Lexical bundles in NS lecturing and NNS lecturing: a corpus-based study

Master's Thesis

MA Linguistics – Language and Communication Coaching
Radboud University

Student:

Evelijn Thijssen

[redacted]

[redacted]

Supervisor:

Sanne van Vuuren

[redacted]

Second reader:

Jarret Geenen

[redacted]

Abstract

This corpus-based study investigated the consequences of English-medium instruction in university lectures. The availability of the Nijmegen English Medium-Instruction Corpus (NEMIC) provided a unique opportunity to investigate lexical differences between the use of formulaic language in two languages produced by the same subject. Data derived from a set of parallel lectures held by the same experienced lecturer in L1 Dutch and L2 English. Differences between the subject's use of four-word lexical bundles in both languages were investigated in terms of their frequency and their functions. A qualitative analysis of the bundles used in each sub-corpus was conducted to investigate the effect of L1 transfer on L2 lexical bundle use. The frequential analysis showed a significantly more frequent use of four-word lexical bundles in L2 lectures compared to L1 lectures. Additionally, a rather restricted bundle variety was observed in L2 lectures, indicating a repeated use of a relatively small set of bundles. The functional comparison of the corpora showed the subject to use significantly more participant-oriented and real-world oriented bundles in non-native lectures compared to native lectures. A preference for participant-oriented bundles was found in both sub-corpora, which is said to be typical of spoken academic language. No evidence was found for L1 transfer effects on L2 lexical bundle use. Even though the study results need to be seen in light of some limitations, the study results indicate linguistic differences between native and non-native instruction at academic universities.

Keywords: English-medium instruction, corpus-based research, spoken academic ELF, formulaic language, bundle frequency variation, bundle function variation, L1 transfer, cross-linguistic influence

Table of contents

1. Introduction

1.1 Background

"Nowadays, millions of people around the world use English as a Lingua Franca (ELF) as a means of international conversation" (Wang, 2017). For exactly this reason, a growing number of universities have made the shift towards English-medium instruction, causing a corresponding research interest in the consequences of using non-native instruction in academic settings (Thøgerson & Airey, 2011). This has resulted in an increasing demand for corpus-based research using data from spoken academic registers such as lectures, seminars, student-student interactions and student-counsellor consultations to investigate linguistic differences between native and non-native lecturing.

A linguistic aspect that has been and continues to be prone to corpus-based linguistic research is the use of lexical bundles. Lexical bundles fall under the umbrella term "formulaic language" and is a term used to refer to bundles of words that often occur together. Lexical bundle studies often aim to compare and explain differences between native speaker (NS) and non-native speaker (NNS) use of lexical bundles. Research into variation in lexical bundle use between NSs and NNSs is important because information about e.g. frequency differences or functional differences between NS and NNS lexical bundle use contributes to our understanding of cross-linguistic influence (i.e. the influence from knowledge of other languages). Differences between native and non-native lexical bundle use furthermore give insight into how prefabricated chunks of language are stored in the mental lexicon (Biber et al., 2004). Additionally, research reporting on cross-linguistic use of formulaic sequences can provide language teachers some insight on how to teach certain linguistic features such as prefabricated chunks and formulaic language (Matsumoto, 2008), and contributes to theories about second language acquisition processes (De Knop, 2018).

1.2 Problem analysis

A few corpus-based studies have tried to explore and explain linguistic differences between native speaker (NS) instruction and non-native speaker (NNS) instruction in academic lectures and university teaching, such as differences in speaking rate (Hincks, 2010; Thøgersen & Airey, 2011) or rhetorical style (Thøgersen & Airey, 2011). However, the number of studies that have investigated NS and NNS differences in spoken language production in academic settings remains scarce. The reason for this scarcity can be attributed to the fact that collecting spoken data for corpus-based research is labour-intensive. First of all because controlling the academic setting during the process of recording spoken language output from one or multiple subjects is difficult, causing the study outcome to be dependent on many variables such as target audience, academic discipline, etc., which are often difficult to control. Besides, the process of recording is time-consuming. Second, multiple recordings, preferably from multiple subjects, are needed in order to generate a representative sample size. Even studies that did manage to use representative data have used relatively small data sets compared to corpus-based studies that have used written data. Third, all recordings need to be transcribed in order to compute text-analyses. This process is again time-consuming and subject to an increase in de number of dependent variables affecting the research outcome, such as individual differences in the application of transcribing conventions. It comes naturally that the collection and investigation of written data for corpus-based research is easier, more time-efficient and statistically more appealing, since it is easier to collect large data sets concerning written (academic) pieces. It is therefore that the number of studies reporting on NS and NNS differences in spoken academic data is still lagging behind on studies reporting on these differences in written academic data (Wang, 2017).

The few studies that do report on NS and NNS differences in spoken academic lexical bundle use mainly compare data from spoken academic registers to data from written academic registers (e.g. Biber, 2006; Biber & Barbieri, 2007; Biber, Conrad & Cortes, 2004). Such studies often aim to compare the use of formulaic sequences (e.g. lexical bundles) across different academic registers or across different academic disciplines. Studies that attempt to exclusively investigate lexical bundle use in spoken academic data often draw entirely upon NS language output, making it impossible to generalise findings about differences in spoken academic use of lexical bundles between NSs and NNSs. If we look at existing literature that aims to fill this gap, we find studies that exclusively report upon NS and NNSs differences in written academic lexical bundle use, without making comparisons to spoken data on lexical bundle use. It can be concluded that there remains a gap between studies that investigated lexical bundles use in spoken academic language output and studies that investigated differences between NS and NNS lexical bundle use. Moreover, corpus-based studies that did report on NS and NNS differences in lexical bundles use in spoken academic registers never managed to exclusively use data from spoken academic university lectures taught by a NS and a NNS. To my knowledge, Wang (2017) is one of the few studies reporting on NNS lexical bundle use in spoken academic settings. However, the data used in Wang's (2017) study was collected from a corpus of spoken academic English as a Lingua Franca (ELF) in university lectures and seminars, making to attempt to compare these NNS results to NS findings. The results of Wang's (2017) study can therefore not inform us about cross-linguistic differences in lexical bundle use. Neither can Wang's results inform us about the effect of NNS lexical bundle use in spoken academic English-medium instruction at universities compared to NS lexical bundle use in spoken academic instruction. So far, only few studies report on NS and NNS differences in lexical bundle use in university lectures alone. Besides, only a very limited number of studies can be found to report on the differences in lexical bundle use between a single subject's native or first language (L1) and that same subject's second language (L2).

Since no studies report on intra-personal differences between native and non-native use of lexical bundles in academic lectures, no sufficient conclusions can be drawn about the linguistic differences between native and non-native instruction in academic settings. Moreover, the scarcity of linguistic research drawing on two languages produced by the same subject leaves a gap in our understanding of cross-linguistic transfer. L1 transfer effects of lexical bundles on the use of English as a second language (ESL) have been investigated using data from L1 speakers and ELF speakers with the same L1 background (e.g. Paquot, 2013). However, limited studies have investigated L1 transfer effects on L2 bundle use by the same subject. Besides, many transfer studies report on cross-linguistic influence using widely spoken languages (e.g. French, Chinese, etc.). Only few studies report on L1 Dutch transfer effects (e.g. Kellerman, 1977), but no studies can be found on L1 Dutch transfer effects on ESL lexical bundle use by the same subject. This study aims to fill these gaps in the literature by investigating frequential and functional differences in lexical bundle use in L1 and L2 spoken academic lectures taught by the same subject. A qualitative analysis of the frequential differences in lexical bundle use is aims to contribute to our understanding of interpersonal L1 transfer effects on ESL lexical bundle use. Existing parallel recordings of L1 Dutch and L2 English spoken academic lectures taught by the same subject at a university in the Netherlands have been made available for this study. Considering the scarcity of spoken academic data for corpus-based research, this data allows for a unique opportunity to investigate NS and NNS differences in the use of lexical bundles in academic lectures. Especially considering the fact that both data sets (i.e. a set of NS lectures and NNS lectures) derive from the same subject, providing an opportunity to investigate intra-personal differences between two languages used by the same subject. This study will therefore not only aim to fill a gap in the existing literature concerning NS and NNS differences in spoken academic language output, it also contributes to studies investigating linguistic differences between different languages within the same subject. The results of this study will therefore expand to our knowledge about linguistic differences between L1 and L2 lecturing and be able to contribute to the consideration of English-medium instruction at universities.

The structure of this thesis is as follows. The following section is used to explain the limitations of previous studies that have led to the three research questions composed for this study. Chapter 2 is used to review existing literature on corpus-based studies that have previously investigated NS and NNS differences in lexical bundle frequency and function and to further explain the research question and hypotheses of this study. Chapter 3 will then describe the datasets used for this study as well as the methodology used for the data analysis. Chapter 4 reports on the results of this study. Chapter 5 is used to discuss the quantitative and qualitative findings, and chapter 6 on the conclusion and implications and limitations of the present study.

1.3 Research question and hypotheses

1.2.1 Bundle frequency variation

The first aim of this study is to investigate differences in four-word bundle frequency between lectures given in the subject's L1 Dutch and lectures given in the subject's L2 English. This has led to the following research question:

RQ1: *Is there a statistically significant difference between the number of four-word lexical bundles used in Dutch NS academic lectures and the number of four-word lexical bundles used in parallel English NNS academic lectures taught by the same subject?*

Contrasting results between previous studies that investigated four-word bundle frequency differences between NSs and NNSs make it difficult to predict bundle frequency variation between two languages produced by the same subject. Adel & Erman (2012) found NNSs to use fewer four-word bundles than NSs did, whereas Bychkovska & Lee found NNSs to use more four-word bundles than NSs did. Besides, both these studies have investigated frequency differences in writing, making it difficult to predict NS and NNS differences in four-word bundle frequency in spoken language production. Wang (2017) is among the first to investigate bundle frequency in academic lecturing. In her study, she found a higher four-word bundle frequency in ELF lecturing than in a corpus containing NS spoken academic data. These results lead to the assumption that four-word lexical bundle use should be more frequent in L2 lecturing than in L1 lecturing performed by the same subject. This assumption has led to the following hypotheses regarding the first research question:

H0: There is no significant difference between four-word lexical bundle frequency in NS Dutch lectures and four-word lexical bundle frequency NNS English lectures taught by the same subject.

H1: Four-word lexical bundle frequency is significantly higher in NS English lectures than in NNS Dutch lectures taught by the same subject.

1.2.2 Cross-linguistic interference

The second aim of this study is to investigate L1 transfer effects. Parallel use of L2 lexical bundles and their congruent forms in the leaner's mother tongue can be an indicator of L1 transfer, which occurs when L2 lexical bundle use is influenced by L1 lexical bundle knowledge. Frequency data is often used to investigate cross-linguistic influence (Biber et al., 2004). Bundles that are highly frequent in the learner's L1 often result in an overuse of the L2 translational equivalent or congruent form (Paquot, 2013). As mentioned before, the number of studies reporting on intra-personal L1 transfer effects remains scarce. Previous studies have mainly investigated L1 transfer effects on lexical bundle use by comparing NS datasets to NNS datasets. Other studies have observed L1 transfer effects on lexical bundle use by comparing L2 data from learners with various L1 backgrounds to L2 data from learners from one specific L1 background (e.g. Paquot, 2013). Moreover, no data exists on L1 Dutch transfer effects on L2 English use by the same subject. In order to investigate whether L1 transfer of highly frequent L1 bundles occurs in L2 production by the same speaker, the following research question has been composed:

RQ2: *Are highly frequent four-word lexical bundles in L1 Dutch lectures transferred to L2 lexical bundle use by the same subject?*

It is expected that patterns of L1 transfer between languages produced by the same subject are similar to patterns of L1 transfer that have been observed between groups of L1 users and groups of EFL learners with the same L1 background. It is therefore assumed that highly frequent bundles in the subject's L1 Dutch will transfer to the use of its congruent forms in L2 lexical bundle use. This assumption has led to the following hypotheses regarding the second research question:

H0: Highly frequent four-word lexical bundles in L1 Dutch lectures are not transferred to L2 lexical bundle use in L2 English lectures taught by the same subject.

H1: Highly frequent four-word lexical bundles in L1 Dutch lectures are transferred to L2 lexical bundle use in L2 English lectures taught by the same subject.

1.2.3 Bundle function variation

The third part of the study aims to compare four-word lexical bundle functions between lectures given in the subject's native language (NL) Dutch and the subject's non-native language (NNL) English. This has led to the third and final research question that will be addressed in this study:

RQ3: *Is the functional distribution of four-word lexical bundles significantly different in English NNS lectures compared to Dutch NS lectures taught by the same subject?*

Previously conducted studies have reported incongruent results regarding the functional differences of lexical bundles in NS and NNS academic writing. Ädel and Erman (2012) have reported NSs and NNSs to demonstrate a similar distribution of lexical bundles over the three main functional categories, whereas Byckovska and Lee (2017) reported contrasting results. They found L2 writers to use almost twice as many discourse organisers than L1 users. Moreover, they found L2 writers to use significantly more stance bundles than L1earners did in academic writing. Literature reporting lexical bundle functions in spoken academic registers report a dominant use of stance bundles in ELF academic lectures (e.g. Wang, 2017) and NS classroom teaching (Biber & Barbieri, 2007). Moreover, Biber and Barbieri (2007) find a preference for stance bundles exceptionally common in classroom teaching compared to other academic registers. Since the previous studies concerning written academic data have reported NNS output to show signs of overuse of certain functional categories, it is expected that this trend will also be visible in spoken NNS output. Based on these findings it is predicted that the subject will use a significantly higher variation in bundle functions in NNL lecturing (English) than in NL lecturing (Dutch). This has led to the following hypotheses regarding the second research question:

H0: There is no significant difference between the functional distribution of four-word bundles in NS Dutch lectures and the functional distribution of four-word bundles in NNS English lectures taught by the same subject.

H1: The functional distribution of four-word bundles is significantly greater in NNS English lectures than in in NS Dutch lectures taught by the same subject

2. Literature review

2.1 Lexical bundles and language learning

Second language learning is a complex process which we can begin to explain by looking at language storage in the mental lexicon. The mental lexicon is considered to be a hypothetical network in our brains in which we store information about words. All information we know about a word, such as individual sounds that are combined to form a word, spelling, meaning, word associations and word use, are all stored in the mental lexicon. But the mental lexicon does not only store receptive knowledge about a word, productive knowledge such as pronunciation and grammatical constraints need to be stored as well. Language learners generally have a wider range of receptive lexical knowledge than productive lexical knowledge (Gass, 2013), which explains why second language production is normally preceded by second language recognition.

Even more complex vocabulary to learn are formulaic sequences. "Lexical bundles" are among the most commonly investigated type of formulaic language. The term "lexical bundle" is used to refer to bundles of words that often occur together (Gass, 2013), such as *deep sigh* and *broad daylight* (Gass, 2013). Lexical bundles have been studied under many different names, such as "lexical chunks", "collocations" and "multi-word units". All are examples of formulaic language but with a slightly different meaning. Lexical bundles are best defined as "sequences of a fixed number of words which tend to recur in a particular register" (Bestgen & Granger, 2018). Lexical bundles are therefore often studied under the name "N-grams", referring to the fixed number of words that make up the lexical bundle. The terms "lexical bundle" and "N-Gram" will be used interchangeably throughout this study. Lexical bundles are not to be confused with collocations. Similar to N-grams, collocations can be a sequence that consists of multiple words. However, the individual words that make up a collocation are strongly associated, e.g. *to take a walk*, whereas association patterns in lexical bundles are often unspecified (Bestgen & Granger, 2018). Additionally, lexical bundles always occur alongside each other whereas the lexical units that make up a collocation can be adjoined but do not necessarily have to be. For example, the individual units that make up the collocation *to take a walk* can be used contiguously but do not necessarily have to occur alongside each other, as is the case in the example expression *to take a long walk*. The term "collgram" has recently been introduced by Bestgen and Granger (2018) to define a combination of collocates and N-Grams. Collgrams are fixed lexical bundles that include the association patterns that occur in collocates.

Lexical bundles also differ from multi-word structures that are idiomatic in meaning (Kashiha & Heng, 2014), such as *to spill the beans*, and multi-word structures that represent a concept that can be described using one word in other languages (Gass, 2013). An example of the latter would be the Dutch collocation *broodmes* and its two-word English equivalent *bread knife*. There are a few characteristics by which lexical bundles distinguish themselves from idiomatic multi-word structures. The main difference between lexical bundles and idiomatic structures is that lexical bundles are transparent in meaning, as opposed to idiomatic multi-word sequences which are often figurative in meaning (Cieślicka, 2015). Transparency refers to the extent to which a lexical structure can be deduced from its literal meaning (Cieślicka, 2015). Moreover, lexical bundles are highly frequent, unlike other idiomatic multi-word structures (Kashiha & Heng, 2014). Cortes (2004) found that lexical bundles occur up to twenty times more often per million words than some well-known idioms. Finally, idiomatic structures are relatively fixed expressions (Bychkovska & Lee, 2017), whereas lexical bundles behave more like collocations, which are compositional (Cieślicka, 2015).

Research shows that L2 learners' knowledge of formulaic expressions is often behind on their general L2 proficiency (Steinel, Hulstein, & Steinel, 2007). This cannot only be explained by the fact that lexical bundles and idioms are language-specific but because L2 learners need to learn such lexical structures as a whole. That is because formulaic sequences such as lexical bundles, collgrams and collocations are prefabricated chunks of words. Whether they are continuous or discontinuous, they need to be stored and retrieved from the memory as a whole (Cieślicka, 2015). This explains the finding that multi-word constructions, such as lexical bundles, are more easily retrieved by NSs than by NNSs (Siyanova & Schmitt, 2008), why lexical bundles are often found to be less common in NNSs' language production than in NSs' language production, and why L2 learners with a wide knowledge of lexical bundles come-across as more proficient or native-like. It is important to note that just because formulaic sequences are stored and processed holistically does not mean that they are non-compositional. Moreover, their non-transparency causes interpretation of formulaic language to be dependent on context (Gass, 2013), which makes acquisition of lexical bundles difficult. In his Lexical Approach, Lewis (1993) describes knowledge of such prefabricated "lexical chunks" to be the main factor in building L2 proficiency and "chunk noticing" in an authentic

L2 environment to be the main method behind the acquisition of such chunks. Only after repeated exposure do we learn that a combination of certain words, such as *to take a bath,* are used as a fixed lexical structure rather than other combinations that would carry a similar meaning, such as *to do a bath* (Gass, 2013). It is therefore that knowledge of formulaic language is considered to be an indicator of high L2 proficiency. A study by Boers et al. (2006) has shown that L2 learners of English who had been exposed to formulaic language using the lexical approach were indeed perceived as more proficient L2 speakers than their controls. It is suggested that this native-like perception of L2 speakers with a wide knowledge of formulaic vocabulary can be explained by the fact that such pre-fabricated chunks of words are ready-to-use and do not require any lexical planning. The time that is saved by using formulaic language is therefore believed to facilitate fluency (Cieślicka, 2015). Speakers who use more formulaic sequences in spoken L2 production therefore come across as more fluent and therefore more native-like than L2 speakers that do not use as much formulaic language. This explanation causes many to consider a wide knowledge of lexical bundles and other formulaic sequences to be the main factor in building fluency in a second language. They furthermore help give meaning and add coherence to a text or speech (Kashiha & Heng, 2014). Lexical bundles are therefore considered to be essential in achieving native-like competence (Bychkovska & Lee, 2017).

2.2 Learner corpus research

Over the last decade, the differences between native speaker (NS) and non-native speaker (NNS) lexical bundle use have been widely investigated within the field of corpus linguistics. That is because differences between NS lexical bundle use and NNS lexical bundle use can help us explain differences between L1 and L2 language production. It can furthermore help us to further determine L2 competency. Learner corpus research is a linguistic methodology that uses electronic collections of naturally occurring language output called “corpora” (Granger & Hung, 2002) that consist of data from language learners (hence the term “learner corpus research”). These corpora allow for computer-based comparisons of (usually) large amounts of language output, whether spoken or written. Differences between corpora (i.e. data groups) can be investigated using text-analysis tools such as concordance software. Sequences of words can be extracted automatically, after which they can be used for linguistic analyses. Learner corpus research is necessary because it helps us to identify what linguistic aspects are difficult for second language learners by investigating differences between native language speaker and second language speaker output. Not only can this method be applied to identify general learner difficulties, corpus linguistics also allows for an identification of linguistic difficulties for a very specific group of learners. Results from corpus-based research therefore helps us to expand our knowledge on language learning pedagogy (De Knop, 2018). Not only can these insights help direct language teachers in deciding what features should be emphasized in foreign language teaching, corpus-linguistics can also provide language teachers some insight on how to teach certain linguistic features and in what order they should be taught (Matsumoto, 2008). Additionally, corpus-based research can be used to contribute to theories about second language acquisition processes (De Knop, 2018). Corpus-based research is not new to the linguistic field, many corpus-based studies have already tried to explain differences between NS and language learner (i.e. NNS) language output using corpus linguistics, such as differences in speaking rate (Hicks, 2010; Thøgersen & Airey, 2011) or rhetorical style (Thøgersen & Airey, 2011).

2.3 Lexical interference

Misuse, overuse and underuse of lexical bundles are all patterns that have been investigated in L2 language output. Some studies attribute these differences between native and non-native lexical bundle use to cross-linguistic influence, or L1 transfer. Cross-linguistic influence refers to interference of other languages that are known during L2 processing or production (Gass, 2013). L1 transfer refers to the effects of the learner's mother tongue on L2 production or reception. Since formulaic sequences are highly language specific, transfer of idiomatic sequences is often difficult. L2 learners are therefore often reluctant in the use of lexical bundles. According to Jarvis (2000), "L1 interference exists when a statistically significant correlation is found between features observed in a learner's L2 performance and their L1 background". Equivalence between L1 and L2 bundles can often be facilitative (Caroll, Conklin & Gyllstad, 2016), causing learners to transfer L1 lexical bundle knowledge of collocational patterns, associational patterns and contextual information to facilitate succesful use of their L2 congruent forms. Research has shown that bundles that overlap entirely led to a greater use in L2 production, whereas bundles that overlapped partially led to avoidance in L2 production (Charteris-Black, 2002). Bundles that showed linguistic overlap but are different in meaning are considered to be most difficult (Caroll et al., 2016), because their use is influenced by negative L1 transfer effects, resulting in misuse of the incongruent bundle type.

According to Jarvis's (2000) framework for the study of L1 interference, there are three indicators of L1 transfer effects:

1. Intra-L1-group homogeneity in the target language (TL) performance
2. Intra-L1 group heterogeneity in learners' TL performance
3. Intra-L1-group congruity between L1 and TL performance

The first indicator, intra-L1-group homogeneity, refers to the notion that a group of TL learners who are from similar L1 backgrounds should exhibit similar TL behaviour towards a certain linguistic feature. Such group behaviour implies L1 transfer (Paquot, 2013). The second indicator, intra-L1 group heterogeneity, refers to the notion that learners from different L1 backgrounds should diverge in their TL behaviour. Intra-L1 group heterogeneity shows that TL behaviour is influenced by L1 knowledge. Different L1 backgrounds should result in different TL behaviour. The third and final indicator of L1 interference (i.e. intra-L1-group congruity) refers to a parallel use of L1 features and their corresponding L2 features. Such behaviour indicates how learners' L1 motivates TL use (Paquot, 2013). In her study on L1 transfer effects on L2 lexical bundle use, Paquot (2013) found collocational, syntactic and functional transfer effects. Moreover, she found a transfer effect of L1 frequency resulting in the assumption that a high L1 bundle frequency results in a repeated use of the L2 translational equivalent or L2 congruent form. However, cross-linguistic transfer effects of lexical bundles can often lead to repeated use of the L2 congruent form (Huang, 2015). Repeated use occurs when congruent bundles (i.e. bundles that are similar in form and meaning) are used in extremely high frequencies in L2 production. But high frequencies are not necessarily better. Huang (2015) investigated highly frequent L2 bundles and found that the number of bundles increases as learners became more proficient in the target language, as had been found by e.g. Lewis (2009). However, the accuracy in which the bundles were used by more proficient leaners remained behind on their frequency (Huang, 2015). An increase in bundle frequency is therefore caused by a repeated, but inaccurate use of the same bundles.

2.4 Bundle frequency variation

Lexical bundle use can differ across corpora in many ways, including frequency (e.g. Conrad & Biber, 2005), disciplinary variation (e.g. Kashiha & Heng, 2014; Wang, 2017), (grammatical) structure (e.g. Biber et al., 2004; Cortes, 2004) and functional variation (e.g. Ädel & Erman, 2012; Biber & Barbieri, 2007; Biber et al., 2004; Wang, 2017). A linguistic topic that is has been widely investigated and is still prone to corpus-based research is frequential variation in the use of lexical bundles between NS and NNS corpora. The frequency data of lexical bundles in a sub-corpus gives insight into the extent to which multi-word sequences are stored as prefabricated chunks (Biber et al., 2004).

Ädel and Erman (2012) studied four-word bundle frequency in NNSs' (L1 Swedish) English academic writings and NS English academic writings. They found that NNSs overall used fewer lexical bundles in academic writing than NSs did. Additionally, Lewis (2009) found that the degree of proficiency correlates significantly with the number of lexical bundles used in second language (L2) production, meaning that L2 learners with a higher L2 proficiency show a more native-like frequency in their L2 lexical bundle use. In a more recent study, Bychkovska and Lee (2017) have found contrasting results. They observed a significantly more frequent use of four-word lexical bundles in NNS (L1 Chinese) university students' academic writings than in NS (English) university students' academic writings. This suggests that NNS rely on the use of prefabricated lexical chunks to greater than NS. Until recently, not much was known about bundle frequency in spoken academic discourse. In 2004, Biber, Conrad and Cortes looked into differences in bundle frequency variation between spoken and written academic registers. They found that NS classroom teaching is structured with lexical bundles to a much greater extent than any other academic register. Additionally, Wang (2017) found a higher proportion of four-word lexical bundles in ELF university lectures than in ELF university seminars. This suggests that academic lectures contain a higher information density, leading to the assumption that lectures look more like academic writing in terms of lexical bundle frequency than other academic registers do (Wang, 2017). Not only was university teaching found to show a higher lexical bundle frequency, out of all university registers NS university teaching also showed the greatest variety in bundle use (Biber et al., 2004). The same results were found for NNSs by e.g. Wang (2017) and Byckovska & Lee (2017). Quite recently, Wang (2017) investigated lexical bundle use in spoken academic lingua franca English (ELFA) in university lectures and seminars from a variety of disciplines. Four-word bundle frequency in the ELFA corpus was compared to existing data on NS four-word bundle frequency patterns from Biber et al. (1999). Results showed that speakers of English as a Lingua Franca (ELF) used significantly more lexical bundles than NSs did (11.000 per million words compared to 5000 per million words). These results are in contrast with previously found frequency differences in academic writing by Ädel and Erman (2012), who found NSs to use more four-word bundles than NNSs.

2.5 Bundle function variation

Formulaic language can be quite diverse in terms of their function and position (Cieślicka, 2015). The three main categories used to distinguish between bundle functions are stance bundles, discourse organizers and referential bundles (Biber et al., 1999). Stance bundles are participant-oriented bundles used to comment on the speaker's personal or impersonal knowledge on the following proposition, e.g. in the bundle '*I don't know what*'. Discourse organizers are text-oriented bundles used for the organisation of the speech, e.g. *'on the other hand'*. Referential bundles are used for identification, reference, specification or quantification, e.g. *'at the end of'*.

Ädel and Erman (2012) have found bundle function distribution to be similar in L1 (English) and L2 (Swedish) university students' writings. Their study showed a similar distribution of bundle functions over the three main bundle function categories (i.e. referential bundles, stance bundles and discourse organizers) for both NSs and NNSs. Contrastingly, Bychkovska and Lee (2017) observed clear differences between NSs and NNSs' writings in bundle function distribution over the three main categories. They observed an unequal distribution in both groups and an underuse of discourse organizers in NNS students' academic writings. In line with earlier findings by Biber at al. (2004), both groups showed a preference for referential bundles but a higher preference for referential bundles was found in NS students' writings. When lexical bundle functions in academic writing are compared to lexical bundle functions in other academic registers, a similar pattern can be observed. Biber et al. (2004) found bundle functions in university classroom teaching to be distributed unequally with a preference for referential and stance bundles. Additionally, Biber and Barbieri (2007) found a similar bundle function distribution in spoken and written academic registers. However, in contrast with written academic registers, Biber and Barbieri (2007) found a preference for stance bundles and discourse organizers over referential bundles. This preference was found to be exceptionally noticeable in classroom teaching (Biber & Barbieri, 2007). A classification of the most frequent four-word bundles in a corpus of spoken ELFA lectures and seminars showed a dominant use of bundles for participant-oriented purposes in university lectures over the use text-oriented (i.e. stance bundles) or real-world (i.e. referential) oriented bundles (Wang, 2017).

Wang (2017) suggests that repetitional bundles (e.g. 'it is it is'), which are said to be a signal of hesitation typical to spoken discourse could be typical to spoken academic ELF, as they are were found in ELF lecturing (Wang 2017) but not in NS spoken academic registers (Biber et al., 2004; Nesi & Basturkmen, 2006). A simple explanation for this suggestion would be that L2 speaking requires more processing time and more time for linguistic planning. Moreover, the language used in academic settings contains a higher informational density (Mauranen, 2012). Repetition allows the speaker to gain time for linguistic planning (Wang, 2017). Another explanation for the difference in frequency and variety of repetitional bundles used between NS and NNS lecturing is that the L2 speaker is trying to assist the listener, who is also an L2 speaker of the language used (Mauranen, 2012). In the latter case, the lecturer is simply trying to "help the speaker to keep track of the information flow and ultimately contribute to successful communication" (Wang, 2017).

2.6 Overview and relevance of the present study

The present study aims to build on the field of NS and NNS differences in lexical bundles use in spoken academic settings by investigating (1) the differences in bundle frequency between NS and NNS academic lecturing, (2) investigating instances of L1 transfer effects on L2 lexical bundle use, and (3) differences in bundle function variation between NS and NNS academic lecturing. Previous research into differences between NS bundle frequency and NNS bundle frequency show inconsistent results, making it difficult to generalize their findings. Not only are previously conducted results about NS and NNS differences in lexical bundle frequency and lexical bundle functions contrasting (Ädel & Erman, 2012; Bychovska & Lee, 2017), results from previous studies are often based on research that exclusively used data from written academic registers. Due to the scarcity of spoken academic data, only few studies have evaluated lexical bundle use in spoken academic registers. The few studies that did report on lexical bundle use in spoken academic registers make no attempt to draw conclusions about NS and NNS differences in lexical bundle functions and frequencies in academic lecturing. These studies either solely draw on NS data from university classroom teaching (Kashiha & Heng, 2014) or academic discourse (e.g. Biber & Conrad, 1999; Conrad et al., 2005), or they focus exclusively on NS and NNS differences in bundle use between different academic registers (e.g. Biber et al., 2004; Biber & Barbieri, 2007), making it impossible to draw conclusions about NS and NNS differences in academic lecturing. Wang (2017) is the only study that was found to investigate NNS lexical bundle use in academic lecturing. However, Wang's (2017) study was designed to compare lexical bundle use in academic lecturing to their use academic seminars, drawing exclusively on ELF data. There is not one study that evaluates the differences in lexical bundle use in academic lecturing between two languages produced by the same subject. Moreover, no literature can be found about L1 Dutch transfer effects on L2 English lexical bundle use by the same subject. It can be concluded that more research into bundle frequency differences between NS and NNS is needed. This research aims to fill this gap in the literature by investigating frequential and functional differences in L1 and L2 spoken academic lecturing produced by the same subject. Additionally, a qualitative study will be conducted to investigate interpersonal L1 transfer effects. Existing parallel recordings of L1 Dutch and L2 English spoken academic lectures performed by the same subject at a university in the Netherlands have been made available for this study. Considering the scarcity of spoken academic data for corpus-based research, this data allows for a unique opportunity to investigate NS and NNS differences in academic lectures. Especially considering the fact that both data sets (i.e. a set of NS lectures and NNS lectures) derive from the same subject.

3. Methodology

3.1 Data

The corpus that has been made available for this study is the Nijmegen English Medium of Instruction Corpus (NEMIC). The NEMIC corpus contains spoken data from seven parallel Dutch-English university lectures taught by the same lecturer. The subject is a native speaker of Dutch (L1 = Dutch) and a second language speaker of English (L2 = English) at CEFR C2 level (council of Europe, 2001). All lectures have been given in both the subject's L1 and their L2. This means that the NEMIC corpus is divided into two sub-corpora. The first sub-corpus (i.e. NEMIC_DUTCH) contains video recordings of seven lectures given in the subject's L1 Dutch. The second sub-corpus (i.e. NEMIC_ENGLISH) contains seven video recordings of parallel lectures given in the subject's L2 English. The content of the Dutch and English lectures is exactly the same. Both courses were part of a bachelor's programme in marketing communication studies at Radboud University. Both courses consisted of seven weekly two-hour lectures. There were no additional seminars or tutorials that were part of either one of the courses. Three parallel lectures taken from the main NEMIC corpus, i.e. three Dutch lectures and the corresponding English lectures, were used for this study.

Table 1. Overview of the data used

	NEMIC_DUTCH	**NEMIC_ENGLISH**
	No. of words	**No. of words**
Lecture 1	12 208	10 489
Lecture 3	8 330	8 274
Lecture 4	5 284	2 885
Total	**25 822**	**21 648**

3.4 Transcripts

The lectures were all transcribed using ELAN version 5.5 for Windows (2019). All lectures within a corpus were transcribed separately. A lecture was first divided into segments of 10.000 milliseconds using the segmentation mode in ELAN (2019). Each segment was then transcribed using the transcription mode. Videos and audio fragments that were shown in a lecture were not transcribed as this study aims to exclusively investigate language produced by the lecturer. Student answers were excluded from the transcripts for the same reason.

Filled pauses (e.g. *uhm*) were not transcribed as they could disturb the N-gram search at the bundle identification stage of this study, e.g. when a transcription of a filled pause occurs between the first two and the last two words of a four-word bundle, causing the concordance tool to miss the bundle in an identification or frequency search. Catch phrases that occurred in the middle of an expression were excluded from the transcriptions in order to limit disturbance of the N-gram search and because they are specific to the individual speaker, making it impossible to generalize findings for other speakers. Transcriptions of stuttering were excluded for the same reason. Shortened word combinations such as 'won't' and 'it's' were transcribed as separate words (i.e. *will not* and *it is*) in order to make sure that such combinations would be counted as separate words when identifying four-word bundles in Antconc (Anthony, 2019). Abbreviations that were pronounced as individual letters were transcribed as acronyms (e.g. *PR*) because the number of words in such cases was under four or because their low occurrence was unlikely to affect this study.

3.5 N-gram search

Lexical bundles are identified by measuring the most frequently recurring series of words in a sub-corpus (Biber & Bibieri, 2007) and usually consist of three to four words (Wang, 2017). The cut-off point for lexical bundle frequency varies across studies. Cut-off points normally range from 10 to 40 times per million words (Wang, 2017). The dispersion criterion that is often used in corpus-based bundle frequency studies refers to the number of (spoken) texts a bundle needs to occur in in order to reduce the effect of individual speaker preferences (Wang, 2017). Since this study contains a considerably smaller data set compared to previous studies and because in contrast with other studies there cannot be any influence of speaker bias, it was decided to revise the generally accepted cut-off points for frequency and dispersion. Previous studies used corpora that were on average at least ten times bigger than the corpus used in this study. For instance, Wang (2017) used a 200,000-word sub-corpus resulting in maximum frequency levels between 18 and 43 occurrences per million words (PMW) for four-word bundles. The same normalization criterion (i.e. an occurrence between 10 and 40 times per million words) would lead to frequency levels between 39 and 658 in NEMIC_DUTCH and levels between 46 and 508 in NEMIC_ENGLISH. This would mean that even four-word bundles that would occur only once in a sub-corpus would meet the frequency cut-off point used in previous studies. Use of the generally accepted frequency cut-off point would furthermore lead to a total outcome of 25.145 four-word bundles in NEMIC_DUTCH and 20.634 in NEMIC_ENGLISH. That is because even four-word sequences that only occurred once would meet the frequency cut-off point, resulting in the fact that every four-word sequence would be identified as a four-word lexical bundle. Since the primary focus of this study is to compare NNS data to NS data produced by the same subject, it has been decided to normalise the frequency of the four-word sequences resulting from the N-Gram search so that both corpora (i.e. NEMIC_DUTCH and NEMIC_ENGLISH) of different sizes could be compared. Because both corpora contain between 20.000 and 30.000 words, the normalisation factor was set at 25.000. This means that the normalised frequency of each bundle that resulted from the N-Gram search was calculated per 25.000 words, allowing for a fair frequential comparison between the two corpora. The generally accepted cut-off point for dispersion is that a bundle needs to occur in at least 3 to 10 percent of the data used (Hyland, 2008a). Since both sub-corpora in this study contained only three lectures each, maintaining a minimum dispersion criterion of 3 to 10 percent would result in the fact that bundles that occur in only one of the three lectures in a sub-corpus would meet the cut-off point for dispersion as an occurrence in one out of three lectures equals an occurrence in 33% percent of the data in a sub-corpus. It has therefore been decided not to set a dispersion cut-off point for this study. All four-word lexical bundles that meet the frequency cut-off point are therefore considered for analysis, no matter the amount of texts they occur in. In order to minimalize topic-specific four-word bundles, the minimum frequency cut-off point was set at four to make sure that a bundle occurred twice in at least one lecture in a sub-corpus. This means that only four-word sequences that showed a minimum occurrence of four times in a sub-corpus would be identified as a lexical bundle.

For this study, four-word bundles in each sub-corpus were identified using the concordance tool Antconc (Anthony, 2019). Four-word bundles were considered in order to make sure that the results from this study would be comparable to existing literature. Moreover, three-word bundles are often subsumed in four-word bundles (Ädel & Erman, 2011) and the latter are "within a more manageable size for manual categorization and concordance checks" (Chen & Baker, 2010). Antconc was chosen because it has word and keyword frequency generators and tools for cluster/N-gram analyses. Previous studies have shown Antconc to be an effective tool for the identification of multi-word bundles that meet a certain frequency cut-off point (e.g. Bychkovska & Lee, 2017). Since the aim of this study is to investigate differences between the two corpora (i.e. NEMIC_DUTCH and NEMIC-ENGLISH), the two corpora were analysed separately. Four-word bundles from each were retrieved using the clusters/N-grams tool. An N-gram search makes it possible to find common bundles within a sub-corpus without needing to specify a search term. The N-gram size was set at four words to make sure that only four-word bundles would result from the N-Gram search. Since no dispersion criterion was set, no minimum range was set for the N-Gram search. In order to identify all four-word bundles in each sub-corpus that meet the criteria set above, the minimum frequency for the N-gram searches was set at four.

The results from the N-Gram search were checked manually for topic-specific bundles, context-dependent bundles and bundle overlap. Topic-specific and context-dependent bundles such as the bundle *'is the marketing* communication', are bundles that are dependent on context and topic (Bychkovska & Lee, 2017) and are therefore not representative of the speaker's L2 bundle vocabulary. The N-Gram search was also checked for overlap that could possibly inflate the number of N-Gram types. Bundles that overlapped so that one is subsumed within the other were combined. In such cases, lower-frequency bundles were combined into the higher frequency bundle as to avoid inflation of quantitative results (Bychkovska & Lee, 2017). For example, the bundles *'If you want to'* and *'Do you want to'* that resulted from the N-Gram search on NEMIC_ENGLISH had an individual normalized occurrence of 9 and 6 in NEMIC_ENGLISH. Due to overlap these bundles were combined into the most frequent bundle among the two:*'If you want to'*. The total frequency of the bundle *'If you want to'* was calculated by adding up the normalized frequencies of *'If you want to'* and *'Do you want to'*. The total number of occurrences of the N-Gram *'If you want to'* was therefore 15 times in NEMIC_ENGLISH.

3.5 Bundle frequency measurements

The total number of lexical bundle types that met the criteria explained in section 3.4 (i.e. N-Gram types) were listed separately for each sub-corpus as well as the frequency of occurrence of each individual four-word bundle within a sub-corpus (i.e. N-Gram tokens). The total number of N-Gram tokens in a sub-corpus was calculated by adding up the normalized frequencies of all identified bundles in each sub-corpus. The calculations are provided in Table 2 below. In order to answer the first research question *'is lexical bundle use in Dutch NS spoken academic lectures significantly more frequent than in English NNS spoken academic lectures taught by the same subject?'*, the difference between the total number of N-Gram tokens between NEMIC_DUTCH and NEMIC_ENGLISH was calculated. The results were tested for significance using IBM SPSS (2017). The test statistics found the frequencies in which the identified bundles occurred to be distributed unequally. A normal distribution should show skewness and kurtosis levels between -1 and 1, whereas the test statistics showed the frequency distribution to have a skewness value of 1,61 and a kurtosis value of 1,45, causing a positively skewed, leptokurtic distribution. Since the data was found to be distributed unequally, it was decided to deviate from using an independent samples t-test, which requires the data to be distributed normally. A non-parametric test was used instead. Since the data was categorical and because the two sub-corpora contained data derived from the same subject, a Wilcoxon signed-rank test was used to test whether the frequencies in which the four-word bundles occurred in each sub-corpus were statistically different. The experimental or dependent variable bundle frequency was subject to one independent variable, i.e. sub-corpus (NEMIC_DUTCH or NEMIC_ENGLISH).

Table 2. N-Gram token calculations

NEMIC_DUTCH		**NEMIC_ENGLISH**	
N-GRAM type	**N-GRAM tokens**	**N-GRAM type**	**N-GRAM tokens**
Aan de ene kant	21	It is it is	13
Ik weet niet of	30	Don't know if you	28
Aan de andere kant	12	If you want to	27
En dit is een	6	That is what is	15
Een heel belangrijk onderdeel	5	What is meant by	14
Is in ieder geval	14	I would like to	7
Video video video video	5	Very important part of	30
Als het goed is	4	It is not the	28
Dat betekent niet dat	8	Of course it is	12
Dat is dat is	4	On the basis of	7
Dus het is niet	4	You have to take	12
Een voorbeeld van een	4	It is much more	6
En dan kun je	4	What do you think	12
Even naar me toe	4	You go to the	6
Ik denk niet dat	4	Here you can see	5
In het hoofd van	4	How many people are	5
Je hebt natuurlijk ook	4	In a in a	5

Op een gegeven moment	4	Is a way to	5
		Is used a lot	5
		Not the only part	5
		Put your name on	5
		The idea is that	10
		There are a lot	5
		We will talk about	5
		You do not have	5
Total	**141**		**277**

3.6 Bundle function differences

3.6.1 Taxonomies for bundle function identification

In order to investigate functional variation between the four-word bundles analysed in NEMIC_DUTCH and the four-word bundles that were analysed in NEMIC_ENGLISH, all bundles were classified according to their function. This requires a functional classification of all target bundles that were identified in each sub-corpus. Biber (Biber et al., 1999) was among the first to study functional differences in lexical bundle use between two corpor. Biber's functional taxonomy (Biber et al., 1999; Biber at al., 2004) is therefore a generally accepted taxonomy for bundle function identification in the field of corpus linguistics. This taxonomy was originally developed to classify the discourse functions in conversation and academic prose (Biber et al., 2004) and distinguishes between three major functions of lexical bundles: (1) stance expressions, (2) discourse organizers and (3) referential expressions. Stance expressions provide information about the proposition that immediately follows the stance expression (Biber et al., 2004). Stance bundles can be epistemic or attitudinal/modality bundles. Epistemic stance bundles are used to comment on the speaker's personal or impersonal knowledge of the following proposition. For instance, *I don't know what* in the expression 'I don't know what time it is' expresses uncertainty, whereas *are more likely to* in the expression 'boys are more likely to be aggressive than girls' shows certainty. Attitudinal/modality bundles are used to express speaker attitudes towards the action or event that follows in the proposition (Biber et al., 2004). Attitudinal/modality stance bundles are divided into four subcategories. The first subcategory of attitudinal/modality stance expressions are desire bundles that show the speakers' personal expression of stance. An example of a desire bundle is *I don't want to* in the expression 'I don't want to walk to school today'. The second subcategory of stance bundles express obligation or directives. These bundles have a second person pronoun as their subject, which differentiates them from personal stance bundles that have a first-person subject (Biber et al., 2004). An example would be *you have to do* in the expression 'all you have to do is work on it'. The third subcategory includes intention and prediction bundles such as *is going to be*. The last category classifies bundles that express ability, such as *to be able to*. Discourse organizers are bundles that are either used to introduce a topic and to put focus on a topic or to elaborate or clarify. Introduction/focus bundles are often used by a speaker to announce a new topic. An example of this bundle type would be *want to talk about*. An example of a discourse organizing bundle for elaboration would be *has to do* with, whereas a discourse organizing bundle for clarification is often used to indicate a comparison or contrast, as in *as well as* and *on the other hand* (Biber et al., 2004). Referential bundles are used (a) for identification/focus, (b) to indicate imprecision, (c) to specify a certain aspect, or (d) to refer to time, place or text. Identification bundles are especially common in classroom teaching (Biber et al., 2004), for example in the bundle *those of you who*. Focus bundles can be used in classroom teaching to introduce a discussion topic (Biber et al., 2004). Imprecision bundles are used by the speaker to refer to imprecise references, such as *something like that*. Specifying referential bundles are used to identify specific characteristics of the head noun in the following proposition. Such bundles can be used to specify quantities, topics, size or form, abstract characteristics or logical relationships. The last subcategory of referential bundles includes bundles that refer to time, place or text. Text-deixis bundles only occur in written text.

After Biber et al. (1999), a few others have modified Biber's framework for bundle function identification, including Hyland (2008a) and Cortes (2004). However, all frameworks were initially developed to identify bundle functions across a variety of academic genres or with the aim to identify bundle functions in data that exclusively concerned academic writings. In order to specify the original taxonomy for the identification of bundle functions in spoken academic data, Wang (2017) created a framework for lexical bundle function identification based on Biber's taxonomy (Biber et al., 1999; Biber et al., 2004) as well as Cortes (2004) and Hyland (2008a). Figure 1 below shows Wang's framework for bundle function identification. Wang (2017) designed this framework specifically to investigate bundle function use in ELF lecturing, whereas previously developed frameworks were al developed to identify bundle function differences in academic discourse and/or academic prose. It was therefore decided that Wang's framework for bundle function identification would be most suitable for a functional classification of four-word bundles in the present study. That is because all data used in this study are transcripts of spoken academic lecturing, similar to the data used by Wang (2017) and exactly the type of data that the framework was designed for. Moreover, Wang (2017) designed the framework to identify four-word bundles, as is the case in the present study.

Real-world oriented: referring to real-world properties.

i. Time/place/personal reference, e.g. *at the end of, the rest of Europe*
ii. Identification/descriptive attribute, e.g. *the first half of, the name of the*
iii. Quantity specification, e.g. *a lot of er, a little bit of*

Text oriented: signalling the organisation of the speech and the elements of an argument.

i. Transition signals: establishing logical links between elements, e.g. *on the other hand, so that we can*
ii. Framing signals: situating arguments by specifying limiting conditions, e.g. *in the case of, on the basis of*

Participant oriented: focusing on the interaction between the speaker and the listener.

i. Stance markers: expressing epistemic stance, e.g. *er it is not*, or the speaker's attitudinal/modality stance, e.g. *I don't know if, it has to be*
ii. Engagement signals: addressing the hearer directly, often involving fragments of questions, e.g. *if you want to, what do you think*, or expressing agreement/ disagreement, e.g. *no no no no, yeah mhm hm yeah*
iii. Procedure signals: indicating actions and the organisation of the lecture/seminar, e.g. *I would like to, you are going to*
iv. Fillers: meaningless repetition of single words or sounds, e.g. *the the the the, of the of the*

Figure 1 Framework for lexical bundle function identification (Wang, 2017)

Even though Wang's (2017) framework is specified for the identification of four-word lexical bundles in spoken university lectures, the bundle classification that results from this framework will still be comparable to literature reporting on bundle function identification that has been conducted through the use of earlier frameworks (e.g. Biber et al., 1999 or Hyland, 2008a). That is because previous studies have mainly reported on the distinction between three main categories, which remain the same but under different names. Real-world bundles (Wang, 2017) correspond with referential bundles (Biber et al., 1999), text-oriented bundles (Wang, 2017) correspond to discourse organizers (Biber et al., 1999), and participant-oriented bundles (Wang, 2017) correspond with stance bundles (Biber et al., 1999). It is the sub-categorization of each category that is specified for spoken academic data. Since this study aims to investigate the differences in bundle distribution over the three main categories, the results will be comparable to literature reporting on bundle function variation in written (academic) registers.

3.6.2 Analytical steps for functional variation between corpora

The contextual functions of all four-word bundles that were identified in each sub-corpus were classified according to Wang's (2017) framework of functions. The function of lexical bundles is dependent on context (Hyland, 2008a). It is therefore possible that bundles fell into more than one category in the framework (Wang, 2017). In such cases, the most prototypical bundle function was chosen, and the bundle was classified accordingly. The identified bundles in both NEMIC_DUTCH as well as NEMIC_ENGLISH were classified separately. After the functions of all lexical bundles from each sub-corpus were identified, the number of four-word bundle tokens in each of the three main functional categories (i.e. real-world oriented, text-oriented and participant-oriented) was calculated by adding up the number of bundle tokens of each bundle type that fell into a functional category. The difference between the number of N-Gram tokens in each of the three main functional categories was measured using IBM SPSS (2017). The frequency distribution was found to be unequal in all three functional categories, showing skewness and kurtosis values higher than 1. The statistical difference between the number of bundle tokens in NEMIC_DUTCH and NEMIC_ENGLISH was therefore tested using a Wilcoxon signed rank test. The experimental or dependent variable bundle frequency was subject to one independent variable, which was the sub-corpus (i.e. NEMIC_DUTCH or NEMIC_ENGLISH).

3.7 Qualitative analysis

Since qualitative analyses on L2 bundle performance are sometimes considered to be more reliable and convincing than quantitative comparisons (Huang, 2015), the statistical results were supported by a qualitative analysis. Four aspects of the data and/or results that could potentially explain the obtained quantitative results and/or instances of L1 transfer were investigated: bundle variety, bundle functional variety in bundle use, bundle overlap, and the most frequent bundles used in each sub-corpus.

3.7.1 Bundle variety

First, bundle variety was investigated by comparing the type/token ratio in NEMIC_DUTCH to the type/token ratio in NEMIC_ENGLISH. The type/token ratio is used to illustrate the variety of bundle use and is suitable for (sub-)corpora that are comparable in length (Huang, 2015), as is the case in this study. The type refers to the number of different four-word bundles that was obtained in a sub-corpus, while the tokens refer to the number of occurrences of each individual bundle (Huang, 2015). The type-token ratio therefore indicates to what extent the same bundles have repeatedly been used in a sub-corpus. The type/token ratio was calculated separately for each sub-corpus by dividing the number of bundle types by the number of tokens in each sub-corpus. A higher type/token ratio indicates a more varied bundle use, i.e. that a larger set of bundles was used.

3.7.2 Functional variety in bundle use

In addition to the quantitative functional distribution that was computed, the lexical variation of the bundles used in each functional category was investigated as well. For both NEMIC_DUTCH and NEMIC_ENGLISH, the type/token ratio was calculated for each functional category. The difference between the type/token ratio in NEMIC_DUTCH and NEMIC_ENGLISH was analysed for all three functional categories. A higher type/token ratio indicates a more varied use of lexical bundles within a category.

3.7.3 Bundle overlap

In order to explain L1 transfer effects on L2 bundle use, a qualitative approach is required (Paquot, 2013). Overlapping bundles were identified in two steps. First, the translational equivalents of all bundles in NEMIC_DUTCH and NEMIC_ENGLISH were identified in context. Second, patterns of congruent forms were compared and described. Topical bias was ruled out during N-Gram selection (see section 3.5).

3.7.4 Most frequent bundles

The bundles that occurred most frequently within a sub-corpus (i.e. the bundle types with the most bundle tokens) were listed separately for NEMIC_DUTCH and NEMIC_ENGLISH. A comparison between the most frequent bundles in NEMIC_DUTCH and NEMIC_ENGLISH was conducted in order to explain cross-linguistic differences and to identify instances of L1 transfer effects in lexical bundle use between the subject's L1 Dutch and L2 English.

3.7.5 Functional differences between participant-oriented bundles used

The qualitative analysis resulted in the finding that participant-oriented bundles were preferred in both NEMIC_DUTCH and NEMIC_ENGLISH, but that the preference for this bundle function was significantly higher in L2 English. In an attempt to explain this result, it was decided to analyse the functional difference between participant-oriented bundles used in NEMIC_DUTCH and NEMIC_ENGLISH. In order to do so, the distribution of the bundles that were categorised as participant-oriented over the four functional sub-categories (i.e. stance markers, engagement signals, procedure signals and fillers) was analysed.

4. Results

4.1 Quantitative analysis

4.1.1 Lexical bundle identification

After manual calculations of the normalised frequencies of the four-word sequences that resulted from the N-Gram search on NEMIC_DUTCH and NEMIC_ENGLISH in Antconc (Anthony, 2017), bundles that that did not meet the normalised frequency cut-off point of four were eliminated. The remaining bundles were checked manually. Four-word bundles that were misidentified, topic-specific, context-dependent or overlapping were eliminated or subsumed. Out of the 28 N-Gram types that resulted from the N-Gram search on NEMIC_DUTCH, two bundles were eliminated, and several bundles were subsumed. The bundles *'Die vier P's'* and *Point of purchase communication'* were identified as context and discipline-specific and therefore left out of the analysis. The bundles *'Ik weet niet of', 'Weet niet of jullie'* and *'Weet niet of het'* showed a considerable amount of overlap and were therefore combined into the most frequent bundle with the widest range: *'Ik weet niet of'*. For the same reason, the bundles *'In ieder geval een'*, *'Is in ieder geval'* and *'Of in ieder geval'* were combined into *'In ieder geval een*'. The bundles *'Aan de andere kant'* and *'En aan de andere'* were subsumed and combined into *'In ieder geval een'*. Finally, the bundles *'Dat betekent niet dat'* and *'Maar dat betekent niet'* were combined into *'Dat betekent niet dat'.* After manual checking, 18 four-word bundles in NEMIC_ENGLISH were submitted for further analysis. In NEMIC_ENGLISH, a total of 55 N-Gram types resulted from the N-Gram search in Antconc (Anthony, 2017). 7 Of these bundles were found to be discipline and context specific. The bundles *'Of your target group', 'Is this marketing communication', 'The marketing communication's', 'Communication but is a', 'Marketing communication's objectives are', 'Marketing communication who says'* and *'The product that is'* were therefore excluded from further analysis. A number of bundles that were partly overlapping have furthermore been subsumed into the most frequent bundle among them with the widest range. The bundles *'Do not know if', 'I don't know'* and *'Not know if you'* were combined into *'Do not know if'*. The bundles *'If you want to', 'Do you want to', 'You want to say'* and *'If you have a'* showed a considerable amount of overlap and were therefore subsumed. For the same reason, the bundles '*That is what* is' and *'that is that is'* were combined into *That is what* is', the bundles *'What is meant by'* and *'is what is meant'* were combined into *'What is meant by'*, the bundles '*Is a very important', 'Very important part of', 'It is a very', 'A very* *important part’* and *‘Important part of the’* were combined into ‘*Is a very important’*, the bundles *‘It is not the’, ‘Is not the only’, ‘But it is not’, ‘It is not just’* and *‘So it is not’* were combined into *‘It is not the’,* the bundles *‘Of course it is’, ‘Of course in the’* and ‘*Of course you can’* were combined into ‘*Of course it is’*, the bundles *‘You have to take’, ‘You have to do’* and *‘Do not have to’* were combined into *‘You have to take’,* the bundles *‘What do you think’, ‘ When you think of’* and *‘You think of a’* were combined into *‘What do you think’* and the bundles ‘*That is the idea’* and *‘The idea is that’* were combined into the former. After manual checking, 25 four-word bundles d in NEMIC_ENGLISH were submitted for further analysis. An overview of the bundles in NEMIC_DUTCH and NEMIC_ENGLISH that were submitted for further analysis is provided in table 3 below.

Table 3. Identified four-word lexical bundles in NEMIC_DUTCH and NEMIC_ENGLISH

NEMIC_DUTCH	**NEMIC_ENGLISH**
Aan de ene kant	It is it is
Ik weet niet of	Do not know if
Aan de andere kant	If you want to
En dit is een	That is what is
Een heel belangrijk onderdeel	What is meant by
In ieder geval een	I would like to
Video video video video	Very important part of
Als het goed is	It is not the
Dat betekent niet dat	Of course it is
Dat is dat is	On the basis of
Dus het is niet	You have to take
Een voorbeeld van een	It is much more
En dan kun je	What do you think
Even naar me toe	You go to the
Ik denk niet dat	Here you can see
In het hoofd van	How many people are
Je hebt natuurlijk ook	In a in a
Op een gegeven moment	Is a way to
	Is used a lot
	Not the only part

	Put your name on
	The idea is that
	There are a lot
	We will talk about
	You do not have

4.1.2 Frequential differences in bundle use

In order to investigate whether the subject used significantly more lexical bundles in NS Dutch lecturing than in NNS English lecturing, the total number N-Gram tokens in NEMIC_DUTCH was compared to the total number of N-Gram tokens in NEMIC_ENGLISH. The number of N-Gram tokens in each sub-corpus was calculated manually by adding up the normalised frequencies of all identified bundles in each sub-corpus. Frequencies of overlapping four-word bundles that were subsumed under the most frequent bundle among them were combined. Table 4 shows an overview of the normalised type and token frequencies of four-word bundles in NEMIC_DUTCH and NEMIC_ENGLISH. In NEMIC_DUTCH, 141 N-Gram tokens were identified. In NEMIC_ENGLISH, 277 N-Gram tokens were identified. The difference between the number of N-Gram tokens used in NEMIC_DUTCH and NEMIC_ENGLISH is 136. This difference was tested for significance using IBM SPSS (IBM, 2017). The difference between the number of four-word lexical bundle tokens used in NEMIC_DUTCH and the number four-word lexical bundle tokens used in NEMIC_ENGLISH was found to be statistically significant. The subject was found to use significantly more four-word lexical bundle tokens in NEMIC_ENGLISH, *Mdn* = 7 (*IQR* 5 – 13,5) than in NEMIC_DUTCH, *Mdn* = 4 (*IQR* 4 – 9), $p < 0.001$. The mean frequency of four-word bundles in NEMIC_ENGLISH was 11, whereas the mean frequency of four-word bundles in NEMIC_DUTCH was 8.

Table 4. Overview of N-Gram types and tokens in NEMIC_DUTCH and NEMIC_ENGLISH

Sub-corpus	**N-Gram types**	**N-Gram tokens**
NEMIC-DUTCH	18	141
NEMIC-ENGLISH	25	277

4.1.3 Functional differences in bundle use

In order to determine functional variation between the four-word lexical bundles used in NEMIC_DUTCH and the four-word lexical bundles used in NEMIC_ENGLISH, the distribution of the bundles in NEMIC_DUTCH and NEMIC_ENGLISH across the functional categories was compared. Table 5 below shows the distribution of the lexical bundles over the main functional categories in NEMIC_DUTCH and NEMIC_ENGLISH.

Table 5. Distribution of the lexical bundles over the main functional categories in NEMIC_DUTCH and NEMIC_ENGLISH

Functional category	**Bundle types**				**Bundle tokens**			
	DUTCH		**ENGLISH**		**DUTCH**		**ENGLISH**	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Real-world oriented	4	22%	3	12%	19	13%	41	15%
Text oriented	6	33%	5	20%	53	38%	64	23%
Participant oriented	8	45%	17	68%	69	49%	172	62%
Total	**18**	**100%**	**25**	**100%**	**141**	**100%**	**277**	**100%**

In NEMIC_DUTCH, 22% of all bundle types were identified to function as RWO bundles, whereas in NEMIC_ENGLISH only 12% of bundle types was identified to function as RWO bundles. The same pattern is noticeable in TO bundles. 33% of the bundle types were identified to function as TO bundles in NEMIC_DUTCH, whereas only 20% of the bundle types were identified to function as TO in NEMIC_ENGLISH. A reversed distribution can be observed for PO bundles. The subject used considerably more PO bundle functions in NEMIC_ENGLISH (i.e. 68%) compared to NEMIC_DUTCH (45%). Overall, the subject shows a clear preference for four-word lexical bundles that function as participant-oriented bundles in both NEMIC_DUTCH and NEMIC_ENGLISH, followed by TO bundles. However, with 13% of the bundle tokens in NEMIC_DUTCH to function as RWO, 38% as TO and 49% as PO, the bundle tokens in NEMIC_DUTCH seem to be distributed across the functional categories more evenly than in NEMIC_ENGLISH, in which 15% of the bundles was identified to function as RWO, 23% as TO and 62% as PO. When looking at the functional distribution more closely, a significant difference between the amount of bundle tokens can be observed in two out of three functional categories. An overview of the distribution is provided in figure 2 below.

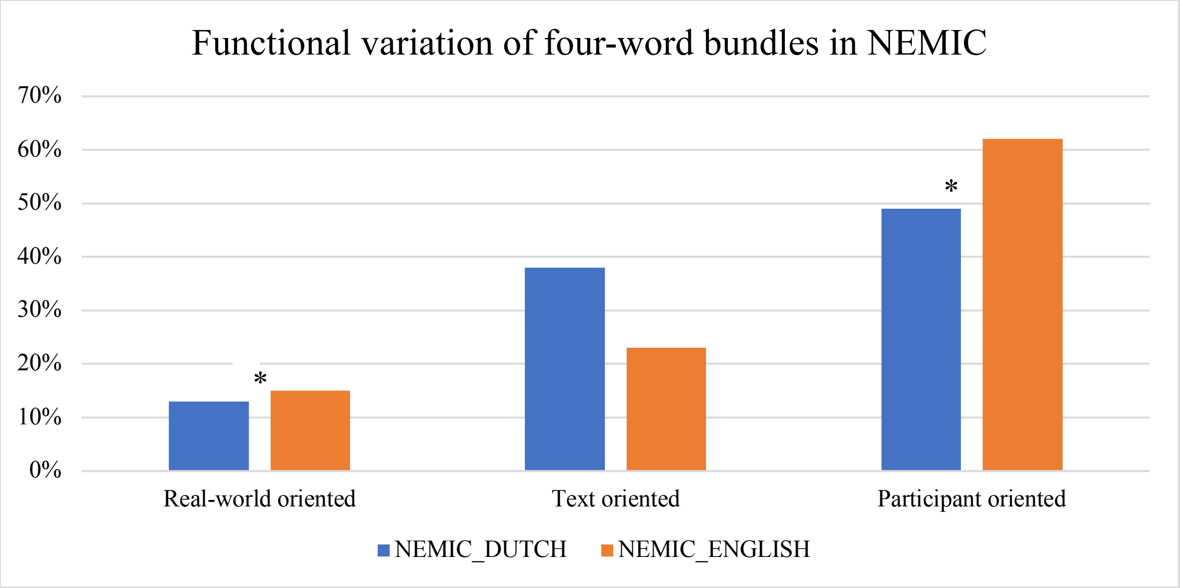


Figure 2. Functional distribution of four-word bundles in NEMIC_DUTCH and NEMIC_ENGLISH

The difference between the number of real-world tokens used in NEMIC_DUTCH and NEMIC_ENGLISH was found to be statistically significant. The subject used significantly more real-world oriented bundles in NEMIC_ENGLISH, *Mdn* = 2 (*IQR* 2 - 2) than in NEMIC_DUTCH, *Mdn* = 1 (*IQR* 1 - 1), *p* = 0.017. A similar effect was found in the number of participant-oriented bundle tokens. The difference between the number of participant-oriented bundle tokens in NEMIC_DUTCH and NEMIC_ENGLISH was found to be statistically significant. The subject used significantly more participant-oriented four-word bundles in NEMIC_ENGLISH, *Mdn* = 2 (*IQR* 2 -2) than in NEMIC_DUTCH, *Mdn* = 0 (*IQR* 0 - 4), *p* = < 0.01. However, the difference between the number of text-oriented bundle tokens used in NEMIC_DUTCH and NEMIC_ENGLISH was not found to be statistically significant.

4.2 Qualitative analysis

4.2.1 Bundle variety

Bundle variety was measured by calculating the type/token ratios in both NEMIC_DUTCH and NEMIC_ENGLISH. The type/token ratio in NEMIC_DUTCH was found to be approximately 0.13, whereas the type/token ratio in NEMIC_ENGLISH was approximately 0.09. Even though the difference between the type/token ratio in NEMIC_DUTCH and the type/token ratio in NEMIC_ENGLISH is only 0,04, the subject was found to show a slightly greater variety in bundle use. This indicates that the subject has used less repetition of the same bundles in NS lecturing compared to NNS lecturing.

4.2.2 Functional variety in bundle use

The functional variety between the four-word lexical bundles used in NEMIC_DUTCH and NEMIC_ENGLISH was analyzed by comparing the type/token ratios between the sub-corpora for each functional category. The results are provided in table 6 below. For all three functional categories, the type/token ratio was found to be larger in NEMIC_DUTCH than in NEMIC_ENGLISH.

Table 6. Overview of type/token ratios in each functional category

	Type-token ratio		
	RWO	TO	PO
NEMIC_DUTCH	0.21	0.11	0.12
NEMIC_ENGLISH	0.07	0.08	0.10

4.2.3 Bundle overlap

A total of 27% of the bundles used in both sub-corpora are shared between NEMIC_DUTCH and NEMIC_ENGLISH. This comes down to three bundles that were found to have equivalent translations in NEMIC_DUTCH and NEMIC_ENGLISH. An overview of the overlapping bundles in NEMIC_DUTCH and NEMIC_ENGLISH is provided in table 7 below. The overlapping bundles account for 30% of the total number of bundles used in NEMIC_DUTCH and their translational equivalents account for 26% of all bundles used in NEMIC_ENGLISH. The three overlapping bundles in NEMIC_ENGLISH show normalised frequencies of 30, 28 and 15 times per 25 thousand words, which seems to be an equal distribution without extreme outliers. The percentage of overlapping bundles in NEMIC_DUTCH however seem to be affected by one particular outlier. The bundles *'Een heel belangrijk onderdeel'* and *'En dit is een'* show relatively low frequencies (i.e 5 and 6 times per 25 thousand words) compared to the bundle *'ik weet niet of'*, which occurred 30 times per 25 million words. The latter is therefore responsible for 21% of the bundle tokens used in NEMIC_DUTCH. The overlapping bundles between the two sub-corpora were all identified to function as either RWO or PO bundles. None of the overlapping bundles were classified as TO bundles. The lack of overlap in TO bundles reflects the quantitative finding that the subject overall used fewer TO bundles in NNS speech than in NS speech.

Table 7. Overview of bundles shared between NEMIC_DUTCH and NEMIC_ENGLISH

NEMIC_DUTCH		**NEMIC_ENGLISH**	
Type	**Tokens**	**Type**	**Tokens**
Een heel belangrijk onderdeel	5	Very important part of	30
Ik weet niet of	30	Do not know if	28
En dit is een	6	That is what is	15

4.3.3 Most frequent bundle types

Table 7 shows the most frequent bundle types in NEMIC_DUTCH and NEMIC_ENGLISH. All bundles that accounted for 10% of the total number of bundle tokens used in a sub-corpus or more were considered to be highly frequent. In NEMIC_DUTCH, this meant that bundles that occurred 14 times or more were considered high frequent. In NEMIC_ENGLISH, bundles that occurred 28 times or more were considered high frequent. Three bundles were identified to be highly frequent in both NEMIC_DUTCH and NEMIC_ENGLISH. One of the top three bundles in both sub-corpora overlapped, this was the bundle *'Ik weet niet of'* and the translational equivalent *'Do not know if'*.

Table 7: Most frequent bundles in NEMIC_DUTCH and NEMIC_ENGLISH

NEMIC_DUTCH		**NEMIC_ENGLISH**	
Type	**Tokens**	**Type**	**Tokens**
Ik weet niet of	30	Very important part of	30
Aan de ene kant	21	Do not know if	28
Is in ieder geval	14	It is not the	28

The top three bundles in NEMIC_DUTCH (i.e. *'aan de ene kant', 'ik weet niet of'* and *'Is in ieder geval'*) account for 46% of the bundles used in NS lectures, whereas the top three bundles in NEMIC_ENGLISH account for only 31% of the bundles used in NNS lectures. The subject therefore used 15% fewer bundles in NNS lecturing compared to NS lecturing.

4.3.4 Functional differences between participant-oriented bundles used

The results in Figure 3 below show the distribution of participant-oriented bundles in NEMIC_DUTCH and NEMIC_ENGLISH differs. The results show a dominant use of participant-oriented bundles to express stance in NEMIC_DUTCH, followed by an equal number of participant-oriented bundles used as fillers or engagement signals. In NEMIC_ENGLISH, most participant-oriented bundles were used as either engagement signals (35%) or procedure signals (29%), whereas the use of participant-oriented bundles to express stance or as fillers was least favourable in non-native instruction. These results indicate a dominant use of participant-oriented bundles for hearer purposes (i.e. engagement or procedural bundles) in NEMIC_ENGLISH and a dominant use of stance bundles in NEMIC_DUTCH.

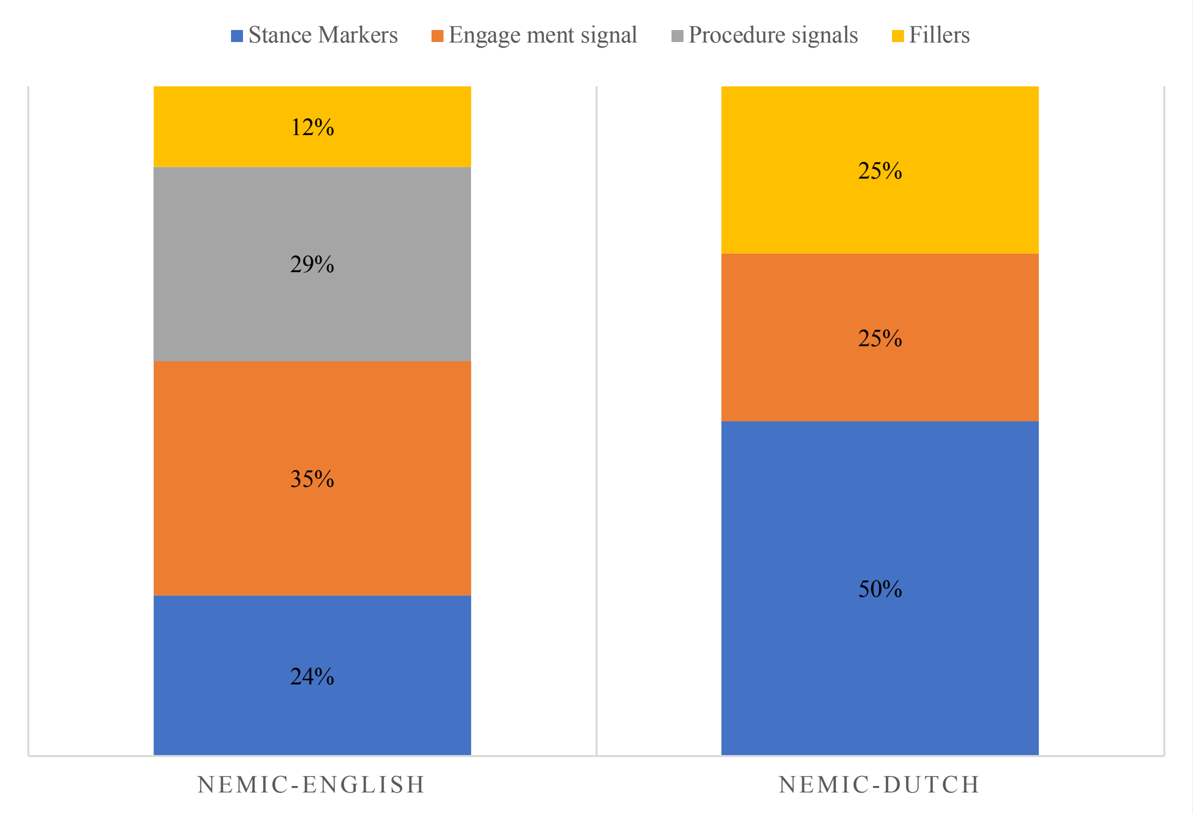


Figure 3. Distribution of participant-oriented function types in NEMIC_ENGLISH and NEMIC_DUTCH

5. Discussion

The analysis described in the previous section aimed to compare a single subject's use of recurrent word combinations in L1 and L2 spoken academic lectures. It compared the use of four-word lexical bundles in a corpus containing L1 Dutch spoken academic lectures and a corpus containing parallel L2 academic lectures produced by the same subject. The difference between L1 and L2 lexical bundle use was investigated in terms of their frequency and function. Frequential differences and a qualitative analysis of overlapping results were used to explain patterns of L1 influence on L2 lexical bundle use.

5.1 Frequential variation

The frequential results of the present study demonstrated that significantly more lexical bundles were used in NNS speech compared to NS speech. The subject has been found to use 141 four-word bundles per 25,000 words in NS Dutch lecturing and 277 four-word bundles per 25,000 words in NNS English lecturing. The subject therefore used almost twice as many bundles in L2 lecturing compared to L1 lecturing. This difference was found to be significantly different. Previous research concerning written data from ESL learners in academic settings showed incongruent results. Our finding supports Bychkovska & Lee (2017), who found a significantly more frequent lexical bundle use in NNSs' writings than in NSs' writings. However, the result found in the present study is in contrast with Ädel and Erman (2012), who found a reversed effect in written academic ESL learners' writings. Comparison to Wang's (2017) study concerning non-native ELF lectures results in congruent findings. Wang (2017) reported ELF speakers to use significantly more four-word lexical bundles than ENL speakers did in spoken academic registers. The first research question: *'Is there a statistically significant difference between the number of four-word lexical bundles used in Dutch NS academic lectures and the number of four-word lexical bundles used in parallel English NNS academic lectures taught by the same subject'* can therefore be answered positively. Four-word lexical bundle frequency was found to be significantly higher in NNS English lectures than in NS Dutch lectures taught by the same subject. However, a higher type/token ratio was found in NEMIC_DUTCH compared to NEMIC_ENGLISH, indicating that the subject used a slightly wider variety of four-word lexical bundles in L1 lectures compared to L2 lectures. Nonetheless, the number of four-word lexical bundles that remained post modification in NEMIC_ENGLISH was higher (*n* = 25) than in

NEMIC_DUTCH (*n* = 18), which demonstrates a reversed pattern. One explanation for the fact that fewer bundles were submitted for analysis in NEMIC_DUTCH than in NEMIC_ENGLISH is that the subject showed a wider variety in L1 bundle use, causing many bundles to not make the minimum frequency cut-off point. However, the mean four-word bundle frequency was found to be higher in NEMIC_ENGLISH than in NEMIC_DUTCH, which corresponds with the found difference in type/token ratio between the two sub-corpora. The results presented in this study therefore contribute to a pattern that has been observed in previous literature (e.g. Ädel & Erman, 2012; Chen & Baker, 2010), which is that NNSs use a more restricted repertoire in the use of lexical bundles than NSs do. However, if the same bundles are used repeatedly in L2 language production, it raises the question whether it is really the case that NNSs use more bundles than NSs do. The results of this study show that higher bundle frequencies in NNS can simply be attributed to the fact that NNSs' range is more restricted, causing NNSs to use the same bundles more repeatedly. In order to draw a solid conclusion of whether NNSs use significantly more lexical bundles in spoken academic registers than NNSs do, frequential measurements should be paired with vocabulary range tests. Instead of the subject being assessed as highly proficient, a vocabulary range test should indicate whether or not the subject's mental set of acquired lexical bundles is similar to the L1 knowledge of lexical bundles.

A higher frequency in L2 spoken academic language production than in L1 spoken academic language production suggests that NNSs rely on the use of prefabricated chunks to a greater extent than NSs do. As learners' L2 language proficiency increases, the number of four-word clusters used in the target language tends to decrease (Lewis, 2009). Moreover, previous studies have found learners with a high L2 proficiency to use significantly fewer bundles in L2 production than low proficient L2 learners (Huang, 2015). This pattern is explained by an increase in vocabulary knowledge, allowing more proficient learners to use a more varied set of lexical bundles. Use of a more varied set of bundles should on its turn result in more native-like use of lexical bundles and thus lower individual bundle frequencies. L2 learners with a higher L2 proficiency should therefore show more native-like frequential patterns in their L2 lexical bundle use than low proficient L2 learners. Since the subject, who was assessed to be a highly proficient EFL speaker (i.e. at Cambridge Proficiency C2-level), has shown to use almost double the number of lexical bundles and a more restricted bundle variety in L2 English lectures compared to L1 lectures, there is reason to assume that perceived native-like L2 proficiency is not necessarily reflected in oral L2 lexical bundle frequency. This finding is therefore in contrast with findings from written academic registers, in which a higher L2 proficiency was found to equal more native-like lexical bundle frequencies.

5.2 Cross-linguistic transfer

The bundles shared between the subject's L1 Dutch and L2 English language production showed divergent frequential patterns of use in NEMIC_DUTCH and NEMIC_ENGLISH. Two patterns were observed that are incongruent with studies reporting evidence of L1 transfer effects in L2 lexical bundles use.

Only 27% of the bundles used in NEMIC_DUTCH and NEMIC_ENGLISH overlapped. This means that out of all 18 bundle types used in the subject's NL Dutch, only three bundle types were used in a translational equivalent form in the subject's L2 English. Since the dataset used in this study was considerably smaller compared to datasets used in previous studies, it is difficult to determine whether this constitutes a large or a small portion of the data. Chen and Baker (2010) found 16% of all bundles used to overlap. However, their data derived from various disciplines. Since the data used in this study exclusively concerns parallel spoken academic lectures a larger amount of bundle overlap is to be expected. Ädel & Erman (2012) used NS and NNS data from the same discipline and found 22% of the bundles to overlap. Compared to this finding, the overlap found in the present study is relatively low, especially considering the fact that the data derived from the same subject. Little overlap therefore suggests that two separate sets of lexical bundles were used in the subject's L1 and L2. This finding suggests that transfer from L1 bundle knowledge (e.g. contextual information, appropriacy, etc.) to L2 bundle use was very limited. The fact that bundles are shared between corpora does not necessarily mean that these bundles are used equally frequently (Ädel & Erman, 2012). Two out of three bundles that were shared between the subject's L1 and L2 were highly frequent in NEMIC_ENGLISH, whereas only one was found to be among the most frequent bundles in NEMIC_DUTCH. Out of the most frequently occurring bundle types in L1 lectures and L2 lectures, only one bundle type was shared. This concerned the bundle *'do not know if'* and the L1 equivalent *'Ik weet niet of'*. This bundle type is said to be typical of academic settings. Wang (2017) found that it is predominantly used in EFL speech to express "a momentarily loss for words" or "as an indicator of insufficient knowledge about the topic of discussion" in spoken academic ELF. ENL speakers on the other hand often use this sequence for hearer-oriented purposes (Baumgarten & House, 2010). Due to their different functions, the most shared bundle *'do not know if'* and the L1 equivalent *'ik weet niet of'* are therefore again no reliable indicator for L1 influence in the dataset used. An explanation for the frequency and overlaps of this particular bundle could be that it can serve many different functions (Wang, 2017), causing a boost in frequential rates in both NS and NNS data. A plausible assumption is that the use of multi-functional lexical bundles is preferred in ELF spoken academic language production as they allow the speaker to use a limited set of formulaic sequences in a high frequency. Further research is needed to test this hypothesis. In order to do so, a more extended framework for identification of lexical bundle function the needs to be developed. Such a framework should allow functional classification of multiple contextual functions served by the same bundle type.

The second pattern that was observed concerns the differences in the number of text-oriented bundles used in L1 lectures and L2 lectures. The finding that the subject overall used fewer text-oriented bundles in NNS speech than in NS speech in combination with the finding that none of the overlapping bundles were found to function as text-oriented bundles suggests that the knowledge of text-oriented bundles in NNS is behind on the subject's L2 knowledge of real-world oriented bundles and participant-oriented bundles. Moreover, this finding does not support the suggestion of L1 transfer effects of text-oriented bundles in spoken language production. Paquot (2015) found the frequency of some text-oriented bundles in EFL learners' written academic language production to be parallel to the frequency of their L1 translational equivalents. If the subject's use of text-oriented bundles in L1 oral language production were to facilitate the use of their translational equivalents in L2 spoken language production, a similar frequency of text-oriented bundles in the two sub-corpora should have been found. Even though the frequential difference between the use of text-oriented bundles in the subject's L1 and their L2 was found to be insignificant, the qualitative analysis of this functional category indicated clear differences in the use of text-oriented bundles. This finding does not support Paquot's (2015) notion of L1 transfer effects on the use of text-oriented bundles.

The data used in this study show no sufficient evidence of L1 transfer effects on L2 lexical bundle use. The second research question: *'are highly frequent four-word lexical bundles in L1 Dutch lectures transferred to L2 lexical bundle use by the same subject?'* can therefore be answered negatively. The present study does not provide evidence to prove that L2 lexical bundle use is facilitated by L1 frequency transfer effects.

5.3 Functional distribution

A detailed analysis of the functional results revealed a preference for participant-oriented bundles in both native and non-native lectures, followed by text-oriented and real-world-oriented bundle types. Wang (2017) reported a similar functional distribution of lexical bundles in ELF lectures. A dominant use of bundles for participant-oriented purposes in university lectures over the use text-oriented (i.e. stance bundles) or real-world (i.e. referential) oriented bundles is said to be typical of classroom teaching (Biber & Barbieri, 2007) and university lectures (Wang, 2017). The found preference for participant-oriented bundles was found to be significantly higher in non-native lecturing compared to native lecturing, which is in contrast with findings by Bychkovska and Lee (2017), who reported a significantly more frequent use of participant-oriented bundles in NS writing than in NNS writing. This difference can be explained by the assumption that a greater use of participant-oriented bundles for in NNS lecturing is used to accommodate to the audience, which is appears to be of greater importance in spoken academic language use compared to written academic language use (Biber & Barbieri, 2007). A more detailed look at the types of participant-oriented bundles used in NEMIC_DUTCH and NEMIC_ENGLISH supports this assumption. The results demonstrate/ that noticeably more participant-oriented bundles were used for listener-directed purposes (i.e. to function as a procedure or engagement signal). In L1 Dutch lectures, only 25% of the participant-oriented bundles were hearer-directed, whereas 64% of the lectures were hearer-directed in L2 English lectures. This difference implies that accommodation of speech to the listener occurs more in NNS speech than in NS speech. A simple explanation for this assumption is that the lecturer is trying to assure comprehension non-native instruction through the use of engagement signals (i.e. in the case of engagement signals) or clarification of procedure (i.e. in the case of procedure signals). Future research using larger data sets is needed to investigate whether lecturers use more participant-oriented lexical bundles for accommodating purposes in non-native lecturing compared to native lecturing. A greater use of accommodating bundles in non-native lectures compared to native lectures could possibly be explained by L2 proficiency. If proficiency is of any influence on the type of participant-oriented bundles that are used in non-native instruction, more proficient learners should demonstrate a participant-oriented bundle use similar to native speakers. Even though the results presented in the present study rely on a small data set, the results show that there is reason to assume that NS lecturers use a more bundles to express stance. Highly proficient L2 lecturers should therefore express more participant-oriented stance bundles and fewer participant-oriented bundles for engagement signalling or procedural functions compared to lower proficient L2 lecturers. Another study design that could be used to test the influence of L2 proficiency on the use of accommodating participant-oriented bundles is to test correlation between the audience's perception and the use of engagement and procedure bundles.

In contrast to participant-oriented bundles, no significant difference was found between the number of text-oriented bundles used in NEMIC_DUTCH and NEMIC_ENGLISH. However, the subject appeared to use slightly more text-oriented bundles in NS lectures compared to NNS lectures. This finding is in line with results presented by Byckovska and Lee (2017), who reported to observe an underuse of text-oriented bundles in NNSs' academic writings compared to NSs' academic writings. This comparison suggests that the use of text-oriented bundles serves similar functions in spoken and written academic registers. However, Biber and Barbieri (2007) found text-oriented bundles to be dominant in NS written academic registers, whereas this study found text-oriented bundles to be the second most preferred bundle function. Even though real-world bundles were the least preferred bundle function, significantly more real-world oriented bundles were used in non-native lectures compared to native lectures. The finding that significantly more bundles were participant-oriented and real-world oriented can again be explained by the fact that a higher bundle variety was observed in the subject's NL, causing frequencies of individual bundles to drop. The more restricted repertoire observed in the speaker's NNL on the other hand caused a repeated use of the same bundles, which brought about an increase in bundle frequencies and possibly an increase in the number of bundles that were categorised as real-world oriented or participant-oriented.

In her framework for the identification of lexical bundle functions, Wang (2017) included a sub-category for repetitional bundles, which are said to be typical of NNS speech. Other frameworks (e.g. Biber et al., 1999; Cortes, 2004; Hyland, 2006) did not include the classification of repetitional bundles. In this study, the subject was found to use a similar number of repetitional bundles (i.e. 6% in both sub-corpora) in L1 lectures and in L2 lectures. This finding suggests that the use of repetitional bundles in academic lectures is not necessarily typical of NNS speech, and equally common in NS speech.

The third and final research question: *'Is the functional variation of four-word lexical bundles significantly different in English NNS lectures compared to Dutch NS lectures taught by the same subject?'* can be answered by accepting the alternative hypothesis. The subject was found to use significantly more participant-oriented and real-world oriented lexical bundles in L2 English lectures compared to parallel L1 Dutch lectures. However, it should be noted that this result is highly affected by a repeated use of a limited set of bundles in the subject's L2 English.

6. Conclusion

6.1 Limitations

The findings of this study have to been seen in light of some limitations. The primary limitation to the generalization of the results found in this study concerns the incomparability of the corpus size to corpora in previous studies conducting research in the area of NS and NNS lexical bundle differences. The corpus that was used in this study was considerably smaller than the corpora used in other studies. Due to the small corpus size, bundle frequencies in this study were normalised per 25,000 words, whereas other studies reported on bundle frequencies that were normalised per million words. Comparison of the results from this study to findings in previous studies is therefore difficult. Besides, other studies report on different bundle sizes, different methods used for automatic retrieval of lexical bundles from a corpus, different settings and different academic disciplines from which data was extracted. Moreover, none of the previous studies report on differences in spoken academic lexical bundle use between two languages produced by the same speaker. Results regarding the comparison to earlier findings in NS and NNS lexical bundle differences should therefore be interpreted with caution. Second, there are a number of limitations to the data which may have affected their analysis. Firstly, the identification of bundle functions was subjective to some extent. Wang's (2017) framework for the identification of lexical bundle functions has been used to guide the classification of bundle functions. However, the classification of the bundles is based on the researcher's interpretation of the bundle functions in the given context. Bundles that fell into more than one category (e.g. *'do not know if'*) were classified according to the most prototypical function, which might have affected the functional bundle analysis. Moreover, prior to this study, the framework used for bundle function identification has only been used in one study regarding lexical bundle use in spoken academic ELF. Comparison of results to studies reporting on bundle functions classified according to other frameworks for bundle function identification is therefore somewhat arbitrary. Secondly, the data was transcribed by two different transcribers. Even though the data was transcribed using the same transcription conventions, this does not mean that individual transcriber differences can be ruled out. Differences in the way lectures were transcribed could have had an effect on the N-Gram search as a result of which bundles might have remained unidentified. Lastly, it needs to be noted that only four-word lexical bundles were considered in this study. The choice to exclusively consider four-word lexical bundles was supported by the fact that most of the previous studies reporting on lexical bundle difference in NS and NNS output also investigated four-word bundles. Logically, the study of four-word bundles made a comparison to previous literature more reliable. However, had two and three-word bundles been considered, a larger number of lexical bundles would have been retrieved, and a more accurate picture of lexical bundle use between two languages performed by the same subject could have been given.

6.2 Future research

The discussion the results presented in the current study have led to four suggestions for future studies in this line of research. Firstly, the results presented in this study suggest that a greater reliance on the use of lexical bundles in non-native instruction is caused by a limited knowledge of lexical bundles in the target language. Additional research is needed in order to test this hypothesis. It is suggested that future studies investigate the relationship between L2 vocabulary range and lexical bundle use in spoken academic settings.

Second, further research is needed to test the hypothesis that multifunctional bundles (e.g. *‘do not know* if’) are preferred in NNS speech as they allow the speaker to use a limited set of formulaic sequences in a high frequency. In order to do so, a more extended framework for identification of lexical bundle function the needs to be developed. Such a framework should allow functional classification of multiple contextual functions served by the same bundle type.

The third suggestion for future research concerns the difference between the functions of participant-oriented bundles used in native instruction in academic settings and non-native instruction in academic settings. The results have led to the assumption that participant-oriented bundles in non-native instruction are predominantly used for accommodating purposes, whereas participant-oriented bundles in native instruction are predominantly used to express stance. In order to contribute to the answer on the question what native-like lexical bundle use looks like in terms of their functions, future research should point out whether highly proficient L2 lecturers express more participant-oriented stance bundles and fewer participant-oriented bundles for engagement signalling or procedural functions compared to lower proficient L2 lecturers. Another question that was raised by the difference that was found in the functional use of participant-oriented bundles is whether the use of accommodating participant-oriented bundles in non-native instruction facilitates understanding. A suggestion for a study design to test this assumption would be to test the correlation between the audience's understanding and the number of accommodating participant-oriented bundles used in non-native instruction.

Lastly, since this study was conducted using a considerably smaller dataset compared to other studies, and because the conclusions were drawn from data produced by only one subject, further research into NS and NNS lexical bundle differences is required to support the findings in this study. In order to give a solid conclusion to the research questions proposed in this study, further research needs to investigate whether similar results are found in larger datasets that includes parallel speech recordings from multiple subjects.

6.3 Implications

The results of this study can be seen as a contribution to the general understanding of the differences between native and non-native instruction in academic settings. Even though listeners and lecturers might not immediately become aware of any differences between NS and NNS instruction, differences in linguistic behaviour do occur, even when the lecturer is a highly proficient L2 speaker. This study contributes to our knowledge of these differences in bundle frequency by supporting previous studies reporting that NNS language production contains significantly more lexical bundles than NS language production does, whether spoken (Wang, 2017) or written (Bychkovska & Lee, 2017). Even though the subject used significantly more lexical bundles in NNS lecturing, the analysis made clear that the observed frequential variation was caused by the fact that a more restricted repertoire was used in NNS lecturing, causing individual bundle types to be used repeatedly. This can be explained by the fact that L2 learners' knowledge of formulaic expressions of often behind on their general L2 proficiency (Steinel et al., 2007). Not only were frequential differences observed, functional differences in bundle use were observed as well. The subject was found to use significantly more participant-oriented and real-world oriented bundles in non-native lectures compared to native-lectures. In line with findings in previously conducted studies, a preference for participant-oriented bundles was observed in both native and non-native lectures. A more detailed analysis revealed that the lecturer mainly used participant-oriented bundles for hearer-oriented purposes in non-native lectures, whereas the preference for participant-oriented bundles in native lectures showed a dominant use of stance bundles. This finding suggests that non-native lecturing is characterised by the use of accommodating bundles, which are assumedly used to check the audience's comprehension in non-native instructional settings. Since no evidence was found for L1 transfer effects from L1 Dutch to L2 English, the results presented in this study suggest that differences between native and non-native instruction in academic settings are language-dependent.

Overall, the results presented in this study suggest that NNS retrieval of prefabricated chunks is easy, since considerably more bundles were used in L2 lectures compared to L1 lectures. However, the discussion of the results in the previous section demonstrated that it is difficult to draw a solid conclusion from the results presented in this study. That is because many intervening factors remain uncontrolled, making it hard to compare the study's results to previous literature. The results therefore need to be interpreted with caution.

References

Ädel, A., & Erman, B. (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. doi:10.1016/j.esp.2011.08.004

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In *Corpora and Lexis* (pp. 277-301). Brill Rodopi.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*(Studies in corpus linguistics, vol. 23). Amsterdam: Benjamins.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, 181-190.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for specific purposes*, *26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405. doi:10.1093/applin/25.3.371

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research*, *10*(3), 245–261. doi:10.1191/1362168806lr195oa

Bychkovska, T., & Lee, J. (2017). At the same time: Lexical bundles in l1 and l2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38-52. doi:10.1016/j.jeap.2017.10.008

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: The influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, *38*(3), 403-443. doi: 10.1017/S027226311500049

Charteris-Black, J. (2002). Second language figurative proficiency: A comparative study of Malay and English. *Applied linguistics*, *23*(1), 104-133. doi:10.1093/applin/23.1.104

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30–49.

Cieślicka, A. B. (2015). Idiom acquisition and processing by second/foreign language learners. *Bilingual figurative language processing*, 208-244. doi:10.1017/cbo9781139342100.012

Conrad, S. M., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes, 23(4), 397-424. doi:10.1016/j.esp.2003.12.001

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Press Syndicate of the University of Cambridge.

De Knop, S., & Meunier, F. (2015). The 'learner corpus research, cognitive linguistics and second language acquisition' nexus: A swot analysis. *Corpus Linguistics and Linguistic Theory, 11*(1), 1-18. doi:10.1515/cllt-2014-0004

ELAN (Version 5.5) [Computer software]. (2019, May 15). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage

Granger, S., & Hung, J. (2002). *Computer learner corpora, second language acquisition and foreign language teaching* (Language learning and language teaching, vol. 6). Amsterdam: Benjamins.

Gass, S. M. (2013). *Second language acquisition: An introductory course*. Routledge.

Hincks, R. (2010). Speaking rate and information content in English lingua franca oral presentations. English for Specific Purposes, 29(1), 4–18. doi:10.1016/j.esp.2009.05.004

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4–21. doi:10.1016/j.esp.2007.06.001

Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. System, 53, 13–23. doi:10.1016/j.system.2015.06.011

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language learning*, *50*(2), 245-309. doi:10.1111/0023-8333.00118

Kashiha, H., & Heng, C. S. (2014). Structural analysis of lexical bundles in university lectures of politics and chemistry. *International Journal of Applied Linguistics and English Literature*, *3*(1), 224-230.

Kellerman, E. (1977). Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin*, 58-145.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

Lewis, M. (2009). *The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency*. Stockholm, Stockholm University.

Matsumoto, N. (2008). Bridges between cognitive linguistics and second language pedagogy: The case of corpora and their potential. *SKY Journal of Linguistics*, *21*, 125-153.

Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers* (The cambridge applied linguistics series). Cambridge: Cambridge University Press.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-304. doi:10.1075/ijcl.11.3.04nes

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi- study perspective. *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes, 64*(3), 429-458. doi:10.3138/cmlr.64.3.429

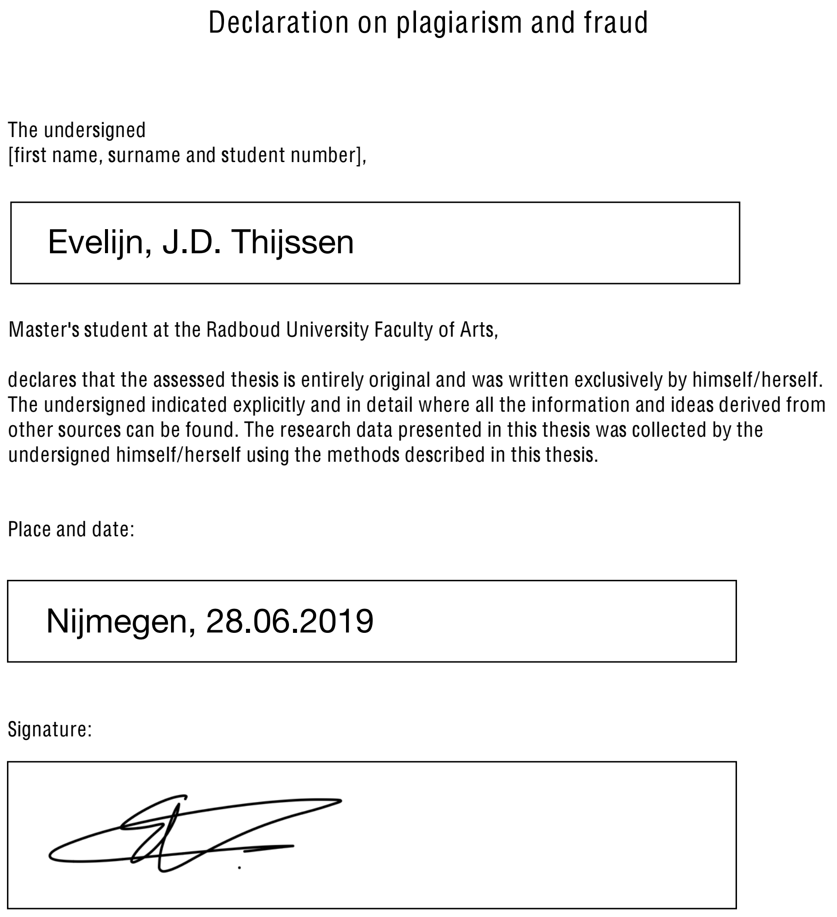
Steinel, M.P., Hulstijn, J.H., & Steinel, W. (2007). Second language learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. Studies in Second Language Acquisition, 29, 449–484.

Thøgersen, J., & Airey, J. (2011). Lecturing undergraduate science in Danish and in English: A comparison of speaking rate and rhetorical style. English for Specific Purposes, 30(3), 209-221. doi:10.1016/j.esp.2011.01.002

Wang, Y. (2017). Lexical bundles in spoken academic ELF. *International Journal of Corpus Linguistics*, *22*(2), 187-211. doi:10.1075/ijcl.22.2.02wan

Appendices

A Declaration of plagiarism and fraud



Declaration on plagiarism and fraud

The undersigned
[first name, surname and student number],

Evelijn, J.D. Thijssen

Master's student at the Radboud University Faculty of Arts,

declares that the assessed thesis is entirely original and was written exclusively by himself/herself. The undersigned indicated explicitly and in detail where all the information and ideas derived from other sources can be found. The research data presented in this thesis was collected by the undersigned himself/herself using the methods described in this thesis.

Place and date:

Nijmegen, 28.06.2019

Signature: