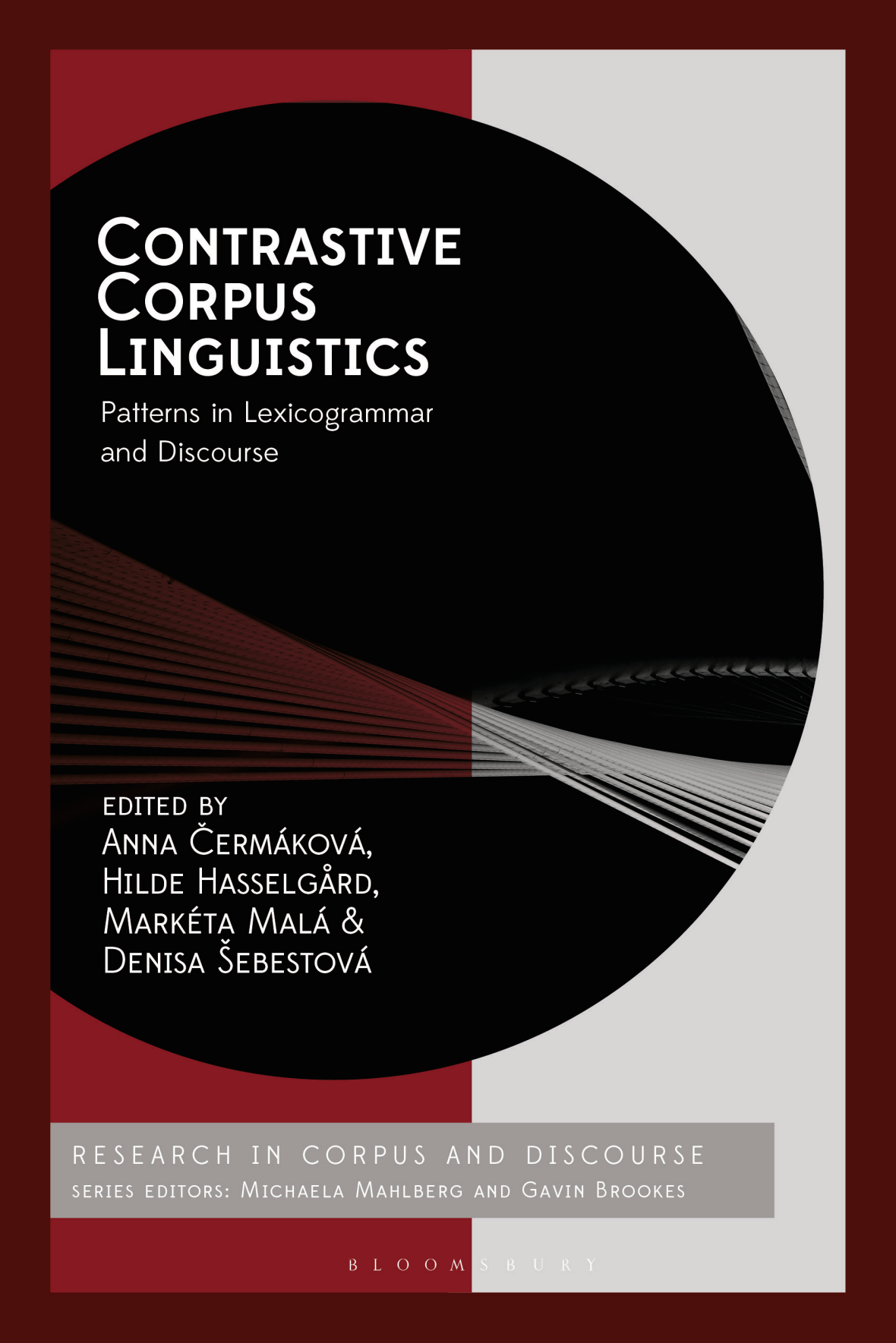




# CONTRASTIVE CORPUS LINGUISTICS

Patterns in Lexicogrammar  
and Discourse

EDITED BY  
ANNA ČERMÁKOVÁ,  
HILDE HASSELGÅRD,  
MARKÉTA MALÁ &  
DENISA ŠEBESTOVÁ

RESEARCH IN CORPUS AND DISCOURSE  
SERIES EDITORS: MICHAELA MAHLBERG AND GAVIN BROOKES

B L O O M S B U R Y

# Contrastive Corpus Linguistics

## Corpus and Discourse

**Series Editors:** Michaela Mahlberg (University of Birmingham, UK) and  
Gavin Brookes (Lancaster University, UK)

**Consulting Editor:** Wolfgang Teubert (University of Birmingham, UK)

### Editorial Board

Paul Baker, Lancaster University, UK  
Frantisek Čermák, Charles University, Prague  
Susan Conrad, Portland State University, USA  
Matteo Fuoli, University of Birmingham, UK  
Maristella Gatto, University of Bari, Italy  
Dominique Maingueneau, Université de Paris XII, France  
Christian Mair, University of Freiburg, Germany  
Alan Partington, University of Bologna, Italy  
Charlotte Taylor, University of Sussex, UK  
Elena Tognini-Bonelli, University of Siena, Italy  
Ruth Wodak, Lancaster University, UK  
Ruihua Zhang, Tianjin University of Science and Technology, China  
Feng Zhiwei, Institute of Applied Linguistics, Beijing, China

Language is ubiquitous. As never before, it is now commonly understood how crucial language is for human interaction, for negotiating and shaping our material and ideational reality. In the digital age, the speed, scale and diversity of forms of communication and language use have grown rapidly. The increasing amount of language data that influences attitudes, decision-making and relationships highlights how the methodology of corpus linguistics together with the explanatory power of discourse analysis are indispensable for deciphering the world around us.

Situated at the interface of corpus linguistics and discourse studies, the *Corpus and Discourse* series publishes innovative research where humanities and social sciences come together to understand the relationship between discourse and society in an increasingly digital world.

### Titles in the Series:

*Academic Vocabulary in Learner Writing*, Magali Paquot  
*Corpus Approaches to the Language of Sports*, edited by Marcus Callies and  
Magnus Levin

*Corpus-Based Approaches to English Language Teaching*, edited by Mari Carmen Campoy, Begona Belles-Fortuno and Maria Lluisa Gea-Valor  
*Corpus, Discourse and Mental Health*, Daniel Hunt and Gavin Brookes  
*Corpus Linguistics and 17th-Century Prostitution*, Anthony McEnery and

Helen Baker

*Corpus Linguistics and World Englishes*, Vivian de Klerk

*Corpus Linguistics in Literary Analysis*, Bettina Fischer-Starcke

*Corpus Stylistics in Heart of Darkness and its Italian Translations*,

Lorenzo Mastropierro

*The Discursive Construction of Economic Inequality*, edited by

Eva M. Gomez-Jimenez and Michael Toolan

*Evaluation and Stance in War News*, edited by Louann Haarman and

Linda Lombardo

*Historical Corpus Stylistics*, Patrick Studer

*Investigating Adolescent Health Communication*, Kevin Harvey

*Keywords in the Press*, Lesley Jeffries and Brian Walker

*Learner Corpus Research*, Vaclav Brezina and Lynne Flowerdew

*Multimodality and Active Listenership*, Dawn Knight

*New Trends in Corpora and Language Learning*, edited by  
Ana Frankenberg-Garcia, Guy Aston and Lynne Flowerdew

*The Prosody of Formulaic Sequences*, Phoebe Lin

*Representation of the British Suffrage Movement*, Kat Gupta

*Rethinking Idiomaticity*, Stefanie Wulff

*Sadness Expressions in English and Chinese*, Ruihua Zhang

*Web as Corpus*, Maristella Gatto



# Contrastive Corpus Linguistics

*Patterns in Lexicogrammar and Discourse*

Edited by

Anna Cermakova, Hilde Hasselgård, Markéta Malá  
and Denisa Šebestová

BLOOMSBURY ACADEMIC  
LONDON • NEW YORK • OXFORD • NEW DELHI • SYDNEY

BLOOMSBURY ACADEMIC  
Bloomsbury Publishing Plc  
50 Bedford Square, London, WC1B 3DP, UK  
1385 Broadway, New York, NY 10018, USA  
29 Earlsfort Terrace, Dublin 2, Ireland

BLOOMSBURY, BLOOMSBURY ACADEMIC and the Diana logo are trademarks  
of Bloomsbury Publishing Plc

First published in Great Britain 2024

Copyright © Anna Cermakova, Hilde Hasselgård, Markéta Malá, Denisa Šebestová  
and Contributors, 2024

Anna Cermakova, Hilde Hasselgård, Markéta Malá, Denisa Šebestová and Contributors have asserted  
their right under the Copyright, Designs and Patents Act, 1988, to be identified as Authors of this work.

For legal purposes the Acknowledgements on p. vii constitute an extension of this copyright page.

Cover design: Elena Durey  
Cover image © Azri Suratmin/Getty

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any  
means, electronic or mechanical, including photocopying, recording, or any information storage or  
retrieval system, without prior permission in writing from the publishers.

Bloomsbury Publishing Plc does not have any control over, or responsibility for, any third-party websites  
referred to or in this book. All internet addresses given in this book were correct at the time of going  
to press. The author and publisher regret any inconvenience caused if addresses have changed or sites  
have ceased to exist, but can accept no responsibility for any such changes.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Cermakova, Anna, 1972- editor. | Hasselgård, Hilde, editor. | Malá, Markéta, 1967- editor. |  
Šebestová, Denisa, editor.

Title: Contrastive corpus linguistics : patterns in lexicogrammar and discourse / edited by Anna  
Cermakova, Hilde Hasselgård, Markéta Malá and Denisa Šebestová.

Description: London ; New York : Bloomsbury Academic, 2024. | Series: Corpus and discourse | Includes  
bibliographical references and index. | Summary: "Marking 30 years of contrastive corpus linguistics,  
this volume provides a state-of-the-art of the field, charting its development over time and pushing the  
boundaries of the discipline. By exploring the application of complex multi-genre multilingual data sets  
and expanding the horizons of contrastive studies, it demonstrates how a juxtaposition of cross-linguistic  
and register variation can deepen our insight into language variation and use"—Provided by publisher.

Identifiers: LCCN 2023048303 (print) | LCCN 2023048304 (ebook) | ISBN 9781350385931 (hardback) |  
ISBN 9781350385979 (paperback) | ISBN 9781350385955 (epub) | ISBN 9781350385948 (ebook)

Subjects: LCSH: Corpora (Linguistics) | Contrastive linguistics. | Discourse analysis. | LCGFT: Essays.

Classification: LCC P128.C68 C636 2024 (print) | LCC P128.C68 (ebook) | DDC 410.188-dc23/  
eng/20231226

LC record available at <https://lcn.loc.gov/2023048303>

LC ebook record available at <https://lcn.loc.gov/2023048304>

ISBN: HB: 978-1-3503-8593-1  
ePDF: 978-1-3503-8594-8  
eBook: 978-1-3503-8595-5

Series: Corpus and Discourse

Typeset by Deanta Global Publishing Services, Chennai, India

To find out more about our authors and books visit [www.bloomsbury.com](http://www.bloomsbury.com)  
and sign up for our newsletters.

# Contents

|                      |     |
|----------------------|-----|
| List of figures      | ix  |
| List of tables       | xi  |
| List of contributors | xiv |

|                                                                                         |   |
|-----------------------------------------------------------------------------------------|---|
| Introduction <i>Hilde Hasselgård, Anna Cermakova, Markéta Malá and Denisa Šebestová</i> | 1 |
|-----------------------------------------------------------------------------------------|---|

|                                                                                                    |    |
|----------------------------------------------------------------------------------------------------|----|
| 1 The Present Status and Recent Trends in Corpus-Based Contrastive Linguistics <i>Karin Aijmer</i> | 11 |
|----------------------------------------------------------------------------------------------------|----|

## Part I Lexicogrammar in Contrast

|                                                                                                                                                                                |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2 Seeing Through Languages and Registers: A Closer Look at the Cognates SEE and SE <i>Signe Oksefell Ebeling</i>                                                               | 29  |
| 3 Periphrastic Genitive Constructions in English and Norwegian <i>Hilde Hasselgård</i>                                                                                         | 62  |
| 4 Double Object Constructions in English and Norwegian: Verbs of SENDING, BRINGING, LENDING and SELLING <i>Thomas Egan</i>                                                     | 87  |
| 5 Prepositional Patterns in English and Czech Newspaper Discourse <i>Denisa Šebestová and Markéta Malá</i>                                                                     | 108 |
| 6 A Cross-Linguistic Study of Journalistic Phraseology <i>Jiajin Xu, Guying Zhou, Xinlu Liu, Yuanyuan Wei, Ruchen Yu and Suhua Zhang</i>                                       | 132 |
| 7 Corpus-Based Contrast in Audiovisual Customization: A Pilot Study on <i>Can/Could</i> and Subject Pronouns in Spanish Dubbing <i>Camino Gutiérrez-Lanza and Rosa Rabadán</i> | 157 |

## Part II Discourse in Contrast

|                                                                                                                                                    |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 8 The Social Functions and Linguistic Patterns of <i>Please</i> and Its Norwegian Correspondences <i>Stine Hulleberg Johansen and Kristin Rygg</i> | 183 |
| 9 Discourse Connectives in English and French: A Contrastive Study on Political Discourse <i>Diana Lewis</i>                                       | 204 |

|    |                                                                                                                                         |     |
|----|-----------------------------------------------------------------------------------------------------------------------------------------|-----|
| 10 | Reporting Verbs in English, Czech and Finnish <i>Anna Cermakova and Lenka Fárová</i>                                                    | 236 |
| 11 | From Dashes to Dashes? – A Contrastive Corpus Study of Dashes in English, German and Swedish <i>Jenny Ström Herold and Magnus Levin</i> | 260 |
|    | Index                                                                                                                                   | 285 |

# A Cross-Linguistic Study of Journalistic Phraseology

Jiajin Xu, Guying Zhou, Xinlu Liu, Yuanyuan Wei, Ruchen Yu and  
Suhua Zhang

## 6.1 Introduction

The identification of a common ground, that is, ‘tertium comparationis’ (TC), is at the outset of most, if not all, contrastive linguistic studies. Contrastive linguists agree that a functional, semantic or conceptual notion may serve as the preferred TC. However, corpus linguistics may be characterized as a form-before-function field of inquiry. Thus, to delineate a comparable unit of analysis for comparing five typologically different languages: English (en), Chinese (cn), Swahili (sw), Arabic (ar) and Malay (ms) (the largest languages in each of the five largest language families, that is, Indo-European, Sino-Tibetan, Niger-Congo, Afro-Asiatic and Austronesian, by number of speakers; Eberhard, Simons and Fennig 2022), is no easy task. When we examine different schools of corpus linguistics, we see that among Birmingham-influenced corpus scholars, the unit of meaning, represented by the phrase, has become a central research focus (Sinclair 1996, 2008). Hence, we take phrase-level units as the TC ad hoc for our current contrastive study. More specifically, we propose to retrieve phrase-frames (henceforth, p-frames) from the five languages as meaning-based TCs for our contrastive analysis. Unlike most previous studies, this study does not focus on one or a few specific linguistic units, but instead on the 100 most frequent phrasal units in our multilingual dataset. We regard phrasal meaning units as a cumulative TC; that is, the aggregate of retrieved chunks plus variable content words (i.e. words in slots) presents a cross-linguistically comparable common entity.

In this study, p-frames serve as a window through which to examine structural and functional characteristics across the five languages. It is hypothesized that

the p-frames of the five languages share similar functional distributions, as per the generic nature of journalism, but different structural patterns, given the morpho-syntactic differences or distances among them. Specifically, these five languages may appear, at least newswise, with different degrees of formulaicity. Some languages may exhibit a higher degree of variability in slot fillers as evidenced by higher type-token ratio (TTR) values; others may exhibit a greater degree of predictability in their use of meaning units, as reflected by higher normalized entropy values. Another very tentative, if not overambitious, aim of this study is to determine whether phraseological formulaicity can serve as an alternative measure for the typological classification of world languages, which has heretofore been informed by, for example, morphology, namely, isolating, agglutinative and inflected languages, as discussed in Sapir (1921), or by word order, namely, the subject-verb-object (SVO) and subject-object-verb (SOV) canonical word orders, as proposed by Greenberg (1963).

## 6.2 Previous studies on phraseological units in contrast

### 6.2.1 Phraseological units in general

As a growing number of researchers have recognized the importance of multi-word units (e.g. Firth 1957; Palmer 1933; Pawley and Syder 1983; Sinclair, Jones and Daley 1970/2004), phraseology has gradually come to the forefront of corpus analysis. This implies that the grammar-lexicon dichotomy may not always hold. As Sinclair (2008: 409) put it, the normal primary carrier of meaning is the phrase, not the word; he considers the phrase ‘quite central and pivotal’ to the description of meaning (2008: 408). Indeed, the term ‘phrase’ is usually loosely used to refer to ‘multi-word combinations’ resulting in varied definitions (see Gray and Biber 2013; Hunston 2022). In this chapter, we take a corpus-driven approach, considering a phraseological unit to be a contiguous and non-contiguous combination of two or more words that has a comparatively high frequency in the corpus and constitutes a semantic unit (cf. Römer 2010: 96).

Corpus studies and phraseology are intimately related, since the prevalence of recurrent multi-word units can hardly be noticed unless one studies texts and corpora (Stubbs 2009: 15). Conceptually rooted in language pedagogy, phraseological units have frequently been studied in corpus-linguistic research on L2 language use. They are often analyzed as having a particular

interest in learner corpus studies, given the significance of acquiring productive knowledge of phraseology for L2 learners. The most common topics addressed in these studies concern the use of phraseological units by non-native and native speakers (e.g. Ädel and Erman 2012; Chen and Baker 2010; De Cock 2004; Juknevičienė 2009; Ren 2022), learners at different proficiency levels (e.g. Garner 2016; Hyland 2008a; Leńko-Szymańska 2014; Römer 2009; Tan and Römer 2022) and learners with different L1 backgrounds (e.g. Juknevičienė and Grabowski 2018; Paquot 2013, 2014; Wang 2016). In addition, phraseological units have received considerable attention in ESP (more specifically, EAP) studies. In this area of research, considerable attention has been focused on extracting and providing lists of pedagogically relevant academic phraseological items (e.g. Ackermann and Chen 2013; Golparvar and Barabadi 2020; Hyland 2008b; Lu, Yoon and Kisselev 2018; Martinez and Schmitt 2012; Simpson-Vlach and Ellis 2010). This line of research also yields important insights into register or genre variation, as it suggests that different genres or registers are characteristic of their corresponding sets of formulaic expressions and vice versa. Thus, such research is mostly concerned with formulaicity across different disciplines (e.g. Cortes 2013; Durrant and Mathews-Aydinli 2011; Omidian, Shahriari and Siyanova-Chanturia 2018), although a series of studies by Biber and colleagues have focused on formulaicity across speech and writing (e.g. Biber 2009; Biber and Barbieri 2007; Biber, Conrad and Cortes 2004; Gray and Biber 2013).

However, the existing literature appears to have given inadequate consideration to phraseology across different languages. A number of studies have been devoted to non-compositional multi-word sequences, such as idioms or metaphors, cross-linguistically; however, as revealed by corpus studies, these opaque units are fairly marginal from a purely quantitative point of view (Colson 2008: 197). Recently, the compositional type has also been the subject of phraseological interest and discussion, particularly among corpus linguists (Ebeling and Ebeling 2013: 2). A growing number of corpus studies have begun to investigate multi-word combinations from a cross-linguistic perspective (e.g. Čermáková and Chlumská 2017; Granger 2014; Tognini-Bonelli 2002; Xiao 2011; Xiao and McEnery 2006). In this chapter, we assume that phraseology should be recognized as a fundamental aspect of language, with different languages encoding it in distinct formulaicities. Specifically, here we take phrase-frames (i.e. p-frames), which have thus far received only limited attention, as our point of departure.

## 6.2.2 P-frames

The corpus-based studies on phraseology mentioned above predominantly focus on continuous multi-word sequences, whereas relatively few investigations have focused on non-continuous phraseological units. Continuous phraseological units, such as lexical bundles, are uninterrupted multi-word strings, as exemplified by *the end of*; by contrast, discontinuous phraseological units such as p-frames feature a variable slot that distinguishes them from lexical bundles. One instance of such phraseological units is *the \* of*, where \* represents a variable slot that can be filled with words such as *rest* or *number*.

An increasing number of studies have focused on discontinuous multi-word sequences. Among others, p-frames allow researchers to account for not only the frequency of recurrent sequences of words but also their fixedness and pattern variability (cf. Forsyth and Grabowski 2015: 540). It has been demonstrated that p-frames can provide valuable insights into the formulaicity of a particular text type (Römer 2010), genre (Vincent 2013), register (Römer 2009) or language (Juknevičienė and Grabowski 2018). In our view, a contrastive analysis of p-frames across different languages allows for a more comprehensive understanding of the formulaic properties of a language.

In the initial stages, Renouf and Sinclair (1991: 128) referred to recurrent discontinuous multi-word expressions as ‘collocational frameworks’, identifying them in a top-down corpus-based way. While this shares similarities with the present-day understanding of p-frames, it differs in that p-frames are derived through a bottom-up, corpus-driven methodology, as exemplified by Fletcher (2002–2007). The predominant trend in this research area is the adoption of frequency-based approaches that utilize automatic identification techniques based on a frequency threshold. In contrast, the present study used a mixed method, combining a corpus-driven approach based on frequency with a manual filtering process that considers semantic completeness when the asterisk-marked slots are filled.

The most commonly examined aspects of p-frames in existing research are their variability, predictability and functions. Variability, also called ‘variant/p-frame ratio’ (VPR), is assessed by computing the TTR of the slot fillers within each p-frame (Römer 2010: 105). The TTR values range from 0 to 1, with values closer to 1 indicating higher variability. A p-frame is considered to exhibit high variability if the ratio of variant types in the slot per p-frame token is high. Predictability, which is used to evaluate the difficulty of determining which variant types will appear in the \* slot of a p-frame, is usually calculated using

normalized entropy (Tan and Römer 2022: 4). It is a measure of the uncertainty of a probability distribution and is computed by dividing the entropy by the logarithm of the number of possible outcomes, thus ranging from 0 to 1. A value closer to 1 indicates an even distribution (i.e. uncertainty), and all variants have an equal likelihood of occurrence. Broadly speaking, these two measures serve as indications of the co-selection between a given p-frame and its filler. In related studies, these measures have been primarily employed to examine whether native and non-native speakers (e.g. Ren 2022), as well as learners at varying levels of proficiency (e.g. Garner 2016; Tan and Römer 2022), demonstrate differential levels of sensitivity with respect to the typical fillers of particular p-frames. Furthermore, p-frames are often analyzed to investigate the relationship between linguistic forms and their rhetorical (e.g. Lu, Yoon and Kisselev 2021) or discourse functions (e.g. Golparvar and Barabadi 2020; Lu, Yoon and Kisselev 2018). Among the various categorization systems, the most commonly used is based on Biber, Conrad and Cortes's (2004) functional taxonomy, originally developed for lexical bundles.

Previous research on p-frames has primarily focused on examining the varying degrees of formulaicity within learner corpora, text types, genres and registers. By utilizing measures such as variability, predictability and functional analysis, these studies shed light on phraseological competence within or across learner groups as well as on formulaic characteristics within or across specific genres of language. Given this background, we assume that p-frames can also reflect varying degrees of formulaicity across languages, which has received limited attention in prior research. In the current study, p-frames are extracted based on frequency thresholds and semantic completeness across the five languages. The variability, predictability and discourse functions of phraseological units are examined to determine the extent to which meaning patterns were shared or distinct across these languages, thus reflecting the elasticity of the basic meaning units of a language.

## 6.3 Corpora and methods

### 6.3.1 Corpora

The data used in this study were newswire texts in Arabic, Chinese, English, Malay and Swahili. The data consists of approximately 176,000 words of

diversely sampled news texts in each language. To maximize data comparability, the Brown corpus sampling frame (Francis and Kučera 1964, 1967) was followed: forty-four reports, twenty-seven editorials and seventeen reviews in five languages were collected.<sup>1</sup> Each language is represented by eighty-eight composite 2000-plus-word news texts, in which short texts were pieced together to form one 2000-word text but saved separately and marked with A, B, C and so on in the filenames. Since the running Chinese texts do not have spaces between words, the Jieba tokenization Python package was used to insert spaces between Chinese words. The genre subtypes were the same across the five languages: politics, sports, finance, society, culture and others. The newspapers in the five languages are broadsheets, mostly national newspapers with some local newspapers. The untagged raw texts were utilized for this study. All articles were published between 2019 and 2022. Hence, news texts in the five languages are highly comparable in terms of genre type, publication date and corpus size.

### **6.3.2 P-frame retrieval and analysis**

Automatic retrieval and manual filtering were performed to collect legitimate three- to five-word discontinuous p-frames with one inner variable slot using AntGram (Anthony 2021). Unlike the morphological and word order approaches to classifying world languages, p-frames can serve as basic meaning units performing basic discourse functions, and they therefore constitute a worthwhile unit of analysis for cross-linguistic and typological generalization. N-grams with initial and final slots were disqualified in line with Garner (2016: 39). P-frames were automatically extracted based on frequency thresholds across the five languages. The length of the p-frames was determined on an incremental, trial basis. In all five languages, two-word phrases do not contain internal slots, and most of them are semantically incomplete. At the other extreme, there were too few instances of p-frames with more than five words in our dataset. Therefore, the top 150 three- to five-word p-frames were used for subsequent manual filtering (see the Appendix for the top 50 p-frames).

Four exclusion criteria were used to manually filter the p-frames: (1) semantically incomplete units; (2) units that cross clausal boundaries; (3) units that only consist of proper nouns, symbols, year or date, or punctuation; and

(4) units that belong to larger p-frames. For example, في \* في ('in \* in'), 月 \* 日 ('month \* day'), to \* the, yang \* dan ('where \* and') and ya \* ya ('of \* of') were excluded. Manual filtering was performed by at least three people on average. Each analyst had to have full professional proficiency in the target language. The second or third analyst double-checked the annotations by the first analyst, as per the exclusion criteria.

Table 6.1 presents examples of frequent p-frames in the five languages. The top 100 p-frames were selected for follow-up analyses. If the 100th p-frame had immediate adjacent p-frames with the same frequency counts, then p-frames with the same frequency were also selected for further analysis. In our dataset, 100 plus 1–5 p-frames were analyzed from the p-frame lists of the five languages.

The variability, predictability and function of p-frames were of central concern in our study. The first two parameters were analyzed quantitatively, and the functional categories were manually annotated and quantified. All three aspects were compared across the five languages.

**Table 6.1** Top 5 p-frames in the five languages

|                | Top 1                                   | Top 2                                      | Top 3                                       | Top 4                                     | Top 5                                   |
|----------------|-----------------------------------------|--------------------------------------------|---------------------------------------------|-------------------------------------------|-----------------------------------------|
| <b>Arabic</b>  | في * من<br>110<br>'in many of'          | من * في<br>96<br>'of its kind in'          | من * أن<br>83<br>'without that'             | من * إلى<br>80<br>'access to'             | على * من<br>76<br>'in spite of'         |
| <b>Chinese</b> | 在 * 的<br>148<br>'with the new'          | 是 * 的<br>122<br>'is ours'                  | 为 * 的<br>105<br>'centred/ oriented'         | 有 * 的<br>91<br>'to some extent'           | 最 * 的<br>90<br>'most important'         |
| <b>English</b> | the * of<br>1152<br><i>the end of</i>   | a * of<br>410<br><i>a lot of</i>           | the * to<br>192<br><i>the right to</i>      | the * that<br>187<br><i>the fact that</i> | the * in<br>183<br><i>the most in</i>   |
| <b>Malay</b>   | dan * yang<br>169<br>'and those who'    | yang * di<br>131<br>'which takes place in' | memberi * kepada<br>102<br>'give effect to' | yang * dalam<br>95<br>'involved in'       | yang * dengan<br>92<br>'related to'     |
| <b>Swahili</b> | na * wa<br>692<br>'and the Minister of' | na * ya<br>628<br>'and some of'            | kwa * ya<br>490<br>'for the sake of'        | katika * ya<br>412<br>'in the case of'    | kwa * wa<br>404<br>'in accordance with' |

*Note:* For non-English p-frames, English translations of the units with the most frequent internal slot words and overall p-frame frequency are provided alongside the original script.

Variability was computed based on the TTR, and predictability was computed based on the normed entropy value of the probability distribution, provided by AntGram. The parameters of variability and predictability are concerned with the extent to which phraseological meaning units in news texts across the five languages adhere to conventionalized meaning patterns, thus reflecting the flexibility or creativity of a language.

The identification of the discourse-pragmatic functions of p-frames was performed qualitatively and independently by at least three people for each of the languages. The functional analysis employed in this study is based on Biber, Conrad and Cortes's (2004) functional taxonomy, which proposes four primary discourse functions: referential, stance, discourse organizer and special conversational expressions. Considering that the language data in the five languages are written text, only the first three categories were identified and assigned to p-frames in the current study. Furthermore, we adhered to the 'variant-based approach' (Lu, Yoon and Kisselev 2018: 79), in which the function of a p-frame was assigned according to the most frequent slot filler. See Table 6.2 for examples of multilingual p-frames with various functions.

Disagreements regarding the inclusion or exclusion of certain p-frames and functional identification were resolved through discussion. Subsequently, we utilized statistical methods based on R (R Core Team 2022) to investigate the structural and functional characteristics of p-frames across the five languages.

**Table 6.2** Examples of discourse-pragmatic functions of p-frames in the five languages

|                | Referential                           | Stance-marking                        | Discourse-organizing              |
|----------------|---------------------------------------|---------------------------------------|-----------------------------------|
| <b>Arabic</b>  | من * إلى<br>'access to'               | لا * أن<br>'it must be'               | من * أخرى<br>'on the other hand'  |
| <b>Chinese</b> | 在 * 中<br>'in the speech of *'         | 最 * 的<br>'the most important'         | 等 * 的<br>'aspects such as'        |
| <b>English</b> | a * of<br><i>a lot of</i>             | was * to<br><i>was supposed to</i>    | as * as<br><i>as well as</i>      |
| <b>Malay</b>   | di * negara<br>'all over the country' | yang * menjadi<br>'which will become' | pada * sama<br>'at the same time' |
| <b>Swahili</b> | katika * ya<br>'in the case of'       | ni * wa<br>'it's time to'             | ili * na<br>'in order to have'    |

*Note:* English translations of meaning units with most frequent slot filler are provided.

To analyze the variability and predictability of the p-frames provided by AntGram, we first created density plots and calculated median values. To determine whether there were statistically significant differences between the languages, a Wilcoxon test was used for pairwise comparisons. To explore the potential association between languages and discourse functions, we employed the chi-squared test and Cramer’s V; additionally, a mosaic plot was used to visualize the strength and direction of the associations between the variables.

6.4 Results and discussion

Table 6.3 summarizes the distribution of identified p-frames across the five languages. For each language, three- to five-word p-frames were selected based on frequency threshold and semantic completeness. When the 100th p-frame had a few immediate adjacent p-frames with the same frequency count, additional p-frames were also included. Therefore, we identified 100(+) p-frames for further analysis for each language.

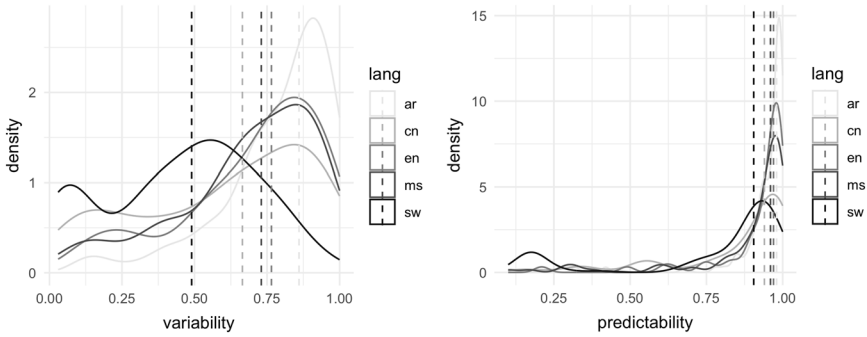
Based on the identified p-frames, the following discussion examines variability, predictability and discourse functions more closely to shed light on formulaic features across the five languages.

6.4.1 Variability and predictability

To attain a holistic understanding of the variability and predictability of p-frames across the five languages, we first used density plots to visualize the distribution of data and provide insights into the underlying data patterns. In Figure 6.1, the density plot on the left reflects the variability of p-frames across

Table 6.3 Summary of the 100 top 3–5-word p-frames of the five languages

|                 | Arabic | Chinese | English | Malay | Swahili |
|-----------------|--------|---------|---------|-------|---------|
| 3-word p-frames | 100    | 96      | 97      | 103   | 90      |
| 4-word p-frames | 0      | 6       | 7       | 2     | 9       |
| 5-word p-frames | 0      | 0       | 0       | 0     | 1       |
| Total           | 100    | 102     | 104     | 105   | 100     |
| Freq. threshold | 14     | 16      | 39      | 21    | 32      |



**Figure 6.1** Density plots of variability and predictability across five languages, two parts.

**Table 6.4** Median variability and median absolute deviation of the 100 top p-frames across the five languages

|                                  | Arabic | Chinese | English | Malay  | Swahili |
|----------------------------------|--------|---------|---------|--------|---------|
| <b>Median</b>                    | 0.8600 | 0.6650  | 0.7650  | 0.7300 | 0.4900  |
| <b>Median Absolute Deviation</b> | 0.1483 | 0.3262  | 0.2298  | 0.2372 | 0.2743  |

the five languages, with the median variability of each language indicated by dashed lines. Similarly, the density plot on the right shows the distribution of predictability and the median value for each language.

In each plot, the  $x$ -axis represents the values of variability (TTR) and predictability (normalized entropy), both ranging from 0 to 1, where a value closer to 1 indicates higher variability or uncertainty (i.e. lower predictability). The  $y$ -axis represents the density, that is, the relative frequency of the data points. Roughly, the distributions of variability and predictability take on similar patterns across the five languages; both plots are left-skewed, with a long tail extending to the left of the peak. The five language groups (ar, cn, en, ms and sw) exhibited a distinct central tendency in their variability. Arabic has the highest peak, at around 0.90, English (en) and Malay (ms) shared a similar peak at around 0.87, and Swahili (sw) and Chinese (cn) showed relatively low peaks at around 0.56 and 0.87. The median variability values of the five languages, indicated by the dashed lines in the plot, can be ranked in the following order:  $0 < sw < cn < ms < en < ar < 1$ . Table 6.4 lists the actual median values and median absolute deviations. Regarding predictability, Arabic (ar) exhibited the highest peak at approximately 1.0, indicating the highest degree of uncertainty among the five groups; Swahili (sw) demonstrated the lowest peak at around 0.9, reflecting a

**Table 6.5** Median predictability and median absolute deviation of the 100 top p-frames across the five languages

|                                  | Arabic | Chinese | English | Malay  | Swahili |
|----------------------------------|--------|---------|---------|--------|---------|
| <b>Median</b>                    | 0.9800 | 0.9400  | 0.9700  | 0.9600 | 0.9050  |
| <b>Median Absolute Deviation</b> | 0.0297 | 0.0741  | 0.0297  | 0.0445 | 0.0815  |

comparatively lower level of uncertainty. The peaks for the intermediate groups – English (en), Malay (ms) and Chinese (cn) – exhibited gradually decreasing intermediate values, indicating correspondingly decreasing levels of uncertainty. The median entropy values of the five languages were ranked as follows:  $0 < sw < cn < ms < en < ar < 1$ ; these are also presented in Table 6.5.

Overall, the descriptive statistics suggest differentiated variability and predictability across the five languages. The median values for TTR and normalized entropy indicate the existence of a flexibility or creativity continuum across the five languages, with Arabic occupying the highest end of the spectrum, followed by English, Malay, Chinese and Swahili, in decreasing order of variability and increasing order of predictability.

However, the statistical significance of these findings requires further investigation. To this end, we employed the Wilcoxon test and applied a multiple-comparison procedure that computes all pairwise comparisons while controlling for the type-I error rate (see Kabacoff 2015: 162). With regard to variability, both Arabic (ar) and Swahili (sw) are statistically different from the other three languages: ar and cn ( $W = 3187.5$ ,  $p < 0.001$ ), ar and en ( $W = 3990.0$ ,  $p < 0.01$ ), ar and ms ( $W = 3725.5$ ,  $p < 0.01$ ), ar and sw ( $W = 1396.5$ ,  $p < 0.001$ ), sw and cn ( $W = 3399.5$ ,  $p < 0.001$ ), sw and en ( $W = 2431.5$ ,  $p < 0.001$ ), and sw and ms ( $W = 2613.0$ ,  $p < 0.001$ ). In contrast, there were no significant differences between Chinese (cn), English (en) and Malay (ms): cn and en ( $W = 2613.0$ ,  $p = 0.107$ ); cn and ms ( $W = 4674.0$ ,  $p = 0.228$ ); and en and ms ( $W = 5180.0$ ,  $p = 0.522$ ).

The situation was very similar in terms of predictability. Significant, pairwise, differences were found for Arabic (ar) and Swahili (SW). The distribution of the predictability values of Arabic (ar) was significantly different from that of the other languages studied here except English: ar and cn ( $W = 3415.0$ ,  $p < 0.001$ ), ar and ms ( $W = 4156.5$ ,  $p < 0.05$ ), ar and sw ( $W = 1904.0$ ,  $p < 0.01$ ), and ar and en ( $W = 4403.0$ ,  $p = 0.171$ ). Swahili (sw) was found to exhibit statistically significant differences from all the other languages: sw and ar ( $W = 1904.0$ ,  $p < 0.01$ ), sw and cn ( $W = 3803.0$ ,  $p < 0.05$ ), sw and en ( $W = 2663.0$ ,

$p < 0.01$ ), and sw and ms ( $W = 3084.0$ ,  $p < 0.01$ ). There were no significant differences between Chinese (cn), English (en), and Malay (ms): cn and en ( $W = 4283.0$ ,  $p = 0.066$ ), cn and ms ( $W = 4645.5$ ,  $p = 0.197$ ), and en and ms ( $W = 5107.5$ ,  $p = 0.419$ ).

In summary, it is statistically supported that Arabic (ar) is the most variable and least predictable, while Swahili (sw) is the least variable and most predictable of the five languages. We thus consider these two languages to occupy opposite ends of the creativity and flexibility continuum. The remaining three languages, English (en), Malay (ms) and Chinese (cn), displayed gradually decreasing levels of variability and increasing levels of predictability; however, this ordering is only supported by the density plots and median values and is not statistically significant. Thus, based on both descriptive and analytical statistics, the flexibility or creativity continuum of the five languages can be roughly summarized as  $sw < (cn < ms < en) < ar$ , where brackets indicate statistically non-significant differences.

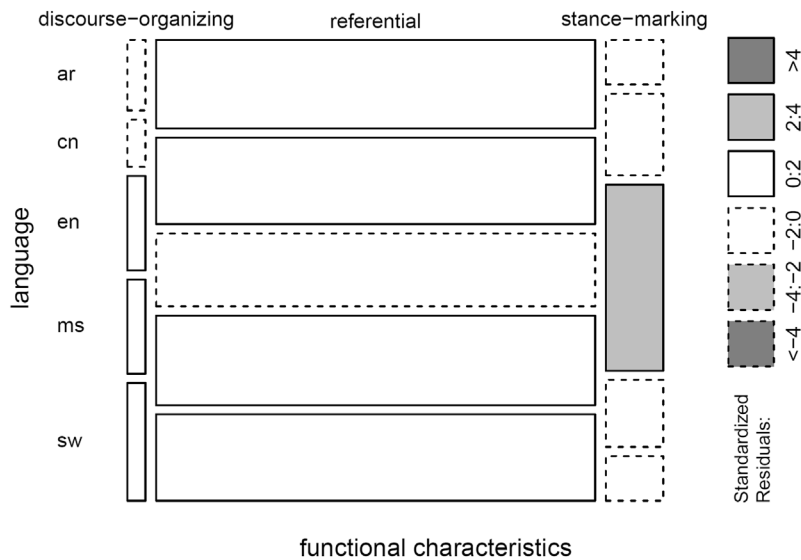
### 6.4.2 Functions

Having examined the variability and predictability of p-frames across the five languages, we now focus on the discourse functions of the identified p-frames. All of the top 100 p-frames were manually classified as either referential, stance-marking or discourse-organizing, following Biber, Conrad and Cortes (2004). The initial round of discourse function labelling of each of the five languages was completed by the author proficient in that language, and all uncertain instances, accounting for about 10 per cent of the data for the five languages, were discussed among the authors and other language experts based on the original p-frames and their English translations until unanimous agreement was reached. By analyzing the functions of these phraseological units, we can gain insights into how they are used cross-linguistically.

Table 6.6 summarizes the functional distributions of p-frames across the five languages. In all languages, referential expressions constitute the largest proportion, followed by stance-marking and discourse-organizing expressions. However, there were statistically significant differences in the functional distribution of p-frames across the five languages, with a small but almost medium effect size ( $\chi^2 = 25.112$ ,  $df = 8$ ,  $p < 0.001$ , Cramer's  $V = 0.157$ ). These results suggest a weak-to-moderate relationship between language and the functional distribution of p-frames.

**Table 6.6** Mean predictability of the 100 top p-frames across five languages

|                      | Arabic | Chinese | English | Malay | Swahili |
|----------------------|--------|---------|---------|-------|---------|
| Referential          | 91     | 89      | 75      | 92    | 89      |
| Stance-marking       | 6      | 11      | 25      | 9     | 6       |
| Discourse-organizing | 3      | 2       | 4       | 4     | 5       |
| Total                | 100    | 102     | 104     | 105   | 100     |



**Figure 6.2** Mosaic plot of functional characteristics across five languages.

Furthermore, a mosaic plot was used to provide a visual representation of the chi-squared and Cramer’s *V* statistics. In particular, a mosaic plot can help visually represent the structure of a contingency table, as well as the strength and direction of associations between the variables. In our case, this will help to identify and illustrate the relationship between discourse functions and different languages.

As shown in Figure 6.2, each rectangle in the mosaic plot represents a cell in the contingency table (see Table 6.6), with its area proportional to the observed frequency. Shading indicates the standardized residual, which reflects the degree to which a cell in a contingency table deviates from what would be expected if the row and column variables are independent. A positive standardized residual value indicates that the observed value in a cell is greater than expected, whereas

a negative standard residual value indicates that the observed value is less than expected, based conventionally on a value of  $\pm 2$  standard deviations. In our case, the mosaic plot demonstrates that the most notable difference between the five languages is the significantly more frequent usage of stance-marking expressions in English than in the other four languages.

Based on our findings, we can draw general conclusions about the distribution of the discourse-pragmatic functions of p-frames in newswire texts across the five languages. First, referential expressions comprise the largest proportion of p-frames in all five languages, followed by stance-marking and discourse-organizing expressions. This trend in the functional distribution of p-frames suggests that referring is a communicative or functionally dominant concept in newswire texts across languages.

Second, our analysis revealed that English employs a significantly higher proportion of stance-marking expressions, which may convey attitudes or assessments of certainty that frame other propositions (Biber, Conrad and Cortes 2004: 384). This difference in language use may be due to different media types and their attributes in the respective countries. For instance, English-language news media are often characterized by commerciality as opposed to state affiliation, which may influence language use in newswire texts. Additionally, the extent to which the English language per se typologically contains more stance-marking expressions than other languages remains in need of further exploration.

The structural variability/predictability and functional categorization of p-frames connect languages on a structural and functional continuum rather than dividing them into discrete islands. Morphological and word-order-motivated language classifications are either too local or syntactic. Phraseological units are the middle ground of unit of language analysis and, most importantly, they are meaning units for which a TC can be better identified for contrastive analysis. Alternatively, the automatically extracted p-frames can be examined using other functional frameworks.

## 6.5 Conclusion

This study demonstrates how a p-frame approach to phraseology can contribute to cross-linguistic comparisons among languages representing the five largest language families. The contrastive analysis sheds light on the structural and

functional characteristics of p-frames in newswire texts across five languages: Arabic, Chinese, English, Malay and Swahili. Overall, we summarized a cross-linguistic variability cline,  $sw < (cn < ms < en) < ar$ , and a cross-linguistic predictability cline,  $sw > (cn > ms > en) > ar$ ; the former represents an increasing degree of creativity, whereas the latter reflects descending idiomaticity across the five languages. Thus, through the window of p-frames, we have attained valuable insights into the complex and dynamic nature of language use, which is in line with how Sinclair (1991: 109–12) described the nature of language: the ‘open-choice principle’ and the ‘idiom principle’. Moreover, regarding discourse functions, we found that the functional distribution of these phraseological units demonstrated a similar pattern across all languages. Specifically, referential expressions constitute the largest proportion of p-frames, followed by stance-marking and discourse-organizing expressions. This indicates the existence of communicative or functional universality in newswire texts across different languages. Our analysis also revealed statistically significant differences in the use of stance-marking expressions, with English employing a significantly higher proportion than other languages.

However, the current study has a major drawback in its generalizability: its reliance on single-genre news-text data. Therefore, the findings must be interpreted with caution, as they may not be representative of the entire language. The extent to which these findings can be generalized as typological features of the five languages, or others, remains to be explored and confirmed. A four-genre corpus project for these five languages is currently underway. The balanced corpora-based p-frame comparison will shed more light on the shared and individualistic phraseological behaviours of the five languages. A methodological limitation should also be noted, as the same frequency thresholds for all p-frame lengths appear to be arbitrary and require better justification in future research.

## Acknowledgements

The work reported in this chapter was supported by the National Social Science Fund of China (Ref. No. 21BYY021) and the support of the National Research for Foreign Language Education, Beijing Foreign Studies University. We would like to thank Yingying Su, Professor of Malay, and Junyu Mao, graduate student

in Arabic at Beijing Foreign Studies University, for their assistance in data collection and analysis. We are also grateful to the reviewers and editors for their constructive comments on the manuscript.

## Note

1. The news corpora for the five languages (i.e. arBrown Press, ToRCH2019, CROWN2021, msBrown Press and swBrown Press) are accessible online at BFSU CQPweb Multilingual Corpus Portal ([http:// 114.251.154.212/cqp/](http://114.251.154.212/cqp/), user ID: test; password: test).

## Appendix. The top 50 phrase-frames across the five languages

### The top 50 phrase-frames in Arabic

| Rank | P-frame    | Frequency | Variability | Predictability | Function             |
|------|------------|-----------|-------------|----------------|----------------------|
| 1    | في * من    | 110       | 0.66        | 0.93           | referential          |
| 2    | من * في    | 96        | 0.94        | 0.99           | referential          |
| 3    | من * أن    | 83        | 0.52        | 0.91           | referential          |
| 4    | من * إلى   | 80        | 0.89        | 0.98           | referential          |
| 5    | على * من   | 76        | 0.66        | 0.88           | discourse-organizing |
| 6    | التي * في  | 67        | 0.93        | 0.99           | referential          |
| 7    | على * في   | 63        | 0.87        | 0.99           | referential          |
| 8    | من * التي  | 59        | 0.9         | 0.98           | referential          |
| 9    | أن * في    | 58        | 0.93        | 0.99           | referential          |
| 10   | في * على   | 53        | 0.81        | 0.97           | referential          |
| 11   | إلى * في   | 52        | 0.96        | 1              | referential          |
| 12   | التي * بها | 51        | 0.75        | 0.96           | referential          |
| 13   | إلى * من   | 50        | 0.8         | 0.98           | referential          |
| 14   | في * التي  | 45        | 0.8         | 0.93           | referential          |
| 15   | في * إلى   | 42        | 0.79        | 0.97           | referential          |
| 16   | الذي * في  | 41        | 0.93        | 0.99           | referential          |
| 17   | من * من    | 41        | 0.98        | 1              | referential          |
| 18   | في * مع    | 40        | 0.68        | 0.89           | referential          |
| 19   | عن * في    | 36        | 0.97        | 1              | referential          |
| 20   | في * ما    | 36        | 0.86        | 0.99           | referential          |
| 21   | لا * أن    | 36        | 0.69        | 0.94           | stance-marking       |

## The top 50 phrase-frames in Arabic (Continued)

| Rank | P-frame      | Frequency | Variability | Predictability | Function       |
|------|--------------|-----------|-------------|----------------|----------------|
| 22   | ما * من      | 36        | 0.86        | 0.98           | referential    |
| 23   | التي * فيها  | 34        | 0.88        | 0.98           | referential    |
| 24   | أن * من      | 33        | 0.73        | 0.94           | referential    |
| 25   | التي * على   | 33        | 0.94        | 0.99           | referential    |
| 26   | في * أن      | 29        | 0.69        | 0.93           | referential    |
| 27   | في * الأخيرة | 29        | 0.41        | 0.86           | referential    |
| 28   | في * العربية | 29        | 0.59        | 0.96           | referential    |
| 29   | لا * في      | 29        | 0.76        | 0.92           | referential    |
| 30   | أن * إلى     | 27        | 0.89        | 0.99           | referential    |
| 31   | بعد * من     | 27        | 0.7         | 0.94           | referential    |
| 32   | قطاع * بنسبة | 27        | 0.44        | 0.94           | referential    |
| 33   | التي * عليها | 26        | 0.96        | 1              | referential    |
| 34   | لا * من      | 26        | 0.69        | 0.89           | stance-marking |
| 35   | ما * في      | 26        | 0.88        | 0.98           | referential    |
| 36   | لا * على     | 25        | 0.72        | 0.91           | referential    |
| 37   | أو * في      | 24        | 0.96        | 1              | referential    |
| 38   | الذي * به    | 24        | 0.83        | 0.98           | referential    |
| 39   | الذي * على   | 24        | 0.88        | 0.99           | referential    |
| 40   | أن * لم      | 23        | 0.96        | 0.99           | referential    |
| 41   | على * أن     | 23        | 0.87        | 0.98           | referential    |
| 42   | في * عن      | 23        | 0.78        | 0.94           | referential    |
| 43   | في * هذا     | 23        | 0.87        | 0.99           | referential    |
| 44   | لم * في      | 23        | 0.87        | 0.99           | referential    |
| 45   | ومن * أن     | 23        | 0.57        | 0.9            | stance-marking |
| 46   | التي * من    | 22        | 0.91        | 0.99           | referential    |
| 47   | في * الذي    | 22        | 0.77        | 0.94           | referential    |
| 48   | في * لا      | 22        | 0.95        | 0.99           | stance-marking |
| 49   | قد * في      | 22        | 0.91        | 0.99           | referential    |
| 50   | مع * من      | 22        | 0.73        | 0.95           | referential    |

## The top 50 phrase-frames in Chinese

| Rank | P-frame   | Frequency | Variability | Predictability | Function             |
|------|-----------|-----------|-------------|----------------|----------------------|
| 1    | 在 * 的     | 148       | 0.78        | 0.95           | referential          |
| 2    | 是 * 的     | 122       | 0.86        | 0.98           | stance-marking       |
| 3    | 为 * 的     | 105       | 0.42        | 0.79           | stance-marking       |
| 4    | 有 * 的     | 91        | 0.87        | 0.98           | stance-marking       |
| 5    | 最 * 的     | 90        | 0.71        | 0.94           | stance-marking       |
| 6    | 对 * 的     | 89        | 0.87        | 0.99           | referential          |
| 7    | 在 * 中     | 75        | 0.68        | 0.95           | referential          |
| 8    | 更 * 的     | 73        | 0.55        | 0.91           | stance-marking       |
| 9    | 中国 * 的    | 70        | 0.73        | 0.97           | referential          |
| 10   | 中国 * 社会主义 | 66        | 0.05        | 0.14           | referential          |
| 11   | 又 * 了     | 61        | 1           | 1              | referential          |
| 12   | 初 心 * 使命  | 60        | 0.05        | 0.76           | referential          |
| 13   | 所 * 的     | 58        | 0.78        | 0.97           | referential          |
| 14   | 文化 * 的    | 55        | 0.6         | 0.93           | referential          |
| 15   | 一个 * 的    | 54        | 0.8         | 0.97           | referential          |
| 16   | 以 * 为     | 51        | 0.65        | 0.87           | referential          |
| 17   | 不 * 的     | 49        | 0.71        | 0.94           | stance-marking       |
| 18   | 在 * 上     | 48        | 0.76        | 0.97           | referential          |
| 19   | 新 * 中国    | 47        | 0.04        | 0.15           | referential          |
| 20   | 一 * 的     | 47        | 0.66        | 0.95           | referential          |
| 21   | 与 * 的     | 47        | 0.94        | 0.99           | referential          |
| 22   | 以 * 的     | 46        | 0.87        | 0.99           | referential          |
| 23   | 发展 * 的    | 41        | 0.76        | 0.95           | referential          |
| 24   | 中华民族 * 复兴 | 38        | 0.07        | 0.52           | referential          |
| 25   | 把 * 的     | 37        | 0.63        | 0.9            | referential          |
| 26   | 也 * 了     | 37        | 0.84        | 0.98           | referential          |
| 27   | 习近平 * 在   | 37        | 0.05        | 0.99           | referential          |
| 28   | 新 * 的     | 36        | 0.43        | 0.8            | referential          |
| 29   | 经济 * 的    | 34        | 0.47        | 0.89           | referential          |
| 30   | 就 * 了     | 34        | 0.91        | 0.99           | stance-marking       |
| 31   | 社会 * 的    | 32        | 0.88        | 0.98           | referential          |
| 32   | 从 * 到     | 32        | 0.94        | 0.99           | referential          |
| 33   | 伟大 * 的    | 32        | 0.19        | 0.54           | referential          |
| 34   | 而 * 的     | 32        | 0.88        | 0.98           | referential          |
| 35   | 都 * 了     | 31        | 0.88        | 0.98           | stance-marking       |
| 36   | 到 * 的     | 28        | 0.9         | 0.99           | referential          |
| 37   | 更加 * 的    | 28        | 0.75        | 0.96           | stance-marking       |
| 38   | 还 * 了     | 27        | 0.82        | 0.98           | discourse-organizing |
| 39   | 很 * 的     | 27        | 0.56        | 0.91           | stance-marking       |
| 40   | 总书记 * 讲话  | 27        | 0.11        | 0.52           | referential          |
| 41   | 被 * 的     | 26        | 0.93        | 0.99           | referential          |
| 42   | 中国 * 地    | 26        | 0.08        | 0.24           | referential          |
| 43   | 时代 * 的    | 25        | 0.62        | 0.81           | referential          |
| 44   | 大 * 的     | 24        | 0.64        | 0.94           | referential          |
| 45   | 世界 * 的    | 24        | 0.67        | 0.93           | referential          |
| 46   | 国家 * 的    | 24        | 0.88        | 0.99           | referential          |
| 47   | 对 * 进行    | 23        | 0.67        | 0.95           | referential          |
| 48   | 中国 * 在    | 23        | 0.52        | 0.86           | referential          |
| 49   | 主席 * 访    | 23        | 0.13        | 0.32           | referential          |
| 50   | 人民 * 的    | 23        | 0.65        | 0.92           | referential          |

**The top 50 phrase-frames in English**

| Rank | P-frame      | Frequency | Variability | Predictability | Function             |
|------|--------------|-----------|-------------|----------------|----------------------|
| 1    | the * of     | 1152      | 0.57        | 0.95           | referential          |
| 2    | a * of       | 410       | 0.55        | 0.91           | referential          |
| 3    | the * to     | 192       | 0.73        | 0.95           | referential          |
| 4    | the * that   | 187       | 0.78        | 0.96           | referential          |
| 5    | the * in     | 183       | 0.89        | 0.99           | referential          |
| 6    | the * is     | 177       | 0.77        | 0.98           | referential          |
| 7    | a * to       | 137       | 0.73        | 0.95           | referential          |
| 8    | the * for    | 135       | 0.78        | 0.96           | referential          |
| 9    | and * to     | 120       | 0.95        | 1              | referential          |
| 10   | is * to      | 120       | 0.56        | 0.93           | stance-marking       |
| 11   | a * that     | 115       | 0.82        | 0.97           | referential          |
| 12   | as * as      | 115       | 0.35        | 0.73           | discourse-organizing |
| 13   | s * to       | 113       | 0.71        | 0.95           | referential          |
| 14   | a * in       | 107       | 0.84        | 0.99           | referential          |
| 15   | the * was    | 103       | 0.86        | 0.99           | referential          |
| 16   | to * in      | 98        | 0.76        | 0.97           | referential          |
| 17   | and * of     | 97        | 0.8         | 0.98           | referential          |
| 18   | in the * of  | 97        | 0.68        | 0.96           | referential          |
| 19   | to * to      | 85        | 0.68        | 0.95           | stance-marking       |
| 20   | of * in      | 84        | 0.92        | 0.99           | referential          |
| 21   | are * to     | 83        | 0.7         | 0.94           | stance-marking       |
| 22   | to * that    | 78        | 0.76        | 0.97           | discourse-organizing |
| 23   | and * in     | 77        | 0.92        | 0.99           | referential          |
| 24   | at the * of  | 76        | 0.47        | 0.9            | referential          |
| 25   | the * has    | 75        | 0.75        | 0.97           | referential          |
| 26   | and * s      | 74        | 0.46        | 0.82           | referential          |
| 27   | to * it      | 74        | 0.78        | 0.95           | stance-marking       |
| 28   | was * to     | 74        | 0.77        | 0.97           | stance-marking       |
| 29   | the * as     | 73        | 0.85        | 0.98           | referential          |
| 30   | in * of      | 72        | 0.53        | 0.88           | referential          |
| 31   | in * to      | 70        | 0.54        | 0.85           | referential          |
| 32   | s * of       | 69        | 0.83        | 0.98           | referential          |
| 33   | be * to      | 68        | 0.71        | 0.94           | stance-marking       |
| 34   | a * for      | 67        | 0.75        | 0.95           | referential          |
| 35   | to * on      | 63        | 0.68        | 0.96           | referential          |
| 36   | the * states | 62        | 0.08        | 0.2            | referential          |
| 37   | it * to      | 61        | 0.69        | 0.92           | discourse-organizing |
| 38   | the * on     | 61        | 0.89        | 0.98           | referential          |
| 39   | for * to     | 60        | 0.72        | 0.97           | stance-marking       |
| 40   | the * said   | 60        | 0.63        | 0.93           | referential          |
| 41   | that * is    | 59        | 0.63        | 0.92           | referential          |
| 42   | to * with    | 59        | 0.59        | 0.94           | referential          |
| 43   | to the * of  | 56        | 0.96        | 1              | referential          |
| 44   | was * in     | 55        | 0.84        | 0.98           | referential          |
| 45   | the * are    | 54        | 0.93        | 0.99           | referential          |
| 46   | on the * of  | 51        | 0.88        | 0.99           | referential          |
| 47   | a * on       | 50        | 0.84        | 0.98           | referential          |
| 48   | i * to       | 50        | 0.56        | 0.9            | stance-marking       |
| 49   | it * not     | 49        | 0.24        | 0.62           | stance-marking       |
| 50   | and * it     | 48        | 0.98        | 1              | discourse-organizing |

**The top 50 phrase-frames in Malay**

| Rank | P-frame            | Frequency | Variability | Predictability | Function             |
|------|--------------------|-----------|-------------|----------------|----------------------|
| 1    | dan * yang         | 169       | 0.91        | 0.99           | referential          |
| 2    | yang * di          | 131       | 0.57        | 0.88           | referential          |
| 3    | memberi * kepada   | 102       | 0.43        | 0.88           | referential          |
| 4    | yang * dalam       | 95        | 0.63        | 0.93           | referential          |
| 5    | yang * dengan      | 92        | 0.64        | 0.96           | referential          |
| 6    | yang * oleh        | 91        | 0.68        | 0.96           | referential          |
| 7    | yang * untuk       | 88        | 0.67        | 0.96           | referential          |
| 8    | di * ini           | 83        | 0.24        | 0.59           | referential          |
| 9    | dengan * yang      | 74        | 0.84        | 0.97           | referential          |
| 10   | baru * ini         | 73        | 0.03        | 0.1            | referential          |
| 11   | dan * di           | 72        | 0.94        | 1              | referential          |
| 12   | dan * untuk        | 70        | 0.94        | 1              | referential          |
| 13   | di * negara        | 67        | 0.27        | 0.7            | referential          |
| 14   | dan * dalam        | 63        | 0.95        | 1              | referential          |
| 15   | yang * pada        | 59        | 0.73        | 0.95           | referential          |
| 16   | dan * dengan       | 57        | 0.89        | 0.99           | referential          |
| 17   | adalah * yang      | 56        | 0.71        | 0.96           | referential          |
| 18   | tidak * dengan     | 53        | 0.77        | 0.97           | referential          |
| 19   | dalam * itu        | 51        | 0.37        | 0.61           | referential          |
| 20   | pada * ini         | 50        | 0.34        | 0.85           | referential          |
| 21   | di * yang          | 49        | 0.65        | 0.94           | referential          |
| 22   | kepada * yang      | 49        | 0.65        | 0.91           | referential          |
| 23   | dalam * yang       | 48        | 0.69        | 0.93           | referential          |
| 24   | berusia * tahun    | 46        | 0.57        | 0.9            | referential          |
| 25   | dalam * ini        | 45        | 0.47        | 0.78           | discourse-organizing |
| 26   | dan * tidak        | 45        | 0.87        | 0.99           | referential          |
| 27   | yang * kepada      | 45        | 0.73        | 0.97           | referential          |
| 28   | yang * tidak       | 44        | 0.77        | 0.98           | stance-marking       |
| 29   | pada * sama        | 41        | 0.05        | 0.17           | discourse-organizing |
| 30   | pada * lalu        | 40        | 0.42        | 0.93           | referential          |
| 31   | ada * yang         | 39        | 0.9         | 0.99           | referential          |
| 32   | dan * negara       | 38        | 0.58        | 0.88           | referential          |
| 33   | di * hari          | 37        | 0.19        | 0.38           | referential          |
| 34   | pada * yang        | 37        | 0.57        | 0.82           | referential          |
| 35   | yang * sebagai     | 37        | 0.76        | 0.96           | referential          |
| 36   | atau * yang        | 36        | 0.94        | 0.99           | referential          |
| 37   | di * itu           | 36        | 0.42        | 0.81           | discourse-organizing |
| 38   | tidak * untuk      | 36        | 0.69        | 0.97           | referential          |
| 39   | yang * ini         | 36        | 0.61        | 0.89           | referential          |
| 40   | negara * yang      | 35        | 0.66        | 0.94           | referential          |
| 41   | satu * yang        | 35        | 0.86        | 0.98           | referential          |
| 42   | sini * ini         | 35        | 0.14        | 0.32           | referential          |
| 43   | di sini * ini      | 35        | 0.14        | 0.32           | referential          |
| 44   | di * hari ini      | 34        | 0.12        | 0.29           | referential          |
| 45   | dan * akan         | 33        | 0.76        | 0.96           | stance-marking       |
| 46   | dan * serta        | 32        | 1           | 1              | referential          |
| 47   | lebih * berbanding | 32        | 0.44        | 0.89           | referential          |
| 48   | dalam * politik    | 31        | 0.55        | 0.93           | referential          |
| 49   | yang * kita        | 31        | 0.74        | 0.96           | stance-marking       |
| 50   | yang * menjadi     | 31        | 0.81        | 0.98           | stance-marking       |

**The top 50 phrase-frames in Swahili**

| Rank | P-frame             | Frequency | Variability | Predictability | Function             |
|------|---------------------|-----------|-------------|----------------|----------------------|
| 1    | na * wa             | 692       | 0.42        | 0.92           | referential          |
| 2    | na * ya             | 628       | 0.42        | 0.93           | referential          |
| 3    | kwa * ya            | 490       | 0.22        | 0.7            | referential          |
| 4    | katika * ya         | 412       | 0.31        | 0.86           | referential          |
| 5    | kwa * wa            | 404       | 0.3         | 0.72           | referential          |
| 6    | na * kwa            | 323       | 0.69        | 0.96           | referential          |
| 7    | kwa * na            | 280       | 0.53        | 0.91           | referential          |
| 8    | ni * wa             | 193       | 0.55        | 0.94           | stance-marking       |
| 9    | katika * wa         | 190       | 0.44        | 0.89           | referential          |
| 10   | katika * za         | 169       | 0.28        | 0.81           | referential          |
| 11   | na * katika         | 158       | 0.84        | 0.98           | referential          |
| 12   | na * la             | 114       | 0.38        | 0.86           | referential          |
| 13   | za * wa             | 107       | 0.71        | 0.96           | referential          |
| 14   | kwenye * ya         | 104       | 0.7         | 0.96           | referential          |
| 15   | katika * la         | 96        | 0.32        | 0.83           | referential          |
| 16   | kuwa * wa           | 96        | 0.56        | 0.89           | referential          |
| 17   | na * kuwa           | 92        | 0.59        | 0.91           | stance-marking       |
| 18   | na * yake           | 91        | 0.52        | 0.91           | referential          |
| 19   | kuwa * ya           | 88        | 0.55        | 0.95           | referential          |
| 20   | na * ni             | 86        | 0.79        | 0.98           | referential          |
| 21   | katika * cha        | 82        | 0.21        | 0.73           | referential          |
| 22   | kuwa * na           | 82        | 0.77        | 0.97           | referential          |
| 23   | kwa * la            | 81        | 0.36        | 0.79           | stance-marking       |
| 24   | kwa * kwa           | 80        | 0.78        | 0.97           | stance-marking       |
| 25   | katika * hiyo       | 79        | 0.53        | 0.95           | referential          |
| 26   | na * mbalimbali     | 76        | 0.59        | 0.91           | referential          |
| 27   | kwenye * wa         | 73        | 0.51        | 0.85           | referential          |
| 28   | na * wake           | 71        | 0.66        | 0.94           | referential          |
| 29   | kwa * wake          | 70        | 0.36        | 0.73           | referential          |
| 30   | kwa * ni            | 66        | 0.53        | 0.86           | referential          |
| 31   | katika * mbalimbali | 64        | 0.48        | 0.87           | referential          |
| 32   | ni * kwa            | 64        | 0.7         | 0.94           | referential          |
| 33   | kwa * za            | 63        | 0.49        | 0.89           | referential          |
| 34   | kwa * katika        | 62        | 0.73        | 0.96           | referential          |
| 35   | na * vya            | 62        | 0.39        | 0.89           | referential          |
| 36   | kwa * hiyo          | 61        | 0.57        | 0.94           | referential          |
| 37   | na * ili            | 61        | 0.95        | 0.99           | discourse-organizing |
| 38   | na * cha            | 53        | 0.45        | 0.93           | referential          |
| 39   | rais * magufuli     | 52        | 0.08        | 0.29           | referential          |
| 40   | rais * tanzania     | 51        | 0.08        | 0.21           | referential          |
| 41   | chama * katika      | 49        | 0.06        | 0.18           | referential          |
| 42   | kuwa * nafasi       | 49        | 0.06        | 0.18           | referential          |
| 43   | na * watatu         | 49        | 0.06        | 0.18           | referential          |
| 44   | naibu * dk          | 49        | 0.06        | 0.18           | referential          |
| 45   | saba * ya           | 49        | 0.06        | 0.18           | referential          |
| 46   | siku * tatu         | 49        | 0.06        | 0.18           | referential          |
| 47   | ili * na            | 48        | 0.71        | 0.96           | discourse-organizing |
| 48   | kupitia * hicho     | 48        | 0.04        | 0.15           | referential          |
| 49   | kutoka * ya         | 48        | 0.54        | 0.92           | referential          |
| 50   | na * hiyo           | 48        | 0.62        | 0.95           | referential          |

## References

- Ackermann, K. and Y. H. Chen (2013), 'Developing the Academic Collocation List (ACL): A Corpus-driven and Expert-judged Approach', *Journal of English for Academic Purpose*, 12 (4): 235–47.
- Ädel, A. and B. Erman (2012), 'Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: A Lexical Bundles Approach', *English for Specific Purposes*, 31 (2): 81–92.
- Anthony, L. (2021), *AntGram* (Version 1.3.0) [Computer Software], Tokyo: Waseda University. <https://www.laurenceanthony.net/software>
- Biber, D. (2009), 'A Corpus-driven Approach to Formulaic Language in English: Multi-word Patterns in Speech and Writing', *International Journal of Corpus Linguistics*, 14 (3): 275–311.
- Biber, D. and F. Barbieri (2007), 'Lexical Bundles in University Spoken and Written Registers', *English for Specific Purposes*, 26 (3): 263–86.
- Biber, D., S. Conrad and V. Cortes (2004), 'If you look at . . . : Lexical Bundles in University Teaching and Textbooks', *Applied Linguistics*, 25 (3): 371–405.
- Čermáková, A. and L. Chlumská (2017), 'Expressing Place in Children's Literature', in T. Egan and H. Dirdal (eds), *Cross-linguistic Correspondences: From Lexis to Genre*, 75–96, Amsterdam: John Benjamins.
- Chen, Y. H. and P. Baker (2010), 'Lexical Bundles in L1 and L2 Academic Writing', *Language Learning and Technology*, 14 (2): 30–49.
- Colson, J. P. (2008), 'Cross-linguistic Phraseological Studies: An Overview', in S. Granger and F. Meunier (eds), *Phraseology: An Interdisciplinary Perspective*, 191–206, Amsterdam: John Benjamins.
- 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.
- De Cock, S. (2004), 'Preferred Sequences of Words in NS and NNS Speech', *Belgian Journal of English Language and Literatures (BELL)*, 2 (1): 225–46.
- Durrant, P. and J. Mathews-Aydinli (2011), 'A Function-First Approach to Identifying Formulaic Language in Academic Writing', *English for Specific Purposes*, 30 (1): 58–72.
- Ebeling, S. and J. Ebeling (2013), *Patterns in Contrast*, Amsterdam: John Benjamins.
- Eberhard, D. M., G. F. Simons and C. D. Fennig, eds (2022), *Ethnologue: Languages of the World*, 25th edn, SIL International. <http://www.ethnologue.com>
- Firth, J. R. (1957), 'A Synopsis of Linguistic Theory, 1930–1955', in *Studies in Linguistic Analysis*, 1–32, London: Blackwell.
- Fletcher, W. H. (2002–2007), *KfNgram* [Computer Software], United States Naval Academy. <http://www.kwicfinder.com/kfNgram/kfNgramHelp.html>
- Forsyth, R. S. and Ł. Grabowski (2015), 'Is there a Formula for Formulaic Language?', *Poznan Studies in Contemporary Linguistics*, 51 (4): 511–49.

- Francis, W. N. and H. Kučera (1964), *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English*, Providence: Brown University Press.  
<http://icame.uib.no/brown/bcm.html>
- Francis, W. N. and H. Kučera (1967), *Computational Analysis of Present-Day American English*, Providence: Brown University Press.
- Garner, J. R. (2016), 'A Phrase-Frame Approach to Investigating Phraseology in Learner Writing across Proficiency Levels', *International Journal of Learner Corpus Research*, 2 (1): 31–67.
- Golparvar, S. E. and E. Barabadi (2020), 'Key Phrase Frames in the Discussion Section of Research Articles of Higher Education', *Lingua*, 236: 1–15.
- Granger, S. (2014), 'A Lexical Bundle Approach to Comparing Languages: Stems in English and French', *Languages in Contrast*, 14 (1): 58–72.
- Gray, B. and D. Biber (2013), 'Lexical Frames in Academic Prose and Conversation', *International Journal of Corpus Linguistics*, 18 (1): 109–36.
- Greenberg, J. (1963), 'Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements', in J. Greenberg (ed.), *Universals of Language*, 73–113, Cambridge, MA: The MIT Press.
- Hunston, S. (2022), *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
- Hyland, K. (2008a), 'Academic Clusters: Text Patterning in Published and Postgraduate Writing', *International Journal of Applied Linguistics*, 18 (1): 41–62.
- Hyland, K. (2008b), 'As Can Be Seen: Lexical Bundles and Disciplinary Variation', *English for Specific Purposes*, 27 (1): 4–21.
- Juknevičienė, R. (2009), 'Lexical Bundles in Learner Language: Lithuanian Learners vs. Native Speakers', *Kalbotyra*, 61: 61–72.
- Juknevičienė, R. and Ł. Grabowski (2018), 'Comparing Formulaicity of Learner Writing through Phrase-frames: A Corpus-driven Study of Lithuanian and Polish EFL Student Writing', *Research in Language*, 16 (3): 303–23.
- Kabacoff, R. I. (2015), *R in Action: Data Analysis and Graphics with R*. Shelter Island, New York: Manning.
- Leńko-Szymańska, A. (2014), 'The Acquisition of Formulaic Language by EFL Learners: A Cross-sectional and Cross-linguistic Perspective', *International Journal of Corpus Linguistics*, 19 (2): 225–51.
- Lu, X., J. Yoon and O. Kisselev (2018), 'A Phrase-frame List for Social Science Research Article Introductions', *Journal of English for Academic Purposes*, 36: 76–85.
- Lu, X., J. Yoon and O. Kisselev (2021), 'Matching Phrase-frames to Rhetorical Moves in Social Science Research Article Introductions', *English for Specific Purposes*, 61: 63–83.
- Martinez, R. and N. Schmitt (2012), 'A Phrasal Expressions List', *Applied Linguistics*, 33 (3): 299–320.
- Omidian, T., H. Shahriari and A. Siyanova-Chanturia (2018), 'A Cross-disciplinary Investigation of Multi-word Expressions in the Moves of Research Article Abstracts', *Journal of English for Academic Purposes*, 36: 1–14.

- Palmer, H. E. (1933), *Second Interim Report on English Collocations*, Tokyo: Kaitakusha.
- Paquot, M. (2013), 'Lexical Bundles and L1 Transfer Effects', *International Journal of Corpus Linguistics*, 18 (3): 391–417.
- Paquot, M. (2014), 'Cross-linguistic Influence and Formulaic Language: Recurrent Word Sequences in French Learner Writing', *EUROSLA Yearbook*, 14 (1): 240–61.
- Pawley, A. and F. H. Syder (1983), 'Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency', in J. C. Richards and R. W. Schmidt (eds), *Language and Communication*, 191–226, London: Routledge.
- R Core Team (2022), 'R: A Language and Environment for Statistical Computing', R Foundation for Statistical Computing. <https://www.R-project.org>
- Ren, J. (2022), 'A Comparative Study of the Phrase Frames Used in the Essays of Native and Non-native English Students', *Lingua*, 274: 1–22.
- Renouf, A. and J. Sinclair (1991), 'Collocational Frameworks in English', in K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics*, 128–43, London: Longman.
- Römer, U. (2009), 'English in Academia: Does Nativeness Matter?', *Anglistik: International Journal of English Studies*, 20 (2): 89–100.
- Römer, U. (2010), 'Establishing the Phraseological Profile of a Text Type: The Construction of Meaning in Academic Book Reviews', *English Text Construction*, 3 (1): 95–119.
- Sapir, E. (1921), *Language: An Introduction to the Study of Speech*, New York: Harvest.
- Simpson-Vlach, R. and N. C. Ellis (2010), 'An Academic Formulas List: New Methods in Phraseology Research', *Applied Linguistics*, 31 (4): 487–512.
- Sinclair, J. M. (1991), *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
- Sinclair, J. M. (1996), 'The Search for Units of Meaning', *Textus*, 9 (1): 75–106.
- Sinclair, J. M. (2008), 'The Phrase, the Whole Phrase, and Nothing but the Phrase', in S. Granger and F. Meunier (eds), *Phraseology: An Interdisciplinary Perspective*, 407–10, Amsterdam: John Benjamins.
- Sinclair, J. M., S. Jones and R. Daley (1970/2004), *English Collocation Studies: The OSTI Report*, ed. R. Krishnamurthy, London: Continuum.
- Stubbs, M. (2009), 'Technology and Phraseology: With Notes on the History of Corpus Linguistics', in U. Römer and R. Schulze (eds), *Exploring the Lexis-Grammar Interface*, 15–32, Amsterdam: John Benjamins.
- Tan, Y. and U. Römer (2022), 'Using Phrase-frames to Trace the Language Development of L1 Chinese Learners of English', *System*, 108: 1–10.
- Tognini-Bonelli, E. (2002), 'Functionally Complete Units of Meaning across English and Italian', in B. Altenberg and S. Granger (eds), *Lexis in Contrast*, 73–95, Amsterdam: John Benjamins.
- Vincent, B. (2013), 'Investigating Academic Phraseology through Combinations of Very Frequent Words: A Methodological Exploration', *Journal of English for Academic Purposes*, 12 (1): 44–56.

- Wang, Y. (2016), *The Idiom Principle and L1 Influence: A Contrastive Learner-corpus Study of Delexical Verb+ Noun Collocations*, Amsterdam: John Benjamins.
- Xiao, R. (2011), 'Word Clusters and Reformulation Markers in Chinese and English: Implications for Translation Universal Hypotheses', *Languages in Contrast*, 11 (2): 145–71.
- Xiao, R. and T. McEnery (2006), 'Collocation, Semantic Prosody, and Near Synonymy: A Cross-linguistic Perspective', *Applied Linguistics*, 27 (1): 103–29.