Contrasting English and Other Languages through Corpora

Contrasting English and Other Languages through Corpora

Edited by

Markéta Janebová,
Ekaterina Lapshinova-Koltunski
and Michaela Martinková

Cambridge
Scholars
Publishing

Contrasting English and Other Languages through Corpora

Edited by Markéta Janebová, Ekaterina Lapshinova-Koltunski
and Michaela Martinková

This book first published 2017

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-9601-2
ISBN (13): 978-1-4438-9601-6

TABLE OF CONTENTS

Part III: Exploring Pragmatic Functions with Parallel Corpora

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In recent decades, significant progress has been made in various areas of corpus-based contrastive analysis. Numerous multilingual corpora, both comparable and parallel (translation), have come into being, accompanied by the development of tools for handling, analyzing and searching these corpora. These resources facilitate contrastive research comparing languages at different levels of description, including morphology, lexico-grammar, semantics, discourse and pragmatics. The elements analyzed include words, phrases, syntactic constructions and discourse elements. Linguistic devices can be compared with regard to form, function and the larger contexts in which they are used. This type of contrastive data helps us to enrich the theoretical description of the languages being compared.

The chapters in this volume result from the contributions to the annual workshop on Language Contrasts at ICAME; all the contributions have undergone a rigorous peer reviewing process, each being assessed by two external reviewers, to whom our thanks are due. There are a number of existing titles that also result from the contributions to the workshop on Language Contrasts in previous years. They include:

Ebeling, Signe Oksefjell, and Hilde Hasselgård, eds. 2015. *Cross-linguistic Perspectives on Verb Constructions*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Aijmer, Karin, and Hilde Hasselgård, eds. 2014. *Cross-linguistic Studies at the Interface between Lexis and Grammar.* Special issue of *The Nordic Journal of English Studies* 14: 1.

Altenberg, Bengt, and Karin Aijmer, eds. 2013. *Text-Based Contrastive Linguistics.* Special issue of *Languages in Contrast* 13: 2.

Similarly to the above-mentioned titles, the present monograph addresses several issues of contrasts between English and other languages. At the same time, the present volume complements the existing ones with respect to the nature of the linguistic phenomena it explores: it covers a wider range of phenomena in lexico-grammar, discourse and pragmatics. As it relies on data from parallel or comparable corpora, it hopes to bring valuable insights into cross-linguistic differences between English and other languages (specifically, Norwegian, Swedish, German, Czech and

Chinese) which might otherwise go unnoticed. The results presented in this volume comprise both quantitative and qualitative analyses of language contrasts and they also demonstrate how contrastive corpus data can contribute to broader theoretical issues.

The volume is subdivided into three parts. The first part contains works on contrasts in phraseology. Signe Oksefjell Ebeling and Jarle Ebeling compare recurrent collocations in English and Norwegian fictional texts, presenting a bottom-up approach to their functional analysis. They use both comparable and parallel corpus data, and their chapter shows that lexico-grammatical differences between English and Norwegian seem to account for many of the differences between the two languages. Moreover, they address a number of challenges emerging as a result of the differences between the languages in terms of recurrence and morphology, as well as the problematic issue of normalizing the frequency of n-grams. Åke Viberg focuses on the uses of the morphemes meaning UP/DOWN and IN/OUT in Swedish and English. The analysis is based primarily on the English-Swedish Parallel Corpus (ESPC). The author classifies all the occurrences of the bound forms in compound verbs in Swedish into semantic fields and then compares them to their translations into English. This contrastive comparison not only serves the analysis of the functions of Swedish verbs, but the results also highlight the general difference in morphological transparency between English and Swedish. The contribution by Hilde Hasselgård is focused on phraseological patterns expressing temporal relations in English and Norwegian. The main question in this analysis is concerned with what phraseological patterns can reveal about how temporal relations are expressed in Norwegian and English. The hypothesis is that English expresses more temporal relations with noun phrases than Norwegian does. The main background for this analysis is previous research showing that Norwegian uses more temporal adverbials than English does.

The second part of the volume includes two studies on contrasts in discourse-related features between English and two other Germanic languages – Norwegian and German. Hildegunn Dirdal concentrates on the analysis of *and* and *og*. The author aims to find quantitative and qualitative differences between their sentence-initial use. Moreover, she also pays attention to the transfer of sentence-initial *and*/*og* in translations and investigates possible reasons why these connectives are often omitted or changed. The other paper in this part (Ekaterina Lapshinova-Koltunski and José Manuel Martínez Martínez) concentrates on the differences and similarities between English and German in terms of various cohesive devices, with a focus on the two modes of production – written and

spoken. The authors use a comparable corpus that contains a number of different registers and apply descriptive statistical methods and techniques derived from automatic text classification approaches. These techniques help them to assess the information which cohesive features contribute to language or mode distinction.

The last part of the volume includes two contributions exploring pragmatic functions in English and other languages with parallel corpora. Markéta Janebová and Michaela Martinková investigate the textual and pragmatic functions of the type nouns *kind* and *sort* using the English-Czech section of the multilingual parallel (translation) corpus InterCorp. Analyzing the NP-internal uses of these type nouns, the authors adopt a different methodology from those used in previous monolingual studies. With the help of the parallel resource, they are able to identify the status of the two type nouns through their correspondences in Czech, which is a typologically different language than English. In the last paper of the volume, Marina Bondi and Danni Yu explore textual and pragmatic functions of direct quotations in Corporate Social Responsibility report speeches in English and Chinese. Observing different ways of voice presentation, the authors aim to understand the motivation of the choices and to identify culture-specific rhetorical preferences (which are also dependent on different argumentative traditions) in the two languages under analysis.

We hope that the present volume will contribute to moving empirical contrastive language studies ahead, as it addresses several issues on contrasts between English and other languages from different research angles. The chapters of this volume are grounded in the latest research. Therefore, the book could be useful both to experts on language studies and advanced students with an interest in linguistics. We also hope that this volume will serve as a catalyst to other researchers and will facilitate further advances in the contrastive analysis of the English language. Last but not least, we hope that the results of the linguistic analyses described in the chapters of the present book will be useful for practical applications in lexicography, language teaching and translation (both human and machine), including translator training.

LIST OF ABBREVIATIONS

BYU-BNC	British National Corpus (BYU interface)
CSR	Corporate social responsibility
Df	Degrees of freedom
DIR	Directional
ENPC	English-Norwegian Parallel Corpus
ENPC+	English-Norwegian Parallel Corpus+
EO	English original
EO	English originals
ESPC	English-Swedish Parallel Corpus
ET	English translation
Eval.	Evaluative
FEI	Fixed expressions and idioms
GECCo	The German-English Contrasts in Cohesion Corpus
GSLC	Gothenburg Spoken Language Corpus
Inform.	Informational
Modal.	Modalizing
MPC	Multilingual Parallel Corpus
N-N	Non-Neuter ("common") gender
NO	Norwegian originals
NT	Norwegian translation
Org.	Organizational
PART	Particle
PD/Q-ADJ	Postdeterminer/qualifying adjective
pmw	per million words
SKT-nouns	Type nouns *sort*, *kind*, *type*
SLB	The Swedish Language Bank
TEI	Text Encoding Initiative
VCV	Verbal Communication Verb

PART I:

CONTRASTS IN PHRASEOLOGY

CHAPTER ONE

A CROSS-LINGUISTIC COMPARISON OF RECURRENT WORD COMBINATIONS IN A COMPARABLE CORPUS OF ENGLISH AND NORWEGIAN FICTION

SIGNE OKSEFJELL EBELING
AND JARLE EBELING

This chapter explores a bottom-up approach to the functional analysis of recurrent word combinations across languages. On the basis of comparable corpus data in English and Norwegian, 3-grams occurring at least 20 times per million words in 25% of the texts were extracted automatically using AntConc. The functional classification, inspired by Altenberg (1998), Moon (1998) and Biber et al. (2004), operates with four main classes: informational, evaluative, modalizing and organizational. The informational 3-grams were further classified into 11 sub-categories. Measuring the effect of language on the function of 3-grams, we show that language plays a role in most functional classes, of which some are favoured in English fiction and others in Norwegian. A more qualitative analysis of evaluative and (informational) spatial 3-grams revealed that lexico-grammatical differences between the two languages seem to account for much of the discrepancy between the two languages. Methodologically, the study reveals a number of challenges, some of which are due to language differences in terms of e.g. recurrence and morphology, while others are due to the problematic issue of normalizing the frequency of n-grams. The study addresses these challenges and an attempt is made to counter some of them by measuring 3-gram token occurrence against 3-gram tokens extracted per text, instead of using a more traditional baseline such as per number of words or s-units.

1. Introduction and Aims

Earlier studies of recurrent word combinations in English and Norwegian fiction texts have found both similarities and differences in the way in which the two languages cluster (Ebeling and Ebeling 2013; Ebeling et al. 2013).[1] Although a bottom-up method was employed to extract the word combinations in these studies, specific combinations were quite rapidly singled out for cross-linguistic analysis, typically on the basis of translational data. Semantic unity was also one of the defining criteria in the case studies reported in Ebeling and Ebeling (2013) and Ebeling et al. (2013).

Thus, little attention has to date been given to a systematic, comprehensive comparison of the actual (functional) nature of recurrent word combinations in English and Norwegian. The aim of this chapter is therefore to offer a more general picture by conducting an analysis of the functions of recurrent word combinations, regardless of semantic unity, in English original texts vs. Norwegian original texts. In other words, it is a study based on comparable data, inspired by previous monolingual studies of English focusing on broad functional categories of sequences of words (e.g. Altenberg 1998; Stubbs and Barth 2003; Biber et al. 2004). At the highest level the functional analysis distinguishes between ideational/informational, interpersonal and textual uses along the lines of Moon (1998), supplemented by a more fine-grained functional classification of the informational category, inspired by Altenberg (1998), in particular. We seek answers to the following questions: to what extent do English and Norwegian original texts overlap in the use of these functional categories? Does any particular function stand out as being typically English or Norwegian? And if so, may this reflect purely linguistic preferences, or may it also point to more culturally entrenched differences? To answer this last question, a closer look at the structure of individual combinations of words will be necessary.

The focus will be on three-word sequences, or 3-grams, extracted from the original texts contained in the English-Norwegian Parallel Corpus+ (ENPC+) of contemporary (comparable) fiction texts. The study is exploratory and experimental in the sense that it is unclear whether the n-gram method will yield valid results between languages. Indeed, as pointed out by Granger (2014, 69): "[C]omparing lexical bundles across languages poses a number of challenges that are not easy to solve." Cases in point are some well-known morphological differences between English

[1] We wish to thank two anonymous reviewers for their suggestions and comments on a previous version of this chapter.

and Norwegian resulting in semantically corresponding sequences not being 3-grams in both languages (see further Section 3.2.1). The background section (Section 2) starts out with a brief survey of some previous cross-linguistic studies of recurrent word combinations, before presenting some relevant frameworks for the classification of such combinations. Section 3 outlines the material and method used, including the classification scheme applied in the study. In Section 4, we discuss some methodological challenges and present some quantitative findings, while some qualitative analyses of the data are presented in Section 5. The concluding Section 6 offers a summary of the findings and their implications.

2. Background

Although using similar automatic, or semi-automatic, methods to extract recurrent word combinations, many previous studies differ in some way or other from the current one. Below we list some notable differences, distinguishing between monolingual and cross-linguistic studies:

- Monolingual studies focusing on
    - varieties of the same language, e.g.
        - genre/register/discipline (e.g. Stubbs and Barth 2003; Biber et al. 2004; Cortes 2004; Hyland 2008; Ebeling 2011);
        - L1 vs. L2 English (e.g. Chen and Baker 2010; Paquot and Granger 2012; Ebeling and Hasselgård 2015);
        - L2 vs. L2 English (e.g. Paquot 2017)
        - translated vs. non-translated texts in the same language (e.g. Baroni and Bernardini 2003; Baker 2004; Xiao 2010; Lee 2013; Ebeling and Ebeling 2017);
    - spoken English (e.g. Altenberg 1998);
    - sequences with semantic unity (e.g. Moon 1998; Baker 2004; Paquot and Granger 2012);
    - sequences of 4 words (e.g. Biber et al. 2004; Hyland 2008);
    - sequences of 3–6 words (e.g. Altenberg 1998);
    - or combinations of two or more of the above.

- Cross-linguistic studies focusing on
    - genre/register (e.g. Granger 2014);
    - sequences with semantic unity (e.g. Ebeling and Ebeling 2013; Ebeling et al. 2013);
    - individual sequences (e.g. Ebeling and Ebeling 2013; Ebeling et al. 2013);
    - sequences with specific functions (e.g. Xiao 2011; Granger 2014);
    - translation as *tertium comparationis* (e.g. Ebeling and Ebeling 2013; Ebeling et al. 2013);
    - sequences of 4 words (e.g. Cortes 2008);
    - sequences of 3–7 words (e.g. Granger 2014).

Some of these are, for different reasons, more relevant to the present study than others, and in the following sections we will discuss in more detail the functional frameworks of word combinations put forward by Altenberg (1998), Moon (1998) and Biber et al. (2004) (Section 2.2), and the "lexical-bundle approach" as applied to comparable data across languages by Cortes (2008) and Granger (2014) (Section 2.1). However, to set the scene, we start Section 2.1 by introducing Teich's (2003) three perspectives on language comparison.

2.1 Background

In her investigation of cross-linguistic variation in system and text, Teich (2003) compares features of English and German from three different perspectives: (i) original vs. original texts in the two languages; (ii) original vs. translated texts in the two languages; (iii) original vs. translated texts in the same languages. For the purpose of the present study we are concerned with the first type of comparison and the relation between 3-grams in English and Norwegian original comparable texts. Importantly, and as pointed out by Teich (2003, 2), once the relation has been established, "it can be used as a basis of comparison for working on the second and third" type. Indeed, results from this study will feed into an investigation conducted in parallel with this one on the function of 3-grams in English original vs. English translated texts (from Norwegian) (Ebeling and Ebeling 2017).

Teich's method differs from the current "knowledge-free method" (Baroni and Bernardini 2003, 85) in taking "previously attested register features" as the starting point to investigate translations and comparable

texts (Teich 2003, 229); she focuses on features such as transitivity, voice and NP complexity rather than strictly functional ones.

Granger (2014) explores a lexical-bundle approach to language comparison in her study of English and French stems across two genres,[2] viz. parliamentary debates and newspaper editorials. Granger discusses some methodological challenges that are also highly relevant for the current study, namely "bundle lengths and minimum frequency thresholds that can ensure the cross-linguistic comparability of the data" (Lefer and Vogeleer 2014, 3). Granger's solution is to include bundles of 3–7 words and, for each bundle length, to select proportionally similar numbers of bundle types across the two languages and genres she investigates. Having decided on these thresholds, she finds both cross-genre and cross-linguistic differences, where genre in some ways seems to play the more significant role in choice of formulaic patterns. While stems that contain a first person (singular) pronoun are dominant in the parliamentary debates in both languages, lexical bundles of this kind are absent in newspaper editorials. Moreover, the sheer number of stem bundles is markedly different across the two genres, viz. "the lists for editorials are much shorter" (Granger 2014, 66).

Cortes (2008) carries out a structural and functional analysis of 4-word lexical bundles in academic history writing in English and Spanish. She finds that the bundle count differs widely between English (87) and Spanish (183), but notes that "there is a high level of agreement among the structural types identified in both corpora" (2008, 48). For example, the most common structural bundle type is prepositional phrase in both English (65%) and Spanish (55%). As for the functions of the bundles analyzed, Cortes finds both similarities and differences between the languages. While most bundles in both languages are referential in nature, discourse organizers were relatively infrequent. However, a slight difference between the languages was noted within the latter category; while the Spanish discourse organizers were typically used for topic-introduction focus, the English ones were used for topic elaboration-clarification (2008, 49). She further observes that English and Spanish have "equivalent expressions" (literal or semi-literal counterparts) at their disposal in 21% of the cases. She suggests that her findings may be valuable to translators and translator trainees, as well as to teachers of academic writing.

[2] The category "stems" includes "sequences that contain a subject and a verb, such as 'I agree with'" (Granger 2014, 59). This is also one of the functional categories that we operate with ("thematic stems"), albeit with a slightly stricter definition than that of Granger; see further Table 1-7.

In contrast to Granger, Cortes chooses a fixed bundle size of 4 words as well as a threshold requiring recurrence of 20 times per million words. In addition, the 4-word bundle has to occur in at least five texts. Both Granger and Cortes address challenges posed by typological and morphological differences between the languages they compare, and both try to minimize their impact. As we have seen, Granger advocates a particular kind of frequency threshold and different bundle lengths to counter some of the challenges. Cortes, on the other hand, groups together lexical bundles containing word forms that show gender and number variation in Spanish "to make the comparison more reliable" (2008, 48).

While our study bears some resemblance to both Cortes (2008) and Granger (2014), the focus is slightly different. Our data are culled from one genre only (fiction), we have opted for a fixed bundle length of 3 and we extract all 3-grams that meet our thresholds, classifying them into four main functional categories, one of which has 11 sub-classes (see Section 3.4). In addition, we aim to statistically measure the effect language has on the functions of n-grams, i.e. does the fact that a text is written in English rather than Norwegian have an effect on the preferred functions of 3-grams?

2.2 Previous Work on the Functional Classification of Word Combinations

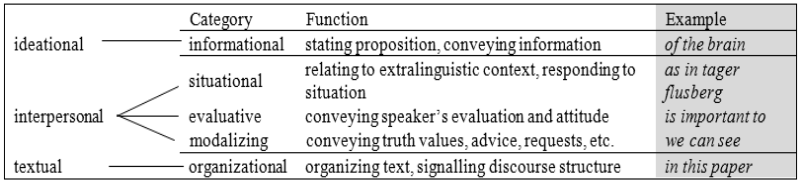
In the following we will present three different frameworks for the classification of sequences of words in text.[3] Two of them include both structural and functional properties (Altenberg 1998; Biber et al. 2004),[4] but, as our main interest is functional, we will keep the structural discussion to a minimum. The third model (Moon 1998) is strictly functional, although the sequences under study – fixed expressions and idioms (FEIs) – have to be structurally "holistic and independent" (1998, 27) in order to qualify as FEIs.

First, to illustrate Moon's (1998) model for functionally classifying FEIs, we refer to Fig. 1-1, which is an adapted version of it used in Ebeling and Hasselgård (2015).

[3] See Cortes (2015, 210) for some other taxonomies.

[4] We will refer to Biber et al. (2004) in our survey of the taxonomy, although earlier, preliminary or similar versions have been presented elsewhere, e.g. Biber et al. (2003), Cortes (2004; 2008), Biber (2006).

Fig. 1-1. The functional classification model (Ebeling and Hasselgård 2015, 93, adapted from Moon 1998, 217)



	Category	Function	Example
ideational	informational	stating proposition, conveying information	*of the brain*
interpersonal	situational	relating to extralinguistic context, responding to situation	*as in tager flusberg*
	evaluative	conveying speaker's evaluation and attitude	*is important to*
	modalizing	conveying truth values, advice, requests, etc.	*we can see*
textual	organizational	organizing text, signalling discourse structure	*in this paper*

Ebeling and Hasselgård (2015, 93) draw attention to the challenge of applying Moon's model to n-grams, which by nature may be neither structurally nor semantically complete units. However, the model is applied to the comparison of 3- and 4-grams in L1 vs. L2 English and across disciplines, yielding interesting results.

Moon explains her model in the following way:

> Ideational, interpersonal, and textual components operate at the highest level – at the level of discourse. In contrast, the text functions of FEIs described here are lower-level functions, reflecting the immediate effects of FEIs within their co-texts . . . (Moon 1998, 218)

Moon's "lower-level" functions correspond to our highest functional level, of which informational and organizational need no further explanation than what is provided in Fig. 1-1, while the three interpersonal functions may need some elaboration. The situational function refers to sequences "typically found in spoken discourse as they are responses to or occasioned by the extralinguistic context", including *excuse me*, *good afternoon*, *thank you very much* (1998, 225–26). Evaluative expressions are "especially associated with the transmission of attitude", while the modalizing ones are "typically epistemic or deontic in nature" (1998, 226). The two latter classes are not always easy to keep apart, as they tend to overlap. However, as Moon allows dual membership for her FEIs, this is not seen to hamper her analysis.

Altenberg (1998) focuses on spoken English and he outlines a comprehensive framework for the classification of "word-combinations consisting of at least three words occurring at least ten times in the corpus [the London-Lund Corpus]" (1998, 102). He starts out by distinguishing between three broad grammatical categories: full clauses (independent and dependent), clause constituents (multiple and single) and incomplete phrases. Within each main (grammatical) type of structure, he recognizes a number of functional types. Table 1-1 shows recurrent word

combinations identified as independent clauses divided into different functional types.

Table 1-1. Some recurrent types of independent clauses (from Altenberg 1998, 104)

Functional type		**Example**
Responses	thanks	*thanks very much*
	reassuring	*it's all right*
	acknowledgement	*oh I see*
	agreement	*(yes) that's right (yes)*
	positive (polar)	*yes it is*
	negative (polar)	*no I haven't*
	disclaimer	*well I don't know*
Epistemic tags		*I don't know*
Metaquestions		*what is/was it*

In our context it is particularly interesting to summarize Altenberg's functional classification of clause constituents, as most of our 3-grams will be shown to fall within this group. Many are also part of the incomplete phrases group, but these are not functionally classified to any large degree by Altenberg.

Most of the word combinations in Altenberg's list of single clause constituents have adverbial functions; they function as temporal or spatial expressions (*at the moment, in this country*), connectors (*first of all*), and qualifying expressions (*more or less*). Other functions include intensifiers/quantifiers (*the whole thing*) and vagueness tags (*sort of thing*).

A more complex framework is offered for the multiple clause constituents, i.e. sequences of two or more clause constituents, such as *and you know*, *and then I*, *there is a . . .*" (1998, 110). Altenberg states that "it is fruitful to see them in a textual perspective and divide them into broad functional categories depending on their function and position in the (unmarked) linear organization of the clause" (ibid.). Inspired by Halliday (1985), Altenberg outlines a positional scheme which includes the following functional slots: Frame, Onset, Stem, Medial, Rheme, Tail, Transition. More than 70% of the multiple clause elements in Altenberg's material (i.e. the London-Lund Corpus) are stems, which have been divided into the functional categories illustrated in Table 1-2.

Table 1-2. Some common types of stems (from Altenberg 1998, 114)

Type	Example
Epistemic	*I don't think (that)*
Existential	*there* BE DET_{INDF}
Reporting	*so I said*
Interrogative	*do you know*
other	*I tried to*

Of the remaining functional positions, Altenberg only discusses Frames (amounting to 7.7% of his multiple clause elements), where some of the most common ones consist of combinations of connectors, response items, discourse items, and modal adverbs, e.g. *and you know*, *well of course*.

Biber et al. (2004) distinguish between three structural types of lexical bundles:[5] VP-based bundles, dependent clause bundles and NP/PP-based bundles (see e.g. Biber 2006, 136), and three primary functional types: stance expressions, discourse organizers and referential expressions (see e.g. Biber et al. 2004, 383ff.). As far as the functional taxonomy is concerned, stance bundles are defined as word sequences that "express attitudes or assessments of certainty", discourse organizers are bundles that "reflect relationships between prior and coming discourse", and referential bundles "make direct reference to physical or abstract entities" (ibid.). Each of the primary functions is divided into several sub-categories, and a condensed overview, including examples, is given in Table 1-3.

Table 1-3. Functional classification of lexical bundles (from Biber et al. 2004, 384ff.)

	Example
I. STANCE EXPRESSIONS	
A. Epistemic stance	
Personal	*I don't know if*
Impersonal	*are more likely to*
B. Attitudinal/modality stance	
Personal	*you want to go*
Impersonal	*it is important to*

[5] "Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their idiomatic status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse" (Biber et al. 1999, 990).

II. DISCOURSE ORGANIZERS **A. Topic introduction/focus** **B. Topic elaboration/clarification**	 *I would like to* *as well as the*
III. REFERENTIAL EXPRESSIONS **A. Identification/focus** **B. Imprecision** **C. Specification** **D. Time/place/text reference**	 *and this is a* *and stuff like that* *the rest of the* *at the same time*
IV. SPECIAL CONVERSATIONAL FUNCTIONS	*thank you very much*

It has been pointed out elsewhere (Ebeling and Hasselgård 2015) that Biber et al.'s (2004) functional classification scheme has clear parallels with that of Moon (1998), where stance expressions, discourse organizers, and referential expressions can be seen to roughly correspond to the interpersonal, textual and ideational functions. Not surprisingly, both Moon's and Biber et al.'s frameworks also have many features in common with that of Altenberg, since all three are in some way or other inspired by Halliday. Their taxonomies thus resemble one another, but the fact that they are arrived at on the basis of different data – including register, size of word sequence and semantic/syntactic unity – they would be expected to operate with different categories. For the purpose of this study, as we shall see below (Section 3.4), we have opted for a mixed functional taxonomy, inspired by all of the above.

3. Material and Method

3.1 The ENPC+

The source of data used in the current study is the English-Norwegian Parallel Corpus+ (ENPC+) (Ebeling and Ebeling 2013). It is a balanced, bidirectional corpus of contemporary fiction in English and Norwegian, containing 39 texts in each language aligned with their translations into the other language. The corpus consists of four relatively equal parts – English originals, Norwegian originals, English translations, Norwegian translations – both in terms of number of texts and number of running words (around 1.3 million in each sub-corpus). We will not exploit the full bidirectional potential of the corpus, as our investigation is confined to a comparison of 3-grams in fiction texts originally written in English and Norwegian. The translations of the texts under study will only to a limited extent be used in the discussion of the results. As our data set is culled

from a subsection of the ENPC+, we will refer to the English texts as the EO sub-corpus and the Norwegian texts as the NO sub-corpus.

3.2 Data Extraction

Due to the relatively small size of our corpus, we chose to focus on 3-grams to get a large, but manageable, set of sequences to analyze, well aware of the challenges we may encounter in the functional classification of such short sequences, and the fact that two 3-gram sequences can be seen to belong in a sense to the same 4-gram.

The 3-grams were extracted using AntConc (Anthony 2014). To ensure recurrence and to avoid idiosyncrasies and topic-related grams, we set quite a conservative threshold, requiring each 3-gram to occur with a frequency of 20 times per million words (= 26 times in 1.3 million words) in at least 25% of the texts (= 10 of the 39 texts in each sub-corpus). We also made some changes to the default settings in AntConc to ensure that: (1) tags/mark-up are not part of the n-grams; (2) apostrophes and hyphens are not word-delimiters, e.g. *n't* is counted as one word and not two;[6] (3) n-grams do not cross s-unit (sentence) borders.[7] By default, AntConc allows n-grams to run across commas, colons and semi-colons (i.e. the program ignores them when calculating word sequences); we chose to do the same. As it turns out, this has relatively little bearing on the number of n-grams extracted for our study, as the high-frequency n-grams (3-grams) tend not to contain punctuation marks anyway.

For the purpose of comparisons across languages, Granger (2014, 61) finds it problematic to set a threshold of x occurrences per million words because "the overall number of n-grams may differ across languages . . . and, as a result, a threshold of 10, for example, may generate a high number of bundles in one language . . . and a much more limited one in another" (ibid.).While we acknowledge that this might pose a problem, we also believe that the fact that this may generate a different number of n-grams in the two languages could point to important cross-linguistic differences. As we have seen (Section 2.1), Granger's "way of countering this effect is to take a similar percentage of the lexical bundles in each

[6] This will not apply to contracted PRON+VERB forms, such as *I'm*, since these have already been split (*I 'm*) in the corpus files.

[7] This meant that the following changes were made to the Global Settings in AntConc: Tags: hide tags and Token definition: User-defined token class; Append following definition: '- for the first two restrictions, while for the third the following change was made to the Tool Preferences: Untick the Replace linebreaks box.

language (variety) for each n-gram size, rather than a specified number of occurrences" (ibid.). For the purpose of this investigation, we have opted for a different approach in order to be able to compare n-grams and n-gram functions across English and Norwegian.[8] In addition to setting the thresholds for data extraction as outlined above, we apply two statistical significance tests to the data, i.e. the *t*-test and the Mann-Whitney-Wilcoxon test. (See further Section 4).

3.2.1 Overview of 3-grams in EO and NO

Table 1-4 gives an overview of our data sets following data extraction with the restrictions specified in Section 3.2.

Table 1-4. Number of 3-gram types and tokens in the ENPC+ with settings (i.e. cut-off at 20 pmw across 25% of the texts)

	3-gram types	**3-gram tokens**
EO	1,408	83,827
NO	895	48,130

We can immediately conclude that the two languages behave differently in terms of 3-gram types (i.e. number of different 3-grams) that meet the threshold requirements. In the same amount of running words (1.3 million), English produces 1,408 3-gram types, while Norwegian only produces 895, and similarly there is a marked difference in the number of 3-gram tokens. The difference between English and Norwegian is statistically significant for both types and tokens when calculated against the number of s-units in each sub-corpus ($p < 0.001$ for both types and tokens, $df = 2$), using a simple pairwise test of equal proportions.[9]

It is striking, albeit perhaps not unexpected, that the number of 3-gram types and tokens in the two languages should differ to the extent shown in Table 1-4. The main explanation for this discrepancy seems to be that Norwegian is generally less recurrent than English. Consider Table 1-5 where the overall number of word types in the two sub-corpora is presented. The Norwegian writers make use of a wider selection of different words (see Table 1-5), which also suggests less recurrence of identical sequences, as is the case in our material. Indeed, if we look at the frequency of 3-grams when no recurrence is required (820,197 vs.

[8] The method advocated by Granger (2014) for English vs. French may be explored for English vs. Norwegian in a future study.

[9] This was performed using the R prop.test (R Development Core Team 2015).

767,169), the Norwegian texts boast a greater number of 3-gram types overall and also a greater number of 3-gram hapaxes. For 3-gram tokens, however, the picture is reversed, i.e. English produces a greater number of 3-gram tokens than Norwegian, on the basis of fewer types; in other words, Norwegian 3-grams recur less often in identical form than English 3-grams.

Table 1-5. Total number of word types and tokens, 3-gram types and tokens and 3-gram hapaxes in the ENPC+

	Word types	Word tokens	3-gram types	3-gram tokens	# of 3-gram hapaxes
EO	37,844	1,316,874	767,169	1,110,564	666,442
NO	67,897	1,310,240	820,197	1,084,100	731,954

This is due to a number of factors including a larger number of accepted spelling/inflectional variants in Norwegian,[10] as well as morphological differences between the two languages. Cases in point are compounding and the encoding of definiteness, i.e. the definite article is a separate word and precedes the noun/NP in English, whereas it is a suffix in Norwegian. For example, *the back door* is a 3-gram that meets the threshold in English, but the corresponding single-word (definite) compound noun in Norwegian, with two spelling variants to encode definiteness: *bakdøren*/*bakdøra* is less likely to be found as part of a recurrent 3-gram. Moreover, the use of *do*-periphrasis in English may also be a contributing factor.

The discrepancy in 3-gram types and tokens following the extraction procedure (Table 1-4) between the two languages poses a challenge for the functional comparison later on, as the frequency of 3-grams in Norwegian will most likely be significantly lower than in English in all functional categories when using number of words or s-units as baseline, i.e. when

[10] These include spelling variants such as *fram*/*frem og tilbake* ("back and forth") and variants of the same inflectional endings such as *snakka*/*snakket* "talked"; both are past/past participle forms of *snakke*, etc. Although both the Norwegian and English sub-corpora contain texts where slang or dialect forms are used, this seems to be more common in the Norwegian texts. Moreover, two of the texts in the Norwegian data are in *nynorsk* (one of the two written standards) and many of the spellings and inflected forms differ from those of the other standard, which is *bokmål*. We were in doubt whether to include texts from both standards in this study, but we opted to include all texts available, as they will all differ from one another in a variety of ways when it comes to spelling, and to some extent also inflections, regardless of language.