of PS + NEG + HCS has been observed in English. The most similar pattern found in the English data is that pattern where an addresser's response consists of *not perhaps* followed by an authoritative assertion (see 7.2). In the English data, sequences such as *you possibly don't know*, used to disclaim a face-threatening implication that the addressee is ignorant, do not occur, though such examples do occur in the Urdu corpus. Another difference is that, in Urdu, MACs do not cooccur with degree adverbs (e.g. *most certainly*), a form which is reasonably prevalent in English. Perhaps in lieu of such reinforcing modification, one HCS MAC, *żarūr* 'definitely' is sometimes used in reduplicated form – reduplication as a morphological operation has a variety of functions in Urdu. Interestingly, however, only one Urdu HCS MAC occurs in reduplicated form (*żarūr*), whereas the data shows that all English HCS MACs may be reduplicated for emphasis (see 7.3).

8 Discussion: Similarities and differences between MACs in English and Urdu

8.1 Chapter overview

The discussion in this chapter is based on analyses undertaken in chapters 5, 6 and 7. The previous literature (see 2.4.2.3) shows that there are two broad groups of modal adverbs of certainty (MACs) in English. The initial analysis determined that there exist equivalent groups of MACs in Urdu (see 4.4.1-4.4.3). One group is *high certainty support* (HCS) and *high certainty support for negative content* (HCSNC); the second comprises *probability support* (PS) and *probability support for negative content* (PSNC) (see 2.4.2.3). In this chapter, I answer my RQ 4: to what degree are English and Urdu MACs similar and in what ways are they used differently?

Although both English and Urdu are part of the Indo-European family, I had at the beginning of my analysis assumed that there would be very marked differences between MACs in English and Urdu. This was due to their different word orders (SVO for English; SOV for Urdu), and the extensive variety and use of modal verbs (MV) in English but not Urdu. My most unanticipated finding is a high level of similarity in the characteristics of MACs in the two languages. However, this finding is consistent with the previous literature (Boye, 2012; 2016), in which it is theorised that any language-specific descriptive category can be compared cross-linguistically by a generalisation of notional meanings (see 2.6). That is, cognitive-conceptual descriptive categories are functionally communicative, and their equivalence, even in geographically and genetically distinct languages, can be determined by their range of semantic, pragmatic, and stylistic factors.

This being the case, in this chapter, I first discuss (in 8.2) the similarities and differences between the preferential distribution of English and Urdu MACs across text types. Then I discuss similarities and differences in the clausal placement of English and Urdu MACs in section 8.3. In section 8.4, I discuss similarities and differences in the semantics of MACs in the two languages especially based on cooccurrence patterns with other elements or sequences. Then, in section 8.5, I discuss similarities and differences in the pragmatic functions of English and Urdu MACs. A summary of the chapter is given in section 8.6.

8.2 The distribution of English and Urdu MACs

In my contrastive analysis of MACs in English and Urdu, I used comparable corpora of several types of newspaper text and online chat. Both English and Urdu HCS and PS MACs occur in both modes of writing (see Tables 4.14 and 4.15). A comparison of the English and Urdu MACs demonstrates some interesting differences in their distribution across text types. However, these differences are not very marked (see 4.4.3). Moreover, as subsequent analyses have shown, the differences in distribution of English and Urdu MACs across these text types apparently do not affect the items' semantic and pragmatic functions, which are more similar than different across the two languages.

If we look only at newspapers versus online chat, relative frequency of both English and Urdu MACs is higher in the former than the latter (see Tables 4.14 and 4.15). Based on this evidence it can be said that in both languages, MACs are equally common in more formal texts (newspaper) than in more casual, possibly speech-like texts (online chat). This similarity in MAC frequency across text types may reflect a similarity in the social contexts represented by those text types in both languages. It may likewise reflect similarity in the

crosslinguistic functional and social-communicative use of MACs as a descriptive category (see 2.6). This finding is consistent with Biber et al.'s (1999, p. 859) analysis that demonstrates differences in frequency of English MACs (labelled *stance adverbials* in Biber et al.'s study) across registers.

8.3 Clause positioning of English and Urdu MACs

The analysis in section 5.6 illustrates a partial similarity in how MACs are positioned within the clause. In both languages, all types of MACs appear both in independent and dependent clauses, and their frequency of occurrence in the two types of clauses is more or less similar (see Tables 5.5 and 5.6). All types of English MAC occur most frequently in clause medial position and least frequently in clause final position. Similarly in Urdu, HCS and PS MACs occur most frequently in clause medial position, while Urdu PSNC MACs are more frequent in clause final position and HCSNC MACs occur most frequently in clause initial position (see the comparison in 5.6 based on Tables 5.5 and 5.6). These differences can be accounted for in terms of a default positioning of the most frequent Urdu PSNC *hō nahīṁ saktā* 'maybe not', which occurs 87 out of 98 times in clause final position, and the most frequent HCSNC MAC *yaqīnān nahīṁ* 'definitely not', which occurs 22 out of 55 times in clause initial position. One of the reasons for *hō nahīṁ saktā* 'maybe not' in clause final position is that in clause final position , they typically occur in verb group and function as an MV instead of a MAC (see 6.8). These MACs are more frequent than the other HCSNC and PSNC MACs in Urdu, and therefore, their preferred positions strongly influence the overall results for their respective categories (see 5.5 and Table 5.6). Moreover, Urdu PSNC and HCSNC MACs have relatively more restricted functions as compared to the equivalent categories in English, which affects their positioning. For instance, in ECC all HCSNC and PSNC MACs occur used to reply to a rhetorical question by the addressers, whereas, in

LUWC, only HCSNC *yaqīnān nahīṁ* 'certainly not' and PSNC *śāyad nahīṁ* 'perhaps not' are used in this way (see 6.5).

However, despite those differences, there is a remarkable similarity between English and Urdu in the usage of MACs at different positions. In initial position, MACs of both languages have scope over the whole proposition (see section 6.2). In both languages, the placement of HCS and HCSNC MACs in initial position always focuses and emphasises the subject or other fronted element of the clause, resulting in an overall intensification of the proposition. On the other hand, English and Urdu PS and PSNC MACs in initial position always focus and downtone the thematic element of the clause, thereby reducing certainty in the truth of the proposition. In both languages, there are instances of clause initial MACs functioning to provide a positive response to a query or request by the previous addresser (see section 6.2).

In clause medial position, MACs in both languages always have in their modal scope the element, typically a main verb, immediately following them in the clause. In that position the modal scope of the MAC is restricted to that verb only. By contrast, English PS and PSNC MACs do not occur immediately before the object of the clause in clause medial position whereas all types of Urdu MACs plus English HCS and HCSNC MACs do. When a MAC occurs before the object in a clause, it has scope over the whole clause. Therefore, unlike other English and Urdu MACs, English PS and PSNC MACs' scope is limited to the immediately following verb and does not extend to the rest of the clause (see 6.3).

Similarly, in clause final position, MACs in both languages have scope over the whole clause they occur in. MACs in clause final position function in both languages as tags, used when an addresser suggests something or invites confirmation from the interlocutor. Hence, in both languages, a MAC in clause final position has a modal scope over that clause.

Thus a clause final MAC adds weight to the proposition by bringing focus of the whole proposition to the end of a clause in both languages (see 6.3). An important difference is that Urdu MACs do not occur in clause final position in interrogative sentences, as English MACs do. This can be attributed to the different word order of the two languages. Otherwise, the propensity of English and Urdu MACs to appear in all three clause positions for equivalent functions, despite the difference in word order, is noteworthy. The tendency of English and Urdu MACs to appear in different clause positions for different functions accords with previous observations (e.g. Boye, 2012; Simon-Vandenbergen & Aijmer, 2007) that, due to their flexibility in clause placement, cross-linguistically MACs demonstrate synchronous polyfunctionality.

The discussion thus far has established that syntactically MACs in both languages tend to occur in same clause initial, medial, and final positions to influence the meaning of the proposition, despite the difference in the word order. The existing literature demonstrates that this ability of MACs to appear in different clause positions in both languages is relatable to their equivalent polyfunctional behaviour with regard to semantic and pragmatic functions (see 2.3 and 2.4). The similarity in characteristics exists because of the similarity in the grammatical category (i.e. MAC) in the two languages.

This finding of more similarities than differences is consistent with previous investigations (Boye, 2012; Croft 2016), which have found that elements in same grammatical category across languages will exhibit, at least to some extent, similarity in the concepts that are considered core characteristics of that grammatical category. The analyses in this thesis support this idea: just as there is an established category in English of modal adverbs that express the addresser's certainty or uncertainty about the information in a proposition, there exists an equivalent grammatical category of Urdu MACs, not hitherto

identified as a specific category of adverbs (see Koul, 2005;Schmidt, 1999), but clearly, on present evidence, having that status.

8.4 Associative meanings of English and Urdu MACs' cooccurrence patterns

The analysis in chapter 6 shows that, semantically, English and Urdu MACs have both similarities and differences in their occurrence in certain sequences that express particular meanings. Most importantly, these sequences involving MACs in both languages are similar in their core functions (i.e. to express certainty or probability). There are also similarities and differences in the meanings expressed by MACs when they occur in different clauses or when they cooccur with various clausal elements. For instance, on one hand, English and Urdu MACs exhibit similarity in meaning when they occur in the *if* and *then* clauses of conditionals. On the other hand, the meanings that English and Urdu MACs express when they occur in interrogatives differ.

In both languages, MACs more readily occur in the apodosis clause of a conditional sentence than in the protasis (see section 6.9). In both languages MACs are rare in *if* clauses. In both languages, in the *then*-clause of a conditional sentence, an HCS MAC confirms the condition given in the *if*-clause. On the other hand, a PS MAC in the *then*-clause expresses low certainty in the truth of the prediction. In sum, English and Urdu MACs have similar functions in epistemic conditional sentences.

Both English and Urdu MACs can cooccur with MVs to enhance or downtone modal meaning. Moreover, MACs cooccurring with MVs exhibit some similarity in usage and meaning (see section 6.6.2). The most remarkable similarity is that in both languages, when an HCS or PS MAC is added to an MV that expresses deontic meaning (e.g. *should*), it

strengthens (or weakens in case it is a PS MAC) the obligation conveyed by the MV (see 6.6).

In English, the meaning of low certainty is expressed when HCS MACs occur together with certain MVs (e.g. *can*, *might*), whereas in Urdu other means are usually employed in lieu of MVs for this function (see Table 4.11). As there is only one epistemic MV, *sak* 'can', these MACs may be used alongside a progressive form (e.g. *yaqīnān* ***khā rahā*** *hō gā* 'certainly will **be eating**') or perfective form (e.g. *yaqīnān* ***kīyā*** *hō gā* 'certainly could have **done**') in order to express low certainty. In fact, in English the use of MACs with MVs is pervasive, whereas this clearly cannot be the case in Urdu, given the relative paucity of MVs overall (see 6.6). Rather than occurring in conjunction with MVs, in Urdu MACs are employed in lieu of MVs. In fact, as existing literature shows that unlike English, preferred use of MACs in lieu of MVs is a more common practice in other Indo-European languages .

Van Olmen and Van der Auwera (2016, p. 369) say that most Indo-European languages, including English, can express epistemic and non-epistemic possibility through verbal constructions and other lexical items that have grammaticalised as modal expressions. But in many Indo-European languages other than English, such as Swedish, German, Norwegian, Persian and Hindi-Urdu, MACs are preferred to MVs to express possibility and certainty (see Aijmer, 1999, Boye, 2016; Van Olmen & Van der Auwera, 2016). The use of MACs instead of MVs in Urdu is, then, consistent with the extensive use of adverbs for this purpose across other Indo-European languages

No study to date has compared the use of different kinds of Urdu items to express modality. However, crosslinguistic comparison of the use of modal adverbs and MVs in English and other languages has been undertaken. Aijmer (1999, pp. 316-17) notes Løken's (1997) comparison of frequencies of MVs and modal adverbs rendered as MVs in translation

from Swedish and Norwegian to English and vice versa. For instance, Løken found that in English translations, 72.2% of Norwegian and 96.5% of Swedish modal expressions were translated as *may* or *might*. On the other hand, English modal expressions being translated as Norwegian or Swedish MVs is much rarer (49.5% and 43% respectively). Aijmer's own analysis confirms the former finding. Similarly, Edmondson et al. (1977, p. 256) point out that in German unlike English, MACs are used more than MVs to express modality, and MVs are particularly associated with formal contexts. Aijmer (1999, p. 319) says that one reason modal adverbs are used to express possibility in Swedish is Swedish's paucity of MVs, compared to English. From these studies, we see that preferring MACs over MVs is a point where Urdu is different from English, but similar to many other Indo-European languages. It is English that shows an idiosyncratic behaviour in its heavy reliance on MVs, rather than Urdu.

Returning to the matter of associative meanings, we see a difference between English and Urdu in the supplementation of epistemic evaluation by one MAC with another MAC. For instance, when an Urdu PS MAC cooccurs with another PS MAC, they indicate tentative inference on the part of the addresser, in both languages. By contrast, in English an HCS MAC cooccurs with another HCS MAC to indicate the addresser's certainty in the inference they have drawn. However, no example of an English PS MAC occurring together with another PS MAC is found in ECC, and no example of an Urdu HCS MAC occurring together with another HCS MAC is found in LUWC. There are examples in both languages of HCS MACs occurring together with PS MACs to express the addresser's low degree of certainty in an inference (see examples in 6.7). Overall, the two languages are similar in terms of the patterns of MACs co-occurring with one another, with a minor difference.

Only English PS MAC *perhaps* occurs in interrogative constructions in my data. In Urdu, an addresser preferentially uses non-adverbial possibility or certainty markers, such as *mūmkīn* 'possible' (adjective), in an interrogative construction.

In both languages, PSNC and HCSNC MACs occur as short responses to queries, whether as means of confirmation or non-committal (see 6.2.4). PSNC and HCSNC MACs also occur as responses to the addresser's own rhetorical questions (see 6.5.3). Despite this similarity, in my comparable corpora, Urdu HCSNC and PSNC MACs are markedly lower in frequency than English HCSNC and PSNC MACs. As aforementioned (see 4.4.1), the parallel data shows that *definitely not* and *must not* get translated with the negators *nahīṁ* 'not' and *hargiz nahīṁ* 'absolutely not' instead of negated MACs. It occurs 457 times in the comparable corpus, far more than *nahīṁ* cooccurring with any HCS MACs (aggregate of Urdu HCS MACs in LUWC is 51). This implies that, in Urdu, use of negators without supplemental MACs projects the same emphatic force (of negation) as it would if MACs were added (see also Genady, 2005 p. 101; Platts, 1784, p.326). Possibly an Urdu MAC cooccurring with a negator would be perceived as redundant, and as not contributing much to the proposition.

8.5 Pragmatic meanings associated with English and Urdu MACs

In both languages, the pragmatic functions of MACs can be categorised as indexical stance (solidarity, authority, politeness, hedging) and rhetorical-pragmatic (emphasiser, expectation, counter-expectation, concession); see sections 7.2-7.9. My analysis demonstrates similarity in pragmatic functions across English and Urdu. For instance, in both languages,

the underlying function of HCS MACs in any proposition is emphasis. HCS and HCSNC MACs enhance the degree of certainty the addresser expresses in the truth value of the proposition; PS and PSNC MACs express a lowered degree of confidence. Both English and Urdu MACs occur on their own or in combination with other MACs (e.g. HCS+PS or PS+HCS) as emphasisers. In addition, in both languages, HCS and PS MACs also occur in similar environments when they occur together with particular elements or sequences to convey a pragmatic meaning. For instance, HCS MACs can also be used to express authority in both languages (7.2). In contrast, the corpus shows that in Urdu a PSNC MAC may also be combined with an HCS MAC to express authority. I did not observe any PSNC + HCS MAC constructions expressing authority in English. No differences like this were observed in the use of MACs in English and Urdu for the functions of solidarity (7.4), expectation (7.5), counter-expectation (7.6), hedging (7.7), and politeness (7.8).

These similarities in use of English and Urdu MACs reflect the fact that, pragmatically, evaluation of meaning is subjective and governed by context-dependent social and functional communicative needs across languages. Therefore, a pragmatic congruence between the English and Urdu MACs signifies that such contextual meanings have become established in both languages. Pragmatic use of MACs is based on conventional inferences of use, drawn by both addresser and addressees. As these inferences are situated in socio-cognitive context, they are generalised and are not specific to any language. However, whether the pragmatic congruence between English and Urdu MACs is also to be observed in social protocols such as (im)politeness or apology is a matter for future research.

8.6 Chapter summary

In this chapter, I have discussed the similarities and the differences between features of English and Urdu MACs based on my analysis in the preceding chapters. Three important outcomes may be noted in summary.

Contrary to expectations, this study did not find any major differences between the usage and functions of English and Urdu MACs. My primary argument has been that, despite the languages' difference in syntactic typology, overall there exist many similarities and few differences. While their word order differs, the two languages demonstrate strikingly similar distribution patterns of MACs (see 8.2), and MACs have similar functions in clause initial, medial and final position (8.3). Consequently, English and Urdu MACs also exhibit remarkable similarities in sequences they tend to occur in, and in their semantic and pragmatic functions (8.4 and 8.5). While they are far from mirror images in their use of MACs, clearly the descriptive categories of MAC in English and Urdu are to a large degree equivalent. This finding broadly supports the theoretical assumptions of Boye's (2012) cross-linguistic study (see 2.3.3.3). These results are in line with a corpus-based contrastive analysis of genetically and geographically similar or different languages (e.g. McEnery & Xiao, 2010) in describing similar concepts in two languages.

Some of the differences between English and Urdu are due to more prevalent use of MVs in English. In fact, there is a strong tendency for English MACs to occur in conjunction with MVs. However, it is encouraging to note that previous researchers found that other Indo-European languages also typically express the concept of epistemic modality through means such as MACs rather than MVs (e.g. Aijmer, 1999; Boye, 2016; Van Olmen & Van der Auwera, 2016). English, therefore, stands unique in its preferred use of MVs, whereas Urdu

resembles other Indo-European languages (e.g. German, Norwegian) in its use of MACs to express epistemic modality.

Other differences, though minor, are not due to the difference in status of MVs. These differences may not be explicable without further research. For instance, I have observed that in Urdu, instead of a MAC, a modal adjective (e.g. *mūmkin* 'possible') is used in interrogative clauses. However, it is not clear why this should be, and specifically why any particular category (here, adjective) is used in lieu of MACs for a certain function. This question is beyond the scope of the present study.

9 Conclusion

9.1 Chapter overview

In this final chapter, I summarise my findings and conclude the thesis. In section 9.2, I lay out the answers to my RQs (see 3.6), based on the contrastive analysis of syntactic placement (Chapter 5), semantics (Chapter 6), pragmatics (Chapter 7), and the implications I have drawn from these corpus-based analyses of modal adverbs of certainty (MACs) in Chapter 8. Then, in section 9.3, I discuss the contribution to the field made by this thesis and its methodological innovations. In section 9.4, I discuss some limitations of the project. Finally, in section 9.5, I discuss possible future research that may follow from what has been accomplished here.

9.2 Summary of findings

RQ 1 was as follows: *On the basis of previous literature and corpus investigation, which lexical items and phrases constitute the set of MACs in English and Urdu?* Based on previous studies including corpus-based investigations (e.g. Biber et al., 1999; Simon-Vandenbergen & Aijmer, 2007), I selected a list of English MACs for this study. I then identified and tabulated these English MACs and translation equivalent Urdu MACs using a parallel corpus (see 4.4.1). The set of English MACs is constituted by *certainly, definitely, obviously, of course, no doubt, undoubtedly, likely, maybe, perhaps, possibly,* and *probably,* and their negative forms *certainly not, definitely not, obviously not, of course not, undoubtedly not, likely not, maybe not, perhaps not, possibly not,* and *probably not*. The set of Urdu MACs is constituted by *beśak* 'undoubtedly', *bilāśubha* 'of course', *koī śubha nahīṁ* 'no doubt', *śak nahīṁ* 'no doubt', *yaqīnān* 'certainly', *yaqīnī tor par* 'certainly', *żarūr*

'definitely', *śāyad* 'perhaps', *hō saktā* 'maybe', and *ġālibān* 'probably', and their negative forms *beśak nahīṁ* 'undoubtedly not', *bilāśubha nahīṁ* 'of course not', *yaqīnān nahīṁ* 'definitely not', *yaqīnī tor par nahīṁ* 'certainly not', *śāyad nahīṁ* 'perhaps not', *hō nahīṁ saktā* 'possibly not', and *ġālibān nahīṁ* 'probably not'. The parallel corpora provide evidence that Urdu MACs cross-linguistically correspond with English MACs, but also that it is common practice to use Urdu MACs to translate English modal verbs (MVs) (see Tables 4.8 and 4.11). My identification of corresponding negative forms of HCS and PS MACs, that is, MAC plus negator (HCSNC and PSNC MACs), in the two languages is supported by evidence from the comparable corpora (see 4.4.2). The congruence in the English and Urdu MACs in parallel and then comparable corpus data demonstrates that MAC as a descriptive category of epistemic modal marker covers most of the meanings of epistemic modality. My consideration of MACs as epistemic modal markers is consistent with Bybee et al.'s (1994, pp. 320-21) definition of epistemic modals as a notion of "degree of commitment of the speaker to the truth or future truth of the proposition", i.e. certainty, uncertainty, possibility and probability. My findings are also consistent with the work within Nuyts' and Boye's approach, under which epistemic support for a proposition can be arranged on a continuum of certainty "which goes from high epistemic support for a proposition over neutral epistemic support to high epistemic support for negative counterpart of a proposition" (Boye, 2016, p. 117); see Figure 2.2.

I established that there exist Urdu MACs that correspond to each of the semantic categories of MAC delineated for English by Boye (2012, p. 46). As mentioned above, Boye's categories classify adverbs according to the degree of confidence in a proposition that they express, and collectively represent a continuum of varying levels of support for a proposition (see 2.3.3.3). I have labelled these categories *high certainty support* or *high certainty support for negative content* (HCS and HCSNC respectively), and *probability*

support and *probability support for negative content* (PS and PSNC respectively), following Boye (2012, p. 46). The results reported above (see 4.4) incrementally support Boye's (2012, p. 46) hypothesis that the use of lexical items as justificatory support for a proposition is a universal phenomenon. Boye (2012, p. 2) defines *justificatory support* as use of lexical items which express an addresser's degree of "epistemic support" for a proposition. Boye (2012, p. 124) lists *certainly, possibly, probably, obviously,* and *likely* as English adverbs whose meaning can be described in terms of justificatory support.

I found only one example of an HCSNC MAC in the English-Urdu parallel corpus (see Tables 4.8, 4.9 and 4.10) and none of any PSNC MACs. But in the comparable corpora, I found instances of HCSNC and PSNC MACs in both English and Urdu. However, HCSNC and PSNC MACs appear to be considerably rarer in Urdu than English (see Tables 4.12 and 4.13). A plausible explanation for this phenomenon is that in Urdu, there is a tendency to use negators alone where in English, a negated MAC would be found. This tendency can be observed in the parallel corpus, in which *definitely not* and *must not* are translated with the negators *nahīṁ* 'not' and *hargiz nahīṁ* 'absolutely not' instead of negated MACs (see 8.4). The marked difference in the frequency of the two aforementioned negators and HCSNC and PSNC MACs in the English and Urdu comparable corpora underlines this tendency of Urdu negators to appear on their own to express negative epistemic commitment (see 4.4.2). Yet it should be underlined that, in spite of such differences of detail, the corpus evidence from parallel and comparable corpora does support the existence and strong cross-linguistic similarity of the category of MAC in English and Urdu.

RQ 2 was as follows: *Is the placement of (English and Urdu) MACs in different clauses similar to what previous literature shows?* The analysis in chapter 5 shows that the placement of both English and Urdu MACs in different independent and dependent clauses is similar to what previous studies have claimed regarding English MACs. The distribution of

MACs in my comparable corpora shows that they can occur in initial, medial and final clause positions, and in various types of clauses, as described in the literature on English MACs (Biber et al., 1999; Simon-Vandenbergen & Aijmer, 2007; Boye, 2012). Being able to appear in clause initial, medial, and final positions is a feature that MACs share with other adverbs, in both languages. However, for all MACs, the medial position is the most typical (see 5.6). Urdu constructions that express certainty or possibility typically involve a MAC instead of an MV; this is attributable to the paucity of distinct Urdu MVs (two) compared to English (nine). In fact, in many cases English MACs cooccur with MVs to emphasise or de-emphasise certainty (see sections 6.6 and 8.4). Despite the difference in the languages' basic word orders, in both English and Urdu, when MACs occur together with a verb, they typically precede it. These examples of MACs before a verb are consistent with Simon-Vandenbergen and Aijmer's (2007, p.86) claim that in English, the verb is the "hub of the proposition" and an HCS MAC preceding it in clause medial position "is likely to emphasise the truth of the proposition as a whole"; my analysis shows that the same holds true for Urdu (see 6.3). By contrast, my analysis also establishes that, in both languages, a PS MAC in clause medial position preceding a verb expresses support for the truth of the proposition but does not function as an emphasiser (see 6.3).

RQ 3 has three parts. The overarching question is: *In what ways do the (English and Urdu) MACs in different clausal positions influence the surrounding elements*? The first part of the question is: *How does the placement of MACs affect their semantic scope over other clause elements, clauses or sentences?* The analysis in chapter 6 establishes that the occurrence of English and Urdu MACs in three clause positions, and the modal scope that they possess in consequence of their clause position, is consistent with previous research on English MACs (Boye, 2012; Simon-Vandenbergen & Aijmer, 2007). The analysis illustrates that in both languages, MACs in initial, medial, and final clause positions can have modal

scope over the whole clause (see 6.2 to 6.5). However, in some positions the scope can be narrower.

The next part of RQ 3 is: *What are the modal semantics of MACs at clause level?* My analysis produced results consistent with previous research on the semantic functions of English MACs (e.g. Pic & Furmaniak, 2012; Simon-Vandenbergen & Aijmer, 2007) (see 6.2 to 6.9). When MACs occur in *then*-clauses of epistemic conditional sentences, they express an addresser's degree of certainty in the predicted state of affairs (see 6.9). The interaction of English and Urdu HCS MACs with MVs with primarily deontic meaning, such as *should, must* and *cāhīē* 'should', strengthens the deontic reading of those verbs, but the HCS MACs themselves do not act as deontic markers (see 6.6; see also Simon-Vandenbergen & Aijmer, 2007). Similarly, when PS MACs occur with MVs such as *should, must* or *cāhīē* 'should', they only lessen the strength of the deontic reading functioning as possibility markers (see 6.6.2). Therefore, MACs as modal expressions primarily function as emphasisers (Quirk et al., 1985, p. 583) or downtoners (Quirk et al., 1985, p. 597): they primarily either add emphasis to or remove emphasis from the degree of certainty, in doing so conveying modal evaluation by an addresser; see section 8.4. MACs also have additional specific functions in particular clause positions. For instance, both English and Urdu MACs at times in clause final position function as a tag question to prompt a response from the interlocutor (see 6.4.3). Moreover, a MAC may also shift the core function of a clause. For instance, in both languages, the use of a MAC in an interrogative clause may result not in emphasis of the interrogative sense, but rather in the overall function ceasing to be a question and becoming a suggestion or hedge (see 7.7).

The third part of RQ 3 is as follows: *What pragmatic functions do MACs perform in English and Urdu?* My pragmatic analysis supports earlier studies of pragmatic features of English MACs, which have described their functions in terms of indexical stance and

rhetorical-pragmatic functions (see 2.4.2, 7.2 to 7.8 7.3). My analysis demonstrates the pragmatic functions of English and Urdu MACs to be similar (see 7.9 and 8.5). The similarity in pragmatic functions across the two languages may possibly be explained by the fact that social-cognitive communication is bound by social and situational contexts which transcend the boundaries of a specific language (see discussion in 8.6). A possible explanation of the pragmatic similarity of English and Urdu MACs is that, as they are used in similar situational contexts, the addresser's intended meaning can be interpreted by the addressee in those social contexts in a manner not specific to any given language. To perform these pragmatic functions involves lexical items (i.e. MACs) with the same semantics (i.e. epistemic (un)certainty). Consequentially, MACs having similar pragmatic functions in English and Urdu is a result of them being used in similar situational contexts. The analysis in chapter 7 shows that HCS MACs are used for pragmatic functions where the addresser wants to assert certainty in the truth of the proposition, such as solidarity, expectation, and authority. In contrast, given the contexts, addressers use both English and Urdu PS MACs pragmatically to convey the modal value of probability instead of high certainty for a proposition, whether to mitigate criticism or to save their own or the addressee's face.

RQ 4 is as follows: *To what extent are the behaviours of the Urdu and English MACs (as established by RQ 2 and 3) similar, and in what ways are they distinct?* The discussion in chapter 8 expanded on the issue of similarities and differences between English and Urdu MACs, which had been covered already to some extent (see RQ 3 answers). The discussion in chapter 8, based on these aforementioned results, shows that MACs in the two languages overall exhibit many similarities and few differences in their characteristics and functions. This is especially surprising because of the structural differences between the two languages, most notably in basic word order. In spite of word order, however, the distribution patterns of MACs in the two languages are similar; consequently, the observed semantic and pragmatic

functions are likewise strikingly similar. The two major differences are: pervasive use of English MACs with MVs, a tendency not replicated in Urdu because it has fewer MVs (as discussed above); and a marked difference in frequency of occurrence between English and Urdu HCSNC and PSNC MACs. However, these distinctions apparently do not lead to any notable difference in the tendency of MACs to appear with particular lexical items or in certain types of clauses. Overall, the similarities in the functions of MACs in English and Urdu strongly imply that they encode a common semantic group of certainty while not being exact mirror images. That is, the functions for which MACs are utilised in English and Urdu are similar, but the MACs making up a given English-Urdu translation equivalence pair may not have precisely the same functional coverage (see Tables 4.8 and 4.11).

9.3 Contribution to the field made by this study

This thesis contributes to the growing literature involving corpus-based study of Urdu. Moreover, the analyses that I have presented bear out the efficacy of the approach of corpus-based contrastive descriptive analysis. This project utilised a combinatory approach to contrastively describe the functions of MACs in English and Urdu, being informed by both prior research and corpus data (see 2.3 to 2.6, and 4.3). The literature consulted was primarily on English MACs, because to date there is not any substantial literature on the descriptive category of MACs in Urdu. On the corpus side, both parallel and comparable corpora were used, following McEnery & Xiao (2010): parallel for the initial identification of the Urdu MACs corresponding to different English MACs; comparable for locating examples to support the analysis of similarities and differences in MACs across the two languages. This contrastive analysis of an Urdu descriptive category, using a comparable corpus, is the first of its kind. Prior to this thesis, no single such descriptive study of a category of Urdu lexical items has incorporated corpus-based analysis. Thus, the first contribution of this study is to

the field of descriptive analysis, and consists of its setting a precedent for future corpus-based descriptive studies of Urdu.

This thesis's contribution is that, as a contrastive study of English and Urdu MACs as a grammatical category, it extends the knowledge of that area achieved by previous research on English MACs (e.g. Boye, 2012; Simon-Vandenbergen & Aijmer, 2007; Van der Auwera et al., 2005). The present study also goes beyond most prior studies of English modal adverbs by putting modal adverbs of certainty and possibility into a single category. Moreover, it has innovated by conjoining corpus-based contrastive analysis (following Simon-Vandenbergen & Aijmer, 2007) and the social-cognitive communication approach (following Boye, 2012). By combining two different theoretical frameworks, the thesis has provided a deeper insight into the equivalent meanings and pragmatic functions of MACs in the two languages than would otherwise have been possible.

Third, this thesis has showcased the ways in which a contrastive analysis using corpus data can help arrive at a better description than is possible using intuition and made-up examples only. On multiple occasions, I personally found, upon looking at an English example, that my instant first reaction was that the construction present in the example existed, and was common, in Urdu too, only to later find that the data showed otherwise. An obvious example of this is the negative forms of Urdu MACs, that is HCSNC and PSNC MACs. The corpus-based analysis establishes that, contrary to my intuition, the frequency of Urdu HCSNC and PSNC MACs is far less than the frequency of the same categories of MAC in English (see Tables 4.12 and 4.13). Interestingly, the analysis also demonstrates that this difference in frequency does not affect the similarity of these items' semantic and pragmatic functions across English and Urdu (see 6.2-6.9, 7.2-7.8, 8.4 and 8.5).

Fourth, my findings have implications for translation studies in that its findings can be a source of information on MACs for translators. For translators, it is easier to translate elements that have equivalent, if not identical, meaning in the original and the translated language. However, translators – or those who create the bilingual lexical resources used by translators – need to avoid dependence on introspection or intuition in identifying translation equivalent lexical items (Lambert & Van Gorp, 2006, p.45). Rather, they should refer to the "vastly superior" evidence provided by the collocations and typical uses of lexical items observable in corpus data (Sinclair, 1991, p. 42). The results of this study can help translators make an informed decision about how a certain example of a MAC should be translated according to its semantic and pragmatic function (see 8.5).

The use of parallel corpora helped in establishing the most common English MACs and translation equivalent Urdu MACs. Comparable corpora established that there exists syntactic congruence between the English and Urdu items. As already noted, this similarity is evident in the occurrence of English and Urdu MACs in initial, medial, and final clausal positions and in different types of both independent and dependent clauses. Given the analysis in chapters 5 and 6, this similarity makes good sense, because similar or equivalent syntactic positioning of English and Urdu MACs correspond to similar semantics. The analysis also provides evidence that translation-equivalent MACs, or sets thereof, should not be treated as mirror images because their behaviour may or may not be precisely similar in the two languages (see McEnery & Xiao, 2010). The interchangeability in context of semantically similar MACs is not identical across the two languages. For instance, of the Urdu HCSNC MACs, only *yaqīnān nahīṁ* 'certainly not' is used as a reply to a rhetorical question, whereas all English HCSNC MACs can have this function (see 6.5.3 and 8.4). The higher frequency of *yaqīnān nahīṁ* 'certainly not' in clause initial position relative to other Urdu HCSNC MACs in clause initial position is one consequence of this lack of

interchangeability of MACs for certain functions. This finding exemplifies how corpus data can help in judging which MAC is the best translation in a particular context with a particular function.

In sum, the contributions made by this study are as follows. A corpus-based approach provides an insight into the correlation between syntactic patterning and semantic meaning. And it demonstrates applicability of combining approaches, namely, corpus-based contrastive analysis and the social-cognitive communicative approach to the pragmatic functions of English and Urdu MACs. My study also provides a source for understanding uses and functions of MACs by the translators. Therefore, from both quantitative and qualitative perspectives, it makes a distinctive and innovative contribution to the knowledge of MACs available to descriptive grammarians and translators.

Moreover, the compilation of an indexed and part-of-speech tagged monolingual Urdu corpus (LUWC), and its being made accessible to the wider group of researchers in linguistics, is a first-of-its-kind accomplishment. Earlier Urdu corpora do of course exist, but they are either not available free of cost to researchers (e.g. CLE Urdu digests corpus[29]) or the corpus compilers have split and scrambled the corpus examples at clause level (e.g. Charles University Urdu Monolingual Corpus, see 4.2.2), which discourages researchers to use such corpora.

9.4 Limitations of the study

As I have argued, the present corpus-based contrastive study has arrived at invaluable findings regarding features of MACs in English and Urdu. However, due to time and space constraints among other practical issues, certain limitations proved unavoidable. For instance,

[29] https://www.cle.org.pk/clestore/index.htm

only the initially selected set of English MACs and corresponding Urdu MACs were analysed. I had designed the study so as to include only those Urdu and English MACs which occur in the English-Urdu parallel corpora, without adding any others by relying on speaker intuition or dictionaries. Simon-Vandenbergen and Aijmer (2007, p. 232) arrive at a list of MACs to analyse by conducting an elicitation experiment in which native speakers were given example sentences with gaps, into which they inserted a MAC of their choice for that context. While I contemplated such a step, I ultimately opted not to incorporate speaker intuitions as an additional factor in the study because once I compiled a considerable sized comparable data, I found that I had corpus-informed evidence of the use of MACs in all possible contexts that I wanted to observe in my thesis.

Therefore, the list of MACs studied did not include those MACs which were not found in the parallel corpus, of which some definitely do exist, for example *in fact*, *surely* and *daraṣ̱al*/*darḥaqīqat* 'in fact'. It would have been possible to amend the study design to include items such as these suggested by my own intuition, or additional MACs found in my comparable corpora. However, I instead adhered to my initial decision to consult only the English-Urdu parallel corpus data for the choice of MACs. The reason for that decision is two-fold: i) the current list of MACs forms the core semantic group of modal adverbs expressing epistemic certainty or uncertainty in each language; ii) the MACs identified later which were not included are used mainly in formal and literary texts. I shall elaborate on each point.

The selected set of English MACs are the core elements in the semantic group of epistemic certainty or uncertainty (see 1.5.1). That is, the selected MACs are the ones that have been previously examined as a part of semantic group of either certainty or possibility in English (see Van der Auwera & Plungian, 1999; Boye, 2012; Pic & Furmaniak, 2012; Simon-Vandenbergen & Aijmer, 2007). By inference, the same is likely true of the Urdu

MACs derived by identifying translated equivalents. Admittedly, I cannot know for certain whether, and to what extent, the overall picture would be different if the unaccounted-for MACs were also part of the analysis.

The government information leaflets in the English-Urdu parallel corpus have been written and translated with much attention to clarity and comprehensibility. They avoid literary and formal style, which in Urdu includes register-specific MACs, such as *vāżeh tōr par* 'avowedly', *bilāḥujjat* 'unquestionably', or *bilāśubha* 'undeniably', as later observed in comparable corpus. Using a list of MACs derived from the parallel corpus thus meant focusing on a core group of MACs in regular and widespread use, rather than those restricted to particular, specialised genres – a decision suitable to the variety of materials in the comparable corpora. Therefore, I did not deem it necessary to change my study design to include such formal and/or literary MACs.

Moving on, the possibility also exists that my analysis has been limited by constraints on the amount of available data. To put it another way, it is plausible, albeit unprovable, that more informative results would have been obtained if there were more, and more extensive, parallel and comparable corpora of English and Urdu available to begin with.

Especially the available parallel corpus is limited in size (see Table 4.1) and arguably non-optimal in composition (see Table 4.2). A possible solution would have been to compile a supplementary parallel corpus, much as I compiled novel comparable corpora. But given the scarcity of accessible machine-readable Urdu text of the relevant kinds, that plan of action could not be contemplated within the scope of this thesis project. Other parallel corpora of English and Urdu do exist, for instance in the form of translations of the Bible and Quran. However, an initial analysis of this data showed that many of the modal adverbs found in the Urdu translation of the Quran exhibit features suggesting they are Arabic loans. The strong

Arabic influence renders this corpus less useful for the study of the category of Urdu MACs (see 4.2.1). Similarly, the status of these texts in both English and Urdu as parallel translations from Hebrew/Greek/Arabic, rather than the source/target language pairs, and as sacred texts subject to the cultural constraints and expectations surrounding that highly specialised genre, limits their utility to a study like this one. Thus, they offer no solution to the problem of limited parallel data.

Similarly, it may be objected that the comparable corpora which I developed, though of considerable size, and readily comparable in terms of their domains and dates of production (see Tables 4.1, 4.2 and 4.3), are restricted in terms of the genres and media that they represent. This criticism is not invalid, but this limitation was a matter of what was practically possible. For Urdu, only certain types of text were easily available as machine-readable data. Typically, Urdu fiction and non-fiction publications exist electronically only as image data in PDF files. Moreover, the present state of the art in optical character recognition is insufficiently advanced for accurate conversion of such images into text. To represent another medium within the data, ideally spoken text would have been included. However, building any spoken (sub-)corpus of appreciable extent is a major undertaking, one which was simply not feasible within the scope of a PhD project not wholly dedicated to corpus design and development. To mitigate this limitation, I deliberately included online chat fora into the design of my comparable corpora; online text chat may be considered a hybrid of spoken and written language (see Chafe & Danielewicz, 1987; Freiermuth, 2011). However, while the social-functional context of online chat is similar to that of informal conversation, the degree of spontaneity is lower. Moreover, features characteristic of spoken language (e.g. turn-taking) may not behave in online conversation precisely as they do in real speech. The mitigation is therefore only partial.

Egbert et al. (2022, p. 160) observe that the objective of using a representative corpus is to get precise "estimates of quantitative-linguistic parameters in the real world-domain". Therefore, they argue, to find accurate estimates, corpus linguists must assess the accuracy of their corpus-based findings as compared to the actual values in the real world. I believe that despite certain limitations of my comparable corpora as mentioned above, my present findings (quantitative and qualitative) do meet the parameters set out in Table 2.2, and the findings can be generalised for further corpus-based descriptive analysis of linguistic features of Urdu using LUWC.

Some aspects of the research were also, to a certain degree, subjective. One example of this is partial subjectivity in my semantic analysis. In the semantic analysis, examples of English MACs being used for response to a rhetorical question raised by the addresser in the previous sentence were easier to locate than some other uses of MACs. For instance, my own introspective judgement was that Urdu MACs can be used in epistemic conditional sentences. But finding examples of Urdu MACs used this way was much more difficult than use of Urdu MACs in other types of conditional constructions (e.g. speech act conditionals). It was necessary to inspect many concordance lines, and their wider context beyond a single clause, to find sure examples of this phenomenon. I had to consider and discard many examples which, after deliberation, seemed not to exemplify this function (see 6.9). The high degree of interpretative activity required in identifying relevant Urdu examples constitutes a subjective element in the analysis.

Similarly, my judgement of the meaning conveyed by English and Urdu MACs collocates in certain clause positions may be considered subjective. For instance, my analysis of MACs used in negative responses (see 6.2.3) included example (197).

197. A: "Just because she's polish doesn't mean she's allowed to touch her"

B: "*Of course* not" (ECC_AskUK201810).

I interpret this example as follows: the response is negative but pragmatically it shows that the replier is agreeing with the original addresser and expressing solidarity with them. This is my subjective understanding. However, a reader might well contest this reading and judge the MAC to have a different function in this context.

Another instance of subjectivity arose in the pragmatic analysis. Pragmatic analysis inherently involves interpretation more abstract than consideration of the meaning of a single linguistic item. At times, I needed to look *beyond* the immediate co-text in order to understand the pragmatic implications of a MAC being used. For instance, interpretation of *perhaps* in example (181), repeated here as (198), required a much wider reading. I found it difficult to understand this MAC's function as a hedge, in a passage which simultaneously defends use of racist language by teachers and distances the writer from that attitude, until I read the whole articles.

198. "Put simply, this means there are now laws in this country preventing teachers from saying certain things- not incitement to violence or stupid rabble-rousing bigotry, just controversial and *perhaps mistaken things* that the dominant elite in our society have decided are offensive" (ECC_mail9021775).

Still, this interpretation is fundamentally *my* interpretation, and no other reader will bring my precise set of attitudes, experiences, and understanding of the text's social context to bear. The point is that any semantic and pragmatic analysis requires interpretation of the examples at hand by the researcher's own linguistic faculty – a process which is inherently subjective but unavoidably so. Leaving out subjectivity would mean leaving out most of the semantic and pragmatic analysis. But subjective does not mean arbitrary. The examples that I have given across various analyses support my arguments with evidence of similarities and

differences in the semantic and pragmatic features of MACs in the two languages. Moreover, this limitation is mitigated by situating such analyses within a context of quantitative findings, especially relating to MACs' clause positioning, and then categorising and comparing the semantic and pragmatic features observed in the two languages.

9.5 Future research possibilities

Given the numerous angles that MACs can be investigated from, it is not surprising that my study is far from comprehensive. In this final section of the thesis, I will briefly outline some possible topics for future research within the area of descriptive and contrastive analysis of Urdu MACs which, if followed up, will lead to a more comprehensive picture emerging. I emphasise in particular the range of different possible avenues of research.

One possible course of action would be to address the limitations discussed previously by improving the representativeness of medium and genre in the comparable corpus data. As I have noted, despite the substantial size of the comparable corpora, they lack diversity in genre and medium by not including spoken data and literary, technical and academic writing. Alternatively, a subsequent analysis might be performed using only spoken corpus data, if it were possible to compile the necessary corpus. Spoken data, representative of spontaneous conversation, might produce different results to those I have presented here. And of course, grammatical choices are known to vary across genres (see Biber et al., 1999; Hasselgård, 2019) as well as between speech and writing.

However, improvements to the parallel corpus component of the data are likely to be contingent on advances in optical character recognition. A plethora of literature in translation, both from English to Urdu and from Urdu to English, exists in the image-based form

discussed previously (see 4.2.4.1). A study that includes a more diverse parallel corpus will possibly shed a light on the usage of those MACs that were not included in this thesis.

Another possible direction of research beyond the present study would be to investigate the use of other epistemic modal markers, such as modal adjectives or nouns. For English, this will also involve a detailed investigation of whether the interaction between MACs and MVs extends to modal markers other than MACs. For Urdu, there is a need for quantitative evidence on whether MACs are the most typical markers of certainty, as opposed to modal adjectives (e.g. *mūmkīn* 'possible') or other categories.

One particularly interesting topic of this kind would be an investigation into which Urdu lexical modal markers are the preferred choice of Urdu speakers to translate English MVs (whether in spoken or written medium). Such preferences may be revealed via a cross-categorial quantitative analysis of frequencies of various types of modal marker. But any such analysis would ideally move on from just frequency to examine associations between epistemic modal markers and semantic and pragmatic functions – as, indeed, the present study has done.

Another valuable contribution would be a detailed investigation of occurrence of MACs in independent and dependent clauses. My analysis here includes the frequency of occurrence of MACs in independent and dependent clauses, statistics based on a huge number of examples. However, it was beyond the scope of the present study to pursue a full comparative discussion of the meanings and pragmatic functions of MACs in different clause types. Yet we might well expect these functions to be affected by whether the MAC is in a dependent or independent clause. A future study would help explain why MACs are preferred in certain grammatical-textual contexts but rarely used in others.

Finally, another possible avenue for future research would be to examine the occurrence of certain MACs in sequences where their meaning is distinct from that which they convey in other contexts (see 6.6). The present study has identified two such Urdu MACs, *hō saktā* 'may be' and *śak nahīṁ* 'no doubt'. But a comprehensive investigation of this phenomenon would have been beyond the scope of the present study. It seems likely that addressing this question in terms of *semantic sequences*, which can be identified by examining the co-text of examples of a specific lexical item (see Hunston, 2008), would be fruitful. On the same subject of collocational or idiomatic combinations, there is also room for a thorough investigation of use of *sak* 'can' with general lexical verbs versus existential *hō* 'be' (see also Genady, 2005; Schmidt, 1999). *Sak* 'can' is normally an ability marker, that is, dynamic modality (see 4.4.1). But the use of *hō* 'be' before *sak* has been grammaticalised so that the phrase functions as a modal adverb expressing epistemic modality. On the other hand, when *sak* follows general lexical verbs (e.g. *kar* 'do'), it always functions as ability marker. Moreover, in my analysis, I did not look for sequences of *saknā* 'can' with auxiliaries (e.g. hai 'is'). For instance in a construction ending with past auxiliary *thā* 'was' *hō saktā* expresses counterfactual meaning. More broadly, an investigation into the patterns of occurrence of *saknā* 'can' with all kinds of verbs will be valuable to determine what meanings are generated through these extended lexical sequences.

Notwithstanding the limitations discussed in the previous section and the points of inquiry that must be left to future research, it is hoped that the present study has provided a good starting point for not only future studies on Urdu MACs specifically but also for corpus-based contrastive analysis of the nature and functions of corresponding descriptive categories in English and Urdu in general.

References

Adams, H., & Quintana-Toledo, E. (2013). Adverbial stance marking in the introduction and conclusion sections of legal research articles. *Revista de lingüística y lenguas aplicadas, 8*(1), 13-22.

Ahmad, R. (2008). Scripting a new identity: The battle for Devanagari in nineteenth century India. *Journal of Pragmatics, 40*(7), 1163-1183.

Aijmer, K. (2008). Parallel and comparable corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 275-291). Berlin: Walter de Gruyter.

Aijmer, K. (2013). *Understanding pragmatic markers.* Edinburgh: Edinburgh University Press.

Aijmer, K. (2015). Analysing discourse markers in spoken corpora: *Actually* as a case study. In T. McEnery & P. Baker (Eds.), *Corpora and discourse studies* (pp. 88-109). London: Palgrave Macmillan.

Aijmer, K., & Altenberg, B. (Eds.). (2013). *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson.* Philadelphia, PA: John Benjamins.

Aijmer, K., & Lewis, D. (Eds.). (2017). *Contrastive analysis of discourse-pragmatic aspects of linguistic genres.* New York, NY: Springer.

Aijmer, K., & Rühlemann, C. (Eds.). (2015). *Corpus pragmatics.* Cambridge: Cambridge University Press.

Aikhenvald, A. Y. (2004). *Evidentiality.* Oxford: Oxford University Press.

Akatsuka, N. (1985). Conditionals and the epistemic scale. *Language, 61*(3), 625-639.

Akram, M. (2008). Speech Acts: A contrastive study of speech acts in Urdu and English. *Asian EFL Journal, 10*(4), 148-172.

Algeo, J. (1987). Review of *A Comprehensive Grammar of the English Language*. By Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. London: Longman. 1985. x+ 1779. *Journal of English Linguistics, 20*(1), 122-136.

Ali, S. S. (2015). Minority language speakers' journey from the mother tongue to the other tongue: A case study. *Kashmir Journal of Language Research, 18*(3), 65-85.

Andersen, G. (2001). *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*. Amsterdam: John Benjamins.

Arnold, E. J., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs newness: the effects of structural complexity and discourse status on constituent ordering. *Language, 76*(1), 28-55.

Badan, L. (2020). Italian discourse markers: The case of *Guarda te*. *Studia Linguistica, 74*(2), 303-336.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233-250). Amsterdam: John Benjamins.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics, 14*(3), 312-337.

Baker, P. (2020). Corpus-assisted discourse analysis. In C. Hart (Ed.), *Researching discourse* (pp. 124-142). London: Routledge.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, P., Hardie, A., McEnery, T., Cunningham, H., & Gaizauskas, R. (2002). EMILLE: a 67-million word corpus of Indic languages: data collection, mark-up and harmonization. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (pp. 819-827). Las Palmas, Canary Islands: European Language Resources Association (ELRA).

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays* (Vol. 1). University of Texas Press.

Barlow, M. (2008). Parallel texts and corpus-based contrastive analysis. In M. Gonzalez, J. Mackenzie, & E. Alvarez (Eds.), *Current trends in contrastive linguistics: Functional and cognitive perspectives* (pp. 101-121). Amsterdam: John Benjamins.

Beg, M. K. (1988). *Urdu grammar: history and structure*. New Delhi: Bahri Publication.

Bhatt, R., Bögel, T., Butt, M., Hautli, A., Sulger, S., & King, T. H. (2011). Urdu/Hindi modals. *Proceedings of the LFG11 Conference* (pp. 47-67). Stanford: CSLI Publications.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic purposes, 5*(2), 97-116.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in contrast, 14*(1), 7-34.

Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Biber, D., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortez, V., Csomay, E., Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. *TOEFL Monograph MS-25*. Princeton, NJ: Educational Testing Service.

Bilal, H. A., Tariq, A. R., Yaqub, S., & Kanwal, S. (2013). Contrastive analysis of prepositional errors. *Academic Research International, 4*(5), 562-570.

Boye, K. (2012). *Epistemic meaning: A crosslinguistic and functional-cognitive study*. Berlin: Walter de Gruyter.

Boye, K. (2016). The expression of epistemic modality. In J. Nuyts, & J. van der Auwera (Eds.), *The Oxford handbook of modality and mood* (pp. 117-140). Oxford: Oxford University Press.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Butt, M. (1995). *The structure of complex predicates in Urdu.* Stanford,CA: Center for the Study of Language (CSLI).

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied linguistics, 16*(2), 141-158.

Caton, C. E. (1966). On the general structure of the epistemic qualification of things said in English. *Foundations of language, 2*(1), 37-66.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe, & N. Johanna (Eds.), *Evidentiality: The linguistic coding of epistemology* (pp. 261-272). Norwood, NJ: Ablex Publishing.

Chafe, W., & Danielewicz, J. (1987). *Properties of spoken and written language. Technical report No. 5.* California University center for the study of writing. Berkeley: Academic Press.

Chesterman, A. (1998). *Contrastive functional analysis.* Amsterdam: John Benjamins.

Coates, J. (1983). *The semantics of the modal auxiliaries.* Surry Hills, Australia: Croom Helm.

Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language* (3rd ed.). New York, NY: Routledge.

Comrie, B., Haspelmath, M., & Bickel, B. (2008). *The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.* Max Planck Institute for Evolutionary Anthropology and University of Leipzig. Retrieved from https://www. eva. mpg. de/lingua/pdf/Glossing-Rules. pdf

Conrad, S. (2010). What can a corpus tell us about grammar? In A. O'Keeffe, M. McCarthy, A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 227-240). New York, NY: Routledge.

Conrad, S., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes*, *12*(1), 33-43.

Coussé, E., & Van der Auwera, J. (2012). Human impersonal pronouns in Swedish and Dutch: A contrastive study of *man* and *men*. *Languages in contrast, 12*(2), 121-138.

Croft, W. (2016). Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology, 20*(2), 377-393.

Culpeper, J. (2011). Politeness and impoliteness. In G. Andersen, & K. Aijmer (Eds.), *Pragmatics of society* (pp. 393-438). Berlin: Walter de Gruyter.

Dancygier, B., & Sweetser, E. (2005). *Mental spaces in grammar: Conditional constructions.* Cambridge: Cambridge University Press.

De Cock, S., & Goossens, D. (2013). Quantity approximation in English and French business news reporting: More or less the same? In K. Aijmer, & B. Altenberg (Eds.), *Advances in corpus-based contrastive linguistics: Studies in honour of Stig Johansson* (pp. 139-156). Philadelphia, PA: John Benjamins.

De Sutter, G., & Lefer, M. A. (2020). On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach. *Perspectives, 28*(1), 1-23.

Diani, G. (2015). Introductory 'IT' patterns in English and Italian academic writing: A cross-generic and cross-cultural analysis. *L'Analisi Linguistica e Letteraria 2008-1, 16*, 343-355.

Do Nascimento, M., Goncalves, J., Pereira, L., & Estrela, A. (2006). The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-Derived Lexicon. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1791-1794). Genoa, Italy: European Language Resources Association (ELRA).

Ebeling, J. (2000). *Presentative constructions in English and Norwegian.* Unpublished doctoral thesis, University of Oslo.

Ebeling, J., & Ebeling, S. O. (2013). *Patterns in contrast.* Amsterdam: John Benjamins.

Ebeling, S., & Ebeling, J. (2020). Contrastive analysis, tertium comparationis and corpora. *Nordic Journal of English Studies, 19,* 97-117.

Edmondson, W., House, J., Kasper, G., & McKeown, J. (1977). *A pedagogic grammar of the English verb: A handbook for the German secondary teacher of English.* Amsterdam: John Benjamins.

Egan, T., & Dirdal, H. (Eds.). (2017). *Cross-linguistic correspondences: From lexis to genre.* Amsterdam: John Benjamins.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness.* Cambridge: Cambridge University Press.

Engel, D. M. (1999). Radio Talk: French and English perfects on air. *Languages in contrast, 2*(2), 255-277.

Farooqi, M. A. (2008). The 'Hindi' of the 'Urdu'. *Economic and Political Weekly, 43*(9), 18-20.

Flowerdew, J., & Scollon, R. (1997). Public discourse in Hong Kong and the change of sovereignty. *Journal of Pragmatics, 28*(4), 417-426.

Foley, W. A. (1986). *The Papuan languages of New Guinea.* Cambridge: Cambridge University Press.

Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics, 14*(2), 219-236.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic and philosophical perspectives* (pp. 159-175). London: Associated University Press.

Freiermuth, M. R. (2011). Debating in an online world: A comparative analysis of speaking, writing, and online chat. *Text & Talk, 31*(2), 127-151.

Fries, C. C. (1957). *The structure of English.* London: Longmans, Green .

Gales, T. A. (2010). *Ideologies of violence: A corpus and discourse analytical approach to stance in threatening conversations.* Unpublished doctoral dissertation, University of California.

Gast, V. (2012). Contrastive linguistics: Theories and methods. In *Dictionaries of Linguistics and Communication Science: Linguistic theory and methodology* (pp. 1-10). Berlin: Mouton de Gruyter.

Gast, V. (2015). Contrastive linguistics. In M. Byram, & A. Hu (Eds.), *Routledge Encyclopedia of language teaching and learning* (2nd ed., pp. 153-158). New York: Routledge.

Genady, S. (2005). *Modality in Hindi*. Munich: LINCOM Europa.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). New York, NY: Routledge.

Goffman, E. (1967). *Interaction ritual*. Chicago: Aldine Publishing.

Gonzalez, M., Mackenzie, J., & Alvarez, E. (2008). Introduction. In M. Gonzalez, J. Mackenzie, & E. Alvarez (Eds.), *Current trends in contrastive linguistics* (pp. xv-xxi). Amsterdam: John Benjamins.

Granger, S. (2008). Learner corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 259-274). Berlin: Walter de Gruyter.

Greenbaum, S. (1969). *Studies in English adverbials*. London: Longman.

Greenbaum, S., & Svartvik, J. (1990). *The London-Lund corpus of spoken English*. Lund: Lund University Press.

Halliday, M. (1970). Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of language, 6*(3), 322-361.

Halliday, M. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 30-43). New York, NY: Routledge.

Halliday, M. (1994). *An introduction to functional grammar* (2nd ed.). London: Arnold.

Halliday, M., & Matthiessen, C. (2004). *An introduction to functional linguistics*. London: Edward Arnold.

Harris, A. C. (2006). Revisiting anaphoric islands. *Language, 82*(1), 114-130.

Hasselgård, H. (2004). Adverbials in IT-cleft constructions. In K. Aijmer, & B. Altenberg (Eds.), *Advances in corpus linguistics* (pp. 195-211). Göteborg: Brill.

Hasselgård, H. (2010). *Adjunct adverbials in English*. Cambridge: Cambridge University Press.

Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. *Languages in Contrast, 20*(2), 184-208.

Hengeveld, K. (1989). Layers and operators in functional grammar. *Journal of Linguistics, 25*(1), 127-157.

Holmes, J. (2013). *Women, men, and politeness*. New York, NY: Routledge.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization.* Cambridge: Cambridge University Press.

Horn, L. R. (2001). *A natural history of negation.* Stanford, CA: CSLI Publications.

Hoye, L. (1997). *Adverbs and modality in English.* New York: Longman.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Hunston, S. (2007). Using a corpus to investigate stance quantitatively and qualitatively. In R. Englebretson (Ed.), *Stancetaking in discourse. Subjectivity, evaluation, interaction* (pp. 27-48). Amsterdam: John Benjamins.

Hunston, S. (2011). *Corpus approaches to evaluation: Phraseology and evaluative language.* New York, NY: Routledge.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies, 7*(2), 173-192.

International Organisation for Standardisation. (2004). *ISO 15919: Information and documentation: Transliteration of Devanagari and related Indic scripts into Latin characters.* Geneva: ISO.

James, C. (1980). *Contrastive Analysis.* New York, NY: Longman.

Jawaid, B., & Zeman, D. (2011). Word-order issues in English-to-Urdu statistical machine translation. *Prague Bulletin of Mathematical Linguistics, 95*(1), 87-106.

Jawaid, B., Kamran, A., & Bojar, O. (2014). Urdu monolingual corpus. *Proceedings of the ninth international conference on language resources and evaluation (LREC '14)* (pp. 2938-2943). Reykjavik: European language resources association (ELRA).

Jesperson, O. (1949). *A modern English grammar on historical principles Parts I-VII.* Copenhagen: Einar Munksgaard.

Johansson, S. (2008). *Contrastive analysis and learner language: A corpus-based approach.* Oslo: University of Oslo.

Johansson, S. (2012). Cross-Linguistic Perspectives. In M. Kytö (Ed.), *English corpus linguistics: Crossing paths* (pp. 45-68). New York, NY: Rodopi.

Johansson, S., & Ebeling, O. S. (1998). *Corpora and cross-linguistic research: theory, method and case studies.* Atlanta, GA: Rodopi.

Johansson, S., & Hofland, K. (1994). Towards an English-Norwegian parallel corpus. In S. P. Botley, A. M. McEnery, & A. Wilson (Eds.), *Multilingual corpora in teaching and research* (pp. 134-147). Atlanta, GA: Rodopi.

Kachru, Y. (2008). Hindi-Urdu-Hindustani. In B. B. Kachru, Y. Kachru & S. N. Sridhar (Eds.), *Language in South Asia* (pp. 81-102). Cambridge: Cambridge University Press.

Kádár, D. Z. (2017). *Politeness, impoliteness and ritual*. Cambridge: Cambridge University Press.

Kärkkäinen, E. (2007). The role of *I guess* in conversational stance-taking. In R. Englebretson (Ed.), *Stance-taking in Discourse: Subjectivity, evaluation, Interaction.* (pp. 183-219). Amsterdam: John Benjamins.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of corpus linguistics* (pp. 487-500). New York, NY: Routledge.

Khan, A. J. (2006). *Urdu/Hindi: An artificial divide: African heritage, Mesopotamian roots, Indian culture & British colonialism.* New York, NY: Algora Publishing.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X: Papers* (pp. 79-86). Phuket: Thailand.

Koul, O. N. (2008). *Modern Hindi grammar.* Springfield, VA: Dunwoody Press.

Kučera , H., & Francis, W. N. (1967). *Computational analysis of present day American English.* Providence, RI: Brown University Press.

Kunz, K., Lapshinova-Koltunski, E., Martinez, J., Menzel, K., & Steiner, E. (2018). Shallow features as indicators of English–German contrasts in lexical cohesion. *Languages in contrast, 18*(2), 175-206.

Lambert, J., & Van Gorp, H. (2014). On describing translations. In T. Hermans (Ed.), *The manipulation of literature* (pp. 37-49). London: Routledge.

Lauridsen, K. (1996). Text corpora and contrastive linguistics: Which type of corpus for which type of analysis? In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast* (pp. 63-71). Lund: Studentlitteratur.

Leach, J. (2000). Rhetorical analysis. In M. Bauer, & G. Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 207-226). London: Sage.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105 -122). Berlin: De Gruyter.

Leech, G. (2006). *A glossary of English grammar.* Edinburgh: Edinburgh University Press.

Leech, G. (2015). Descriptive grammar. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 146-160). Cambridge: Cambridge University Press.

Leech, G., & Leonard, R. (1974). A computer corpus of British English. *Hamburger Phonetische Beiträge, 13*, 41-57.

Løken, B. (1997). Expressing possibility in English and Norwegian. *ICAME Journal, 21*, 43-60.

Lyons, J. (1977). *Semantics.* New York, NY: Cambridge University Press.

Mair, C. (2013). The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide, 34*(3), 253-278.

Malchukov, A. L. (2004). Towards a semantic typology of adversative and contrast marking. *Journal of Semantics, 21*(2), 177-198.

Malchukov, A. L., & Xrakovskij, V. S. (2016). The linguistic interaction of mood and modality and other categories. In J. Nuyts & J. Van Der Auwera (Eds.), *The Oxford handbook of modality and mood* (pp. 196 - 220). Oxford: Oxford University Press.

Malmkjær, K. (1998). Cooperation and literary translation. In L. Hickey (Ed.), *The pragmatics of translation* (pp. 25-40). Philadelphia, PA: Mulitilingual Matters.

Malmkjær, K. (2009). What is translation competence? *Revue française de linguistique appliquée, xiv*(1), 121-134.

Mansoor, S. (1993). *Punjabi, Urdu, English in Pakistan: A Sociolinguistic Study.* Lahore: Vanguard.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). *Building a large annotated corpus of English: The Penn Treebank.* Philadelphia, PA: Scholarly Commons.

Martin, J. R., & White, P. R. (2005). *The Language of Evaluation: Appraisal in English.* Basingstoke: Palgrave Macmillan.

McCarthy, M., & Carter, R. (2001). Size isn't everything: spoken English corpus, and the classroom. *Tesol Quarterly, 35*(2), 337-340.

McCarthy, M., & O'Keeffe, A. (2009). Corpora and spoken language. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An international handbook* (Vol. 2, pp. 1008-1024). New York, NY: Walter de Gruyter.

McEnery , T., & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 33-71). Oxford: Blackwell.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics* (second ed.). Edinburgh: Edinburgh University Press.

McEnery, T., & Xiao, R. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* (pp. 3-4). Lisbon, Portugal: European Language Resources Association (ELRA).

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora. The state of the play. In Y. Kawaguchi, T. Takagaki, N. Tomimori & Y. Tsuruga (Eds.), *Corpus-based perspectives in linguistics* (pp. 131-146). Amsterdam: John Benjamins.

McEnery, T., & Xiao, R. (2010). *Corpus-based contrastive studies of English.* New York, NY: Routledge.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, (pp. 382-398). New York, NY: Routledge.

Meyer, C. F. (2008). Pre-electronic corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An international handbook* (Vol. 1, pp. 1-13). Berlin: Walter de Gruyter.

Narrog, H. (2005). On defining modality again. *Language Sciences, 27*(2), 165-192.

Neumann, S. (2005). Corpus Design. Deliverable of the CroCo project 1. Saarbrücken, Germany. Retrieved from http://fr46.uni-saarland.de/croco/

Neumann, S. (2013). *Contrastive register variation.* Berlin: De Gruyter.

Noël, D. (2003). Translations as evidence for semantics: an illustration. *Linguistics, 41*(4), 757-785.

Nuyts, J. (2001). *Epistemic modality, language, and conceptualization: A cognitive-pragmatic perspective.* Amsterdam: John Bejamins.

Nuyts, J. (2005). The modal confusion: On terminology and the concepts behind it. In A. Klinge (Ed.), *Modality: Studies in form and function* (pp. 5-38). London: Equinox.

Nuyts, J. (2016). Analyses of the modal meaning. In J. Nuyts, J. Van der Auwera (Eds.), *The Oxford handbook of modality and mood* (pp. 31-49). Oxford: Oxford University Press.

Och, F. J., & Ney, H. (2000). Improved statistical alignment models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 400-447). Hong Kong: Association for Computational Linguistics.

Palmer, F. R. (1987). *The English verb.* London: Longman.

Palmer, F. R. (2001). *Mood and modality.* Cambridge: Cambridge University Press.

Palmer, F. R. (2014). *Modality and the English modals.* London: Routledge.

Pešková, A. (2019). Slavic and Romance pro-drop in contrast: Evidence from Czech and Spanish. *Languages in contrast, 19*(2), 310-333.

Pic, E., & Furmaniak, G. (2012). A study of epistemic modality in academic and popularised discourse: the case of possibility adverbs "perhaps", "maybe" and "possibly". *Languages for Specific Purposes, 18*, 13-44.

Platts, J. T. (1874). *A grammar of the Hindūstānī or Urdū language.* London: W.M. H. Allen & Co.

Plungian, V. A., & Van der Auwera, J. (1988). Modality's semantic map. *Linguistic Typology, 2*(1), 79-124.

Poutsma, H. (1926). *A grammar of late Modern English.* Groningen: Noordhoff.

Quirk, R. (1974). *The linguist and the English language.* New York, NY: St. Martin's Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.

Rahman, T. (2011). *From Hindi to Urdu: a social and political history.* Karachi: Oxford University Press Pakistan.

Rühlemann, C. (2019). How long does it take to say 'well'? Evidence from the Audio BNC. *Corpus pragmatics, 3*(1), 49-66.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for specific purposes, 13*(2), 149-170.

Sampson, G. (2001). *Empirical Linguistics.* London: Continuum.

Sampson, G. (2005). Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics, 10*(1), 15-36.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Schmidt, R. L. (1999). *Urdu an Essential grammar.* London: Routledge.

Schwenter, S., & Traugott, E. C. (2000). Invoking scalarity: The development of *in fact*. *Journal of historical pragmatics, 1*(1), 7-25.

Scott, M. (2008). Developing WordSmith. *International Journal of English Studies, 8*(1), 95-106.

Seretan, V., & Wehrli, E. (2007). Collocation translation based on sentence alignment and parsing. In N. Hathout, & P. Muller (Eds.), *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles* (pp. 401-410). Toulouse: ATALA.

Shafique, H., Anwar, B., & Shahbaz, M. (2019). An analysis of interactional metadiscourse markers in Urdu journalistic writings. *Orient Research Journal of Social Sciences (ORJSS), 4*(1), 45-60.

Shahid, M. I., Qasim, H. M., & Hasnain, M. (2020). A Cross-linguistic Study of Metadiscourse in English and Urdu. *Corporum: Journal of corpus linguistics, 3*(1), 33-56.

Shamaa, N. (1978). *A linguistic analysis of some problems of Arabic to English translation.* Unpublished doctoral dissertation, University of Oxford.

Sharma, G. (2010). On Hindi Conditionals. In R. Singh (Ed.), *Annual Review of South Asian Languages and Linguistics* (pp. 107-136). Berlin: De Gruyter.

Sifianou, M. (2011). On the concept of face and politness. In F. Bargiela-Chiappini & D. Z. Kádár (Eds.), *Politeness across cultures* (pp. 42 -59). London: Palgrave Macmillan.

Simon-Vandenbergen, A.-M., & Aijmer, K. (2007). *The semantic field of modal certainty: A corpus-based study of English adverbs.* Berlin: De Gruyter.

Sinclair, J. (Ed.). (1996). *Corpus to corpus: a study of translation equivalence.* Oxford: Oxford University Press.

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1- 24). Amsterdam: John Benjamins.

Sinclair, J. (2004). *Trust the text.* (J. Sinclair, & R. Carter, Eds.) London: Routledge.

Squartini, M. (2016). Interactions between modality and other semantic categories. In J. Nuyts & J. van der Auwera (Eds.), *The Oxford handbook of modality and mood* (pp. 50-67). Oxford: Oxford University Press.

Stefanowitsch, A., & Gries, S. T. (2009). Corpora and grammar. In A. Lüdeling, M. Kytö, A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International handbook* (Vol. 2, pp. 933-952). New York: Walter de Gruyter.

Stockwell, R. P., Schacter, P., & Partee, B. H. (1973). *The major syntactic structures of English.* New York, NY: Holt, Rinehart and Winston.

Storjohann, P. (2005). Corpus-driven vs. corpus-based approach to the study of relational patterns. *Proceedings of the Corpus Linguistics 2005 Conference.* Birmingham: University of Birmingham.Retreived from http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics, 7*(2), 215-244.

Stubbs, M. (2007). Quantitative data on multi-word sequences in English: The case of the word world. In M. Hoey (Ed.), *Text, discourse and corpora: theory and analysis* (pp. 163-190). London: Continuum.

Stubbs, M. (2009). Technology and phraseology: With notes on the history of corpus linguistics. In R. Schulze, & U. Römer (Eds.), *Exploring the Lexis-Grammar Interface* (pp. 15 -32). Amsterdam: John Benjamins.

Stubbs, M. (2013). Sequence and order: The neo-Firthian tradition of corpus semantics. In H. Hasselgård, J. Ebeling & S. O. Ebeling (Eds.), *Corpus perspectives on patterns of lexis* (pp. 13-34). Amsterdam: John Benjamins.

Suzuki, D. (2015). Form and function of the modal adverbs: Recent linguistic change and constancy in British English. *Linguistics, 53*(6), 1365-1389.

Suzuki, D. (2018). The semantics and pragmatics of modal adverbs: Grammaticalization and (inter) subjectification of *perhaps*. *Lingua, 205*, 40-53.

Suzuki, D., & Fujiwara, T. (2017). The multifunctionality of 'possible' modal adverbs: A comparative look. *Language, 93*(4), 827-841.

Temperley, D. (2003). Ambiguity avoidance in English relative clauses. *Language, 79*(3), 464-484.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1-13.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Traugott, E. C. (1999). The rhetoric of counter-expectation in semantic change: a study in subjectification. In A. Blank, & P. Koch (Eds.), *Historical semantics and cognition* (pp. 177-196). Berlin: Walter De Gruyter.

Tucker, G. H. (2001). Possibly alternative modality. *Functions of Language, 8*(2), 183-215.

Van der Auwera, J., & Plungian, V. A. (1988). Modality's semantic map. *Linguistic Typology, 2*(1), 79-124.

Van der Auwera, J., Schalley, E., & Nuyts, J. (2005). Epistemic possibility in a Slavic parallel corpus- a pilot study. *Modality in Slavonic languages. New Perspectives*, 201-217.

Van Olmen, D., & Van der Auwera, J. (2016). Modality and Mood in Standard Average European. In J. Nuyts, & J. Van der Auwera (Eds.), *The Oxford Handbook of modality and mood* (pp. 362-384). Oxford: Oxford University Press.

Verhagen, A. (2000). Concession implies causality, though in some other space. In E. Couper-Kuhlen & B. Kortmann (Eds.), *cause - condition - concession - contrast: Cognitive and discourse perspectives* (pp. 361-380). Berlin: De Gruyter.

Vihla, M. (1999). *Medical writing: Modality in focus.* Atlanta, GA: Rodopi.

Vološinov, V. N. (1995). *Marxism and the Philosophy of Language.* (L. Matjka & I. R. Titunik, Trans.) London: Routledge.

Warner, A. (1993). *English auxiliaries: Structure and history.* Cambridge: Cambridge University Press.

Warsi, J. (2004). Conditions under which English is taught in Pakistan: An applied linguistic perspective. *Sarid Journal, 1*(1), 1-9.

White, P. R. (2003). Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text & Talk, 23*(2), 259-284.

Wulff, S., & Baker, P. (2020). Analyzing Concordances. In M. Paquot, & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 161-179). New York, NY: Springer.

Xiao, R. (2008). Corpus creation. In N. Indurkhya, & F. Damerau, *Handbook of natural language processing* (pp. 147-165). New York, NY: Chapman & Hall.

Zaidi, A. (2010). A postcolonial sociolinguistics of Punjabi in Pakistan. *Journal of postcolonial cultures and societies, 1*(3), 22-55.

APPENDIX

TRANSLITERATION ADAPTED FROM DEVANAGRI AND PERSO-ARABIC ISO 15919

Consonants

Perso-Arabic	Transliteration	Perso-Arabic	Transliteration
ا	a	ژ	ž
آ	ā	س	sa
ب	ba	ش	śa
بھ	bha	ص	ṣa
پ	pa	ض	ża
پھ	pha	ط	ṭa
ت	ta	ظ	ẓa
تھ	tha	ع	‘a
ٹ	ṭa	غ	ġa
ٹھ	ṭha	ف	fa
ث	s̲a	ق	qa
ج	ja	ک	ka
جھ	jha	کھ	kha
چ	ca	گ	ga
چھ	cha	گھ	gha
ح	ḥa	ل	la
خ	kh̲a	م	ma
د	da	ن	na
دھ	dha	ں	ṁ
ڈ	ḍa	و	au; va; ō; ū
ڈھ	ḍha	ہ	ha
ذ	z̲a	ء*	-
ر	ra	ی	ī
ز	za	ے	ai; ē
ڑ	ṛ	◌ّ**	-

*ء ***hamzāh*** is pronounced as hiatus and not transcribed here.

** ***taṣadīd*** is a geminated consonant (doubled sound) and is represented by repeating the letter or letter combination.

Short vowels

Perso-Arabic	Transliteration
◌َ	**a**
◌ِ	**i**
◌ُ	**u**
وُ	**ō; ū**
یِ	**i**
ےٕ	**ai**

Long vowels

Perso-Arabic	Transliteration
اَ	**a**
آ	**ā**
اِ	**i**
اِی	**ī**
اُ	**u**
اُو	**ū**
اے	**ē**
او	**ō**
أے	**ai**
أو	**au**