HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

Tiedekunta/Osasto – Fakultet/Sektion – Faculty
Humanistinen tiedekunta / Kielten osasto

Tekijä – Författare – Author
POROKUOKKA, Jaakko

Työn nimi – Arbetets titel – Title
"This sixth sense or something": The use of formulaic vague expressions by Finnish EFL students

Oppiaine – Läroämne – Subject
Englantilainen filologia

Työn laji – Arbetets	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro gradu -tutkielma	08 2019	56 + liitteet (9 sivua)

Tiivistelmä – Referat – Abstract

Vakiintuneen käsityksen mukaan merkittävä osuus englannin kielestä perustuu valmiiksi tuotettuihin fraaseihin, joiden avulla kielenkäyttäjät voivat edistää ilmaisunsa sujuvuutta sekä ymmärrettävyyttä. Aiemman tutkimustiedon perusteella kaavamaisesta kielestä (engl. *formulaic language*) hyötyvät paitsi englannin kieltä äidinkielenään puhuvat myös kielenoppijat. Aihetta käsittelevä kirjallisuus on kuitenkin antanut viitteitä siitä, että englannin kielelle ominaisen kaavamaisen ja idiomaattisen ilmaisutavan oppiminen on yksi haasteellisimmista osa-alueista kyseistä kieltä vieraana kielenä puhuville oppijoille.

Pro gradu -tutkielmani aiheena on kaavamainen kielenkäyttö suomalaisten lukioikäisten englannin kielen oppijoiden tuottamassa puheessa. Keskityn tutkielmassani yhteen aiemmin määriteltyyn kaavamaisen kielenkäytön luokkaan: epätarkkoihin ilmauksiin (engl. *vague expressions*). Tutkielmani aineistona on kaksi oppijakorpusta, FUSE- sekä Hy-Talk -korpukset, jotka koostuvat suomalaisten lukiolaisten suullista englannin kielen taitoa testaavien kokeiden tallenteista. Tutkielmani sisältää sekä määrällisiä että laadullisia piirteitä. Selvitän ensinnäkin korpusohjelmiston avulla, missä määrin kokelaat käyttävät tarkastelun kohteena olevia ilmauksia vertaamalla tuloksia myös aiempaan tutkimukseen. Tutkielmani laadullisessa osuudessa selvitän epätarkkoja ilmauksia ja niiden tekstiympäristöjä tutkimalla, millaisia käyttötarkoituksia varten kokelaat hyödyntävät kyseisiä fraaseja sekä missä määrin havaitut tarkoitukset ovat linjassa aiemman tutkimustiedon kanssa. Lisäksi tutkin, ovatko kokelaiden Hy-Talk -kokeesta saamat kokonaisarvosanat yhteydessä epätarkkojen ilmauksien tiuhempaan ja monipuolisempaan käyttöön.

Aineiston perusteella suomalaiset kielenoppijat hyödyntävät puheessaan epätarkkoja ilmauksia laajasti ja monipuolisesti jopa englantia äidinkielenään puhuviin verrattuna. Laadullisessa tarkastelussa ilmeni, että kokelaat käyttivät tarkasteltuja ilmauksia pitkälti samoja käyttötarkoituksia varten kuin natiivit, mikä osaltaan antaa viitteitä tällaisten fraasien näkyvästä roolista suomalaisten englannin kielen oppijoiden puheessa. Epätarkkojen ilmauksien määrän ei havaittu muodostavan tilastollisesti merkittävää vastaavuussuhdetta suullisen kielitaidon kokeessa saatuun korkeampaan kokonaisarvosanaan. Kokelaiden käyttämien epätarkkojen ilmauksien monipuolisen käytön puolestaan havaittiin olevan tilastollisesti yhteydessä korkeampaan kokonaisarvosanaan.

Avainsanat – Nyckelord – Keywords
Kaavamainen kieli, epätarkka ilmaus, englanti vieraana kielenä, puhuttu kieli, korpustutkimus (suomeksi)

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopiston kirjasto – Helda / E-thesis (opinnäytteet)

Muita tietoja – Övriga uppgifter – Additional information
Tutkielman nimi suomennettuna

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

“This sixth sense or something”: The use of formulaic vague expressions by Finnish EFL students

Jaakko Porokuokka
Master’s Thesis
English Studies
Department of Languages
University of Helsinki
August 2019

Table of contents

1 Introduction

Multiple-word expressions composed of prefabricated elements have gained increasing attention in linguistics (Schmitt, 2010, pp.8-9). An oft-cited metaphor of the prefabricated nature of language is the idiom principle (Sinclair, 1991) which states that language users draw predominantly on a large set of semi-preconstructed phrases that may superficially be composed of several elements but that nonetheless are retrieved from the linguistic repertoire as if they were single items (ibid., p.110). The formulaic nature of language has been approached from multiple angles and with a wide range of methodologies. It is specifically the corpus-based approach that has added enormous amount of information on the recurrent patterns and pragmatic usefulness of various formulaic expressions. Among others, corpus-based research conducted by Altenberg and Eeg-Olofsson (1990), Altenberg (1998), De Cock, et al. (1998) and Biber, Conrad and Cortes (2004) has revealed that users of spoken and written language make substantial use of formulaic language in various registers.

Besides its apparent frequency both in written and spoken discourse (Schmitt, 2010, pp.9-10; Conklin and Schmitt, 2008, pp.72-73), formulaic language has garnered attention due to its pragmatic usefulness in everyday communication (e.g. Nattinger and DeCarrico, 1992) and suspected effect on the speed of linguistic processing (e.g. Conklin and Schmitt, 2008). Until recently, the presence of formulaic expressions in the field of EFL has been a fairly overlooked area of research since many studies have concentrated on producing descriptions for certain areas of formulaic language such as idioms, phrasal lexemes (e.g. Moon, 1998), or observing the effects of formulaic expressions on first language acquisition (Schmitt, 2004, Preface). Recent years have, however, seen a spur of interest in the use of formulaic language from the point of view of non-native speakers as well (Ellis, 2012, p.21).

Indeed, there are indications that the acquisition of prefabricated phrases may increase EFL learners' chances of being perceived as more proficient users of language (e.g. Boers et al., 2006). At the same time, there is evidence that those non-native speakers who operate at lower skill levels do not avail of such language to the extent that more proficient non-natives and natives do (Conklin and Schmitt, 2012). These observations from prior studies strongly suggest that studying formulaic language from the non-native perspective can add invaluable knowledge about the role such language plays in non-native language use and also how it relates to the construct of foreign language proficiency.

This research aims to provide further answers to the question of how and to what extent formulaic language occurs in the oral language produced by EFL learners. To the best of my knowledge, there has been no previous research on how Finnish upper secondary school learners utilize such language in their speech. The current study sheds light on the topic by applying corpus-based methods for the investigation of two corpora: the FUSE corpus and the Hy-Talk corpus which consist of transcripts of oral tests taken by Finnish upper secondary school learners of English. To limit the scope of the study, the focus was restricted to one specific category of formulaic expressions acknowledged as a relevant feature of spoken language: vague expressions. The following three research questions form the basis of the following inquiry:

> 1 To what extent do Finnish EFL upper secondary school learners use formulaic vague expressions in the observed data?
>
> 2 What kinds of functions do these expressions fulfill in the speech of Finnish upper secondary school learners?
>
> 3 How does the exhibited overall proficiency in a spoken language test correlate with the use of vague expressions?

The first question is mainly quantitative in nature. The quantitative results are compared to the results obtained in a prior study made on formulaic expressions by De Cock, et al. (1998) to add weight to the conclusions made about the frequency of vague expressions. Since any insights gained from numerical data is further supported by qualitative data, the second research question, addressing the observed use of vague expressions from a more qualitative perspective, is aimed to add support for the results obtained from the quantitative analysis. The third research question is answered by inspecting the overall grades examinees have received in the spoken language test contained in the transcripts of the Hy-Talk corpus and by examining whether frequent and/or multifaceted use of vague expressions by the examinees correlates with higher overall test grades.

2 Theoretical background

The theoretical background chapter is divided into five major sections. In the first section, a working definition for formulaic expressions is provided together with some crucial background information. In the second section, the identification of formulaic language is discussed. The third section offers a review on the frequency-related aspects of formulaic language. In the fourth section, the focus is shifted to the specific types of formulaic expressions investigated in this study, i.e. vague expressions. Lastly, some issues pertinent to the use of formulaic language from the perspective of EFL speakers are considered.

2.1 Definitions for formulaic language

Formulaic language is a notoriously fuzzy concept field which becomes evident by the number of different names and criteria given for the phenomenon in the literature (Wray, 2002, p.8). These include, among many others, terms such as *formulae*, *chunks*, *multiword units* (ibid., p.9) and *lexical bundles* (Biber, et al., 1999, p.990). In the midst of this varying terminology, Schmitt (2010, p.119) argues that the term *formulaic language* should be considered as an umbrella term which encompasses all the various senses and terms given for the overarching phenomenon. Addressing the multitude of different names which have slightly different meanings and connotations, Wray and Perkins (2000, p.1) propose the term *formulaic sequence*, defined as follows:

> … a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

This oft-cited definition encompasses two opposite extremes which form the overarching continuum of formulaic language. On the one side, the definition refers to "tightly idiomatic and immutable strings, such as *by and large* which are both semantically opaque and syntactically irregular", and on the other side, the definition also accounts for "transparent and flexible ones containing slots for open class items" (Wray and Perkins, 2000, p.1). Reminiscent of Sinclair's idiom principle (1991), Wray and Perkins define formulaic sequences in terms of their singularity in the minds of their users. Thus, although a formulaic sequence might orthographically be composed of several elements, it nevertheless functions as if it was an individual item in the lexis. Moreover, this definition implies that formulaic language is to some extent capable of eluding the constraints

of the grammar, since otherwise it might be difficult to explain why non-canonical expressions such as *so far so good* and *all in all* (Nattinger and DeCarrico, 1992, p.39) are repeatedly used by natives. Wray's categorization also acknowledges that the items forming formulaic sequences do not have to occur next to each other but may well extend over word boundaries. Wray (2002, p.4) mentions idioms as a useful starting point for describing formulaic elements that are processed as a whole. When phrases such as *pull someone's leg* are interpreted as a whole, they might refer to a different idea or object than the combination of the literal meaning of the words would suggest.

However, idioms form a formulaic category of their own, leaning on the tightly idiomatic and immutable side of the spectrum. Although their salient makeup has garnered much attention from researchers, they are quite infrequent in numbers (Schmitt, 2010, p.118). On the other side of conventionalized language lies recurrent expressions such as *do you want me to* which differ strikingly from invariable expressions such as idioms in terms of their considerably higher frequency and mutability (Biber, et al., 1999, p.989). According to Biber and Barbieri (2007, p.264), these *lexical bundles,* identified simply on the basis of their recurrence in a given register, are "not structurally complete and idiomatic in meaning, but serve important discourse functions in both spoken and written texts". Since the methodology used in the current study does not rely solely on frequency-related information, but also heavily on the pragmatic functions which prefabricated expressions serve, another term is needed. To make comparisons with similar prior research more straightforward, the present study uses the term adopted from De Cock et al. (1998, p.67): *formulaic expressions,* which are "frequently used multiword units that perform pragmatic or discourse structuring functions".

2.2 Identification of formulaic expressions

Although definitions reported above include several useful notions for tracking down the essence of formulaicity, it is clear now that the hyperonym of formulaic language includes several subcategories within it. A further complicating factor for the study of formulaic language is that separating formulaic expressions reliably from novel linguistic items is anything but straightforward. Erman and Warren (2000, p.33) point out several important reasons why identifying prefabricated language from novel language is so difficult. First, the status of formulaicity of any expression may be relative to the individual or speech community in question and may also be subjected to changes through the processes of language change and conventionalization. Furthermore, some formulaic expressions have a more idiomatic and less transparent meaning than others which might cause identification problems in the analysis of data.

Thus, any study dealing with formulaic language must address clearly the question of how formulaic expressions can be distinguished from novel expressions and what kinds of categorization criteria are used.

Another central problem in the identification of formulaic expressions arises from the very fact that we expect them to bypass analytic processing of language. How do we know whether an individual in a given situation retrieves an expression from his/her memory without first constructing it analytically from single elements? According to Erman and Warren (2000, p.33) the expression *I am afraid*, denoting the pragmatically equivalent but longer expression *I regret to have to inform you*, highlights the fact that the expression could come about in either way: it could be produced by the speaker holistically as a single unit, but it could also be crafted analytically and the idiomatic form of the expression could be a product of mere chance. Moreover, the formulaic status of an expression depends on the context in which the expression is used. For instance, an expression such as *keep your hair on* can be used either literally to advise someone not to take his/her wig off or figuratively advising someone to calm down (Wray, 2002, p.31). Indeed, frequency-based information cannot reveal whether such usages are formulaic in nature. Instead, pragmatic clues are needed (ibid., p.31).

According to Wray (2002, pp.19-20), studies concerning formulaic expressions have mostly used two major ways of gathering data, the first concerning experimental, questionnaire-based or other empirical research designs to elicit desired formulaic elements which have been predefined by the authors. The second method, more relevant for the current study, has been to concentrate on a specific data set and search for possible candidates of formulaic strings, basing the decisions again on predefined criteria. Because formulaicity and idiomaticity are often seen as closely related concepts, many studies (e.g. Erman and Warren, 2000) have relied much upon intuition and more specifically on the shared knowledge of a specific speech community to determine what can be considered as formulaic language (Wray, 2002, p.20). This may result in the overconfident use of researchers' intuition which, in general, has been proven to be an inadequate device for detecting recurrent patterns in language use at least since the era of corpus linguistics (ibid., p.21; Sinclair 1991, p.4).

However, as the previous examples should illustrate, an informed research into formulaic language should be able to distinguish between fixed idioms, collocations and structural units (Biber, et al., 1999, p.995). According to Nattinger and DeCarrico (1992), the key separating factor between formulaic expressions which they call *lexical phrases* on the one hand and clichés, idioms and

collocations on the other is that the items belonging to the former category have specific functions which their users have learned as part of their pragmatic competence. For example, formulaic expression such as *how do you do?* is used to greet someone in an informal conversation (ibid., p.36). The notion that certain kinds of expressions are repeatedly retrieved and used by language users in specific registers for specific functions has also been acknowledged by Biber and Barbieri (2007, p.283). This evident connection between certain formulaic expressions and the related functions for which they are repeatedly used form one of the guiding principles of this study.

2.3 Frequency-based criteria for detecting formulaic expressions

Identifying formulaic language solely on the basis of intuition carries the risk of imposing more subjectivity on the identification process than is perhaps desired in a scientific inquiry. The same problem may persist even if a researcher outsources the identification procedures to well-informed native judges who might make inconsistent choices on the formulaicity of items (Wray, 2002, pp.22-23). The corpus-informed method, which carries the benefit of yielding quantitative information about any linguistic data under investigation (e.g. Sinclair, 1991), has led many researchers to consider formulaic language more in terms of the recurrent and pragmatic nature of these expressions (e.g. Altenberg, 1998; Biber and Barbieri, 2007). This kind of frequency-based treatment of formulaic language disregards the consideration of the idiomaticity or semantic opaqueness of an utterance, and instead shifts the focus to the recurrent nature and the functions for which multiple-word units are retrieved by language users.

One of the first researchers to take this kind of approach for investigation of the lexis of spoken language English was Altenberg and Eeg-Olofsson (1990). In a subsequent study on the phraseology of spoken language, Altenberg (1998) examined the 0.5 million-word London-Lund Corpus of Spoken English and found out that the data set included a wealth of recurrent word-combinations, 201,000 tokens with over 68,000 different types (ibid., pp.101-102). Altenberg estimated that more than 80 percent of the words in the corpus belonged to multiple-word units of different sizes and frequencies. To limit the scope of his study, he included only recurrent expressions that were made of at least three individual words and which came up in the corpus at least ten times (ibid., p.102). He concentrated on “any continuous string of words occurring more than once in identical form” (ibid., p.101). Most of them were rather short, consisting of 3-5 words with a mean length of 3,15 words. By adhering to the blatantly arbitrary criteria which he lamented himself, he was still able to track down 6,692 tokens and 470 types of recurrent word-combinations in the corpus (ibid., p.103).

Following the frequency-based method, Biber, Conrad and Cortes (2004) investigated the use of lexical bundles in the spoken register of classroom teaching and the written textbook register in the university context. They (ibid., p.376) set an arbitrary threshold limit of 40 occurrences per million words for the inclusion of any recurring multiword expression examined in the study. The size of the analyzed set of lexical bundles was also restricted by excluding all bundles that were less than four words long (ibid.). To eliminate the possibility of idiosyncratic usages skewing the results, Biber, Conrad and Cortes. set the minimum range of 5 (i.e. lexical bundles had to occur at least in five different texts) for the bundles to be included in the study (ibid.). Using these strictly frequency-based criteria for the study of the registers, they concluded that lexical bundles serve important functions in the creation of discourse in these settings (ibid., pp.398-399). Biber and Barbieri's (2007) subsequent study on a wider set of university registers (including instructional registers, student advising registers etc.) established further support for these observations. Biber and Barbieri suggest that the rich usage of lexical bundles for fixed or semi-fixed functions in both written and spoken registers indicate that these expressions are more likely instances of prefabricated language, memorized and retrieved as a whole from memory, although precise experimental methods are still needed to corroborate this (Biber and Barbieri (2007, p.283).

The corpus-informed research concentrating on the frequency of multiple-word units undoubtedly helps researchers to escape the confines of subjectivity to some extent. As Altenberg points out (1998, p.102), frequency is a measure that may be used as a signpost for detecting important and interesting features of spoken language registers. Moreover, frequency-based data can be utilized to detect recurrent aspects of texts which might be difficult to detect if researchers had to comb through massive amounts of data manually (Biber, Conrad and Cortes, 2004, p.376).

However, to begin searching for items, a researcher first has to decide what the focus of interest is which will inevitably affect the results of the study to some extent (Wray, 2002, p.28). For instance, in a study dealing with differences between the use of formulaic expressions by natives and EFL learners, De Cock, et al. (1998) decided to look for continuous combinations of words that were two to five words of length (ibid., p.71). Complying with the general observation that the more frequently an expression occurs, the shorter it tends to be, they set separate frequency thresholds for items of various lengths. They also decided to treat hesitation features such as *erm* and *er* as word items (ibid., p.71). One of the downside effects of these choices may be that they delimit features that might be interesting from the point of view of phraseology (Altenberg, 1998, p.103). For instance, the methodology used by De Cock et al. (1998) and Altenberg (1998) sheds light only on those expressions that are continuous. Thus, discontinuous formulaic phrases, such as the

undoubtedly productive comparative frame *the __ er the ___er* (e.g. *the sooner the better*) (Nattinger and DeCarrico, 1992, p.42), are disregarded.

Not all researchers are unreservedly convinced of the usefulness of frequency-based methods in the study of formulaic language. For instance, Wray (2002, p.31) has stressed that the frequency measures of an expression cannot tell directly whether it is prefabricated or not. This remark is indeed justified at least in the case of highly compositional and/or figurative language. For example, Moon (1998, p.80), studying phrasal lexemes in the Hector corpus, found out that pure idioms appeared infrequently in the data. On the other hand, certain expressions may be frequent simply because different data sets may emphasize particular subject matters (Howarth, 1998, pp.26-27).

Indeed, this disagreement over what should be treated as formulaic language undoubtedly is one of the reasons why studies counting the proportion of formulaic language in both written and spoken texts have arrived at different results. Biber, et al. (1999) concluded that approximately 30% of the material contained in the conversation corpus they analyzed were part of lexical bundles. Altenberg (1998, p.102) estimated that more than 80% of adult native spoken language may be composed of recurrent material. Erman and Warren (2000) investigated how prefabricated multiword combinations and analytically encoded novel lexical items were distributed in both written and spoken native language included in the London-Lund Corpus of Spoken English and Lancaster-Oslo-Bergen Corpus. Their study (ibid., pp.50-51) concluded that, on average, approximately 55 percent of the texts they examined were made of prefabricated material with conversation tending to make use of prefabricated expressions somewhat more than written language.

However, these calculations must be treated with some skepticism. As Erman and Warren (2000, p.50) point out, these figures depend quite much on the way that formulaic language is categorized (ibid.). Nevertheless, many other researchers have come up with similarly high figures using other kinds of methodologies (Martinez and Schmitt, 2012, p.300), which lends some credibility to the results presented above. However, not all of the reports have reported such high figures on the salience of formulaic language. Kecskes (2007, p.198) came up with the figure 7.6% in his study which examined the use of formulaic language in an ELF setting. As the majority of the prior works on formulaic language have reported high figures of its presence especially in the spoken language mode, the scope of the current research was restricted to only one specific type of formulaic expressions to gain more profound insights about the subject. In the next section, this type of formulaic language is discussed more in detail.

2.4 Taxonomy for formulaic vague expressions

Several taxonomies for the identification of formulaic language have been proposed. Some of them (e.g. Erman and Warren, 2000) have been more heavily influenced by the use of researchers' own intuition while others (e.g. Altenberg, 1998) have used a corpus-driven approach to investigate how recurrent lexical patterns emerge from the data. It is particularly the latter method that is used in the quantitative part of this research, although the current approach does include certain amount of subjectivity as well, concerning especially the criteria used for the detection of formulaic vague expressions and interpretation of various pragmatic functions these items serve in the speech of Finnish EFL learners.

All of the taxonomies referred to in this section focus both on the structural properties of formulaic expressions as well as on the various pragmatic functions for which they are used. One of the constantly emerging observations across different treatments of formulaic expressions, regardless of the term used for these expressions, is that they are used for various pragmatic functions (e.g. Altenberg, 1998; Nattinger and DeCarrico, 1992). Indeed, besides offering a processing benefits for both speakers and hearers, the second major function of formulaic language is related to its usefulness for promoting various socio-interactional functions (Wray, 2000, pp.474-477). These same macro functions are centrally related to the use of vague expressions as well. What follows is a brief overview on the taxonomies that serve as the basis for the identification of formulaic vague expressions and their related pragmatic functions in the subsequent analysis sections.

2.4.1 Vagueness tags

Vagueness tags are multiple-word expressions indicating impreciseness. This category of formulaic language includes phrases such as *and stuff like that, and everything* and *or something* which generally occur in the clause-final position to expand otherwise complete phrases or sentences (Altenberg, 1998; pp.117-118; Overstreet and Yule, 1997, p.250). According to Cambridge Dictionary (2019), vague expressions are particularly frequent in spoken language. Several corpus-based studies have confirmed their salient role in spoken language as well as their functional utility. In a study concentrating on recurrent word combinations in a spoken language corpus, Altenberg (1998, pp.117-118) noted that these kinds of multiword units reflect the demands pertinent to real-time conversation in general. According to Altenberg (ibid.), vagueness tags are utilized to signal the lack of precision typical for spoken language registers. Biber, Conrad and Cortes (2004) use the term *imprecision bundles* for various vagueness tags and classify them under the main label of

referential expressions which refer to lexical bundles identifying entities and highlighting certain features related to them (ibid., p.393). In their work on university registers, imprecision bundles were found to have two major functions: to mark the information accompanied with these expressions as somewhat imprecise or to allude to the fact that other examples could be provided in addition to the one referred to (ibid., p.394). In a similar vein, Nikula (1996, p.52) has suggested that the use of modifiers such as *and everything* and *kind of* can be used by speakers to express uncertainty in terms of the proposition uttered. On the other hand, they can also be used to mark the meaning of an utterance as vague for strategic purposes (ibid).

While the use of vagueness tags has sometimes been viewed as stigmatized language indicating lower-class status of their users, closer examination on how these items are utilized by native users has revealed their usefulness particularly in social interaction (Overstreet and Yule, 1997, p.250). According to Overstreet and Yule (ibid.), vagueness tags, which they refer to as *general extenders*, are applicable for attending interpersonal concerns in conversation such as expressing solidarity towards conversationalists, drawing on assumed common experience and maintaining rapport between speakers. By comparing two American English corpora, one containing telephone and face-to-face conversations in more informal settings between native speakers known to each other and the other exchanges between unacquainted people in more formal speech events, Overstreet and Yule (1997, p.252) concluded that the former data set exhibited more frequent and multifaceted use of vagueness tags. Moreover, informal exchanges exhibited a preference for different types of vagueness tags such as *or whatever* and *and stuff*, whereas formal exchanges often favored vagueness tags such as *et cetera* and *and so on*.

Research on formulaic language has demonstrated that vague language is at least to some extent related to individuals' ability to produce discourse in a fluent way. For instance, the interrelation between fluency and vagueness tags has been acknowledged by Nattinger and DeCarrico (1992, p.64) who categorize vagueness tags such as *and so on* as discourse devices which serve the distinct function of building fluency. Since vagueness tags are used extensively by proficient natives to address interactional needs (e.g. Overstreet and Yule, 1997), this leads to an interesting question: how are these items used by non-natives who generally are in a disadvantaged situation in terms of reaping the processing benefits of formulaic language? De Cock, et al. (1998), delving into formulaic aspects of native and non-native speech with a corpus-based methodology, found out that advanced French adult learners use significantly less vagueness tags than native speakers of British English (ibid., p.77). According to De Cock, et al. (ibid., pp.77-78), this may influence the way speakers are perceived as learners and may also help to illuminate why their speech sounds more

unnatural to the native ear. De Cock (2004) repeated these observations in a subsequent corpus-based study concentrating on recurrent word-combinations in two spoken-language corpora consisting of native and non-native speech. According to De Cock (2004, p.236), the narrow set of vagueness tags utilized by French learners such as *and so on* reflect rather formal than informal qualities of language, making the speech of learners seem more out of place and less natural than native speech. Based on this evidence, it may be hypothesized that learners, in general, seem to be more insensitive to the advantageous roles these expressions serve in building fluency and rapport in interpersonal communication.

2.4.2 Two markers of imprecision: *kind of* and *sort of*

The current study also includes two markers of imprecision, *kind of* and *sort of,* which are particularly interesting from the point of view of comparing native language to learner language. These expressions are not vagueness tags per se, but work functionally very much in the same way, expressing uncertainty, downtoning statements and hedging information (Aijmer, 2002, pp.199-201; Erman and Warren, 2000, pp.43-44) and contributing to the overall vagueness of messages (De Cock, et al., 1998, p.78). Moreover, these expressions are important devices for establishing common ground between the speaker and the hearer (Aijmer, 2002, p.209). They also enable speakers to distance themselves from the contents of utterances and add a sense of informality and congeniality in the conversation (ibid.). Corresponding with the observation of non-natives using less vagueness tags in speech than natives, these items have been found to be more infrequent in the speech of learners than in native speech (De Cock, et al., 1998; 2004). Moreover, non-natives seem to use these items for different pragmatic functions than natives (ibid.). For instance, learners may occasionally use these devices to express their lack of linguistic knowledge of English equivalents, as in *sort of braderie* (De Cock, et al., 1998, p.78).

Henceforth, the term *vague expression* is used as a hyperonym to denote both vagueness tags and the aforementioned two markers of imprecision. Wherever necessary, the term *vagueness tags* is used for denoting the group of vagueness tags discussed in the beginning of this section, excluding the two expressions, *kind of* and *sort of,* henceforth referred to as *markers of imprecision.*

2.5 The utility of formulaic language for native and non-native speakers

From the native speaker perspective, the processing benefits of using formulaic language for both the hearer and speaker seem to be considerable (Wray, 2002, pp.15-18). When speakers utilize prefabricated multiple-word units, not only does it reduce the effort needed to create the stretch of

language but also relieves the recipient of the burden of decoding the utterance into smaller units, decreasing the possibility of misunderstandings (Wray 2002, pp.94-95). Pawley and Syder (2000, p.195), addressing the notion that human language capacity seems to be severely limited to cope with structures larger than one clause at a time, suggest that prefabricated expressions together with ready-made syntactic frames provide speakers with a 'crutch' by which they may bypass these apparent limitations of the human cognition.

Indeed, Pawley and Syder (1983, p.192) estimate that native speakers of English have a considerably larger stock of more or less conventionalized sentence-length expressions (perhaps hundreds of thousands) than individual lexical items. Other examinations on the recurrent nature of language have arrived at similar kinds of conclusions (e.g. Biber, et al., 1999). As Kecskes (2016, p.6) points out, this is largely due to the fact that any speech community sharing the same language begins to generate its own normative set of rules concerning language and lexical choices, resulting from linguistic "core common ground development" (ibid., p.17). Nevertheless, the use of formulaic language is by no means confined to native speech. Prefabricated expressions have been found to play part in successful acquisition of foreign language, particularly in the initial stages of learning where much support and encouragement are needed (Wray 2002, pp.147-148).

As we may induce from our own learning experiences, none of us proceed exactly along the same pathways when learning a foreign language. By the same token, it would be sensible to approach formulaic language by acknowledging the fact that it is not learned, used and processed similarly by all types of EFL learners. Wray (1999, pp.222-223) argues that in order to gain a deeper insight of the way learners use formulaic language in order to match their interactional and processing needs, researchers should consider their individual variation in terms of their different ages, learning abilities, attitudes, and learning styles. According to Wray (2002, p.144) these factors have been somewhat disregarded in studies, presupposing that learners do not have other variables besides their learner status. By appointing to the lack of precision in terms of accounting for these idiosyncratic differences in the existing literature, Wray has classified EFL learners to three distinct groups: young children (up to the age of nine), teenagers and adults learning with formal classroom tuition and teenagers and adults learning without formal classroom tuition (ibid., p.223). This study takes heed of these observations by acknowledging the fact that any results only reflect formulaic aspects in the language of those particular types of learners investigated in this research and cannot be generalized to other learner groups.

2.5.1 Formulaic language and foreign language acquisition

Research has added some invaluable insights into how formulaic language may offer advantages for non-native speakers. According to Boers, et al. (2006, pp.246-247), formulaic language is advantageous to learners for three main reasons. Firstly, many standardized formulaic expressions cannot be built analytically from their constituent parts due to their idiomatic and sometimes even ungrammatical nature. Thus, learning such expressions offers learners the means to become more proficient in their repertoire. Second, utilizing preprocessed and preformulated expressions offers learners means of bypassing the need to create utterances from scratch in situations demanding real time processing. According to Boers, et al. (ibid.), the lack of hesitation markers in a stretch of speech could be an indication of the presence of formulaic speech strategy. On the other hand, presence of hesitation markers could indicate where formulaic units begin or end. The third advantage of such expressions for the learners relates to the notion that formulaic items are "zones of safety" (ibid.) where learners may operate without the fear of making linguistic mistakes.

The positive effect of formulaic devices on the fluency of non-native speakers has also been acknowledged by Wood (2006) who investigated the various functions formulaic sequences serve in the speech of second language learners. From the data, consisting of spoken narrative retellings, he identified five categories of usage that each contribute to the fluency of second language speakers: (1) repetitive uses of formulas in a single turn, (2) combinations of formulas to make individual stretches of speech longer, (3) preference for a particular formula used repeatedly by an individual (4) self-directed talk and fillers and (5) reliance to formulas for rhetorical purposes (ibid., p.24). Wood (ibid., p.29) argues that these usage types provide speakers with means to make individual stretches of speech more verbatim, and as a consequence, enhance their speech production.

There have been some claims that learners generally tend to learn formulaic language quite effortlessly during the first steps of their foreign language education. However, the accumulated knowledge of lexical items and grammatical constructions is usually accompanied with a staggering pace in the learning of formulaic material (Wray, 2002, p.182). It is crucial to acknowledge that classroom learners use formulaic language for different functions than native speakers. According to Wray (2002, p.205), classroom learners often do not need to address tangible issues such as their physical, intellectual or emotional state which discourages them to use formulaic material to address these urgencies. Moreover, the instruction generally encourages them to concentrate on the formal aspects of L2 language. Another factor that may have a decreasing effect on the use of formulaic expressions by L2 learners is the way they usually process linguistic input. The instruction material

may be rich in formulaic expressions, but these may go by unnoticed by the L2 learners who are engaged in the processing of new words instead of the larger context surrounding them (Wray, 2002, p.206). Thus, L2 learners tend to form strings of language analytically in a word-by-word fashion, whereas native speakers have available a wider variety of idiomatically sound prefabricated phrases (ibid). To mend this state of affairs, suggestions have been made for instructional methods that would highlight lexical patterns (e.g. Lewis, 1993, as cited in Boers, et al., 2006, pp.247-248; Nattinger and DeCarrico, 1992). The usefulness of the Lexical Approach by Lewis was corroborated in an experimental study by Boers, et al. (2006) where learners were reported being perceived as more proficient users of language by teachers after having been subjected to instructional methods highlighting phrase-noticing.

The usefulness of formulaic language for learners is not without its controversies, however. One of the controversial issues in the field relates to the proposed connection between learning of formulaic language and development of grammatical competence (Myles, Hooper and Mitchell, 1998, p.327). In addition to helping learners become more native-like, use of simple formulas may also be a communication strategy by which learners may evade the use of more verbose strategies (ibid.). Some have claimed (e.g. Nattinger and DeCarrico, 1992, p.114) that learning formulaic phrases through simple means of imitation enables learners to use those phrases creatively later in the learning process as they have acquired the skills to break them down into smaller constituents. However, the benefits of learning formulas may become a burden in certain situations for learners of foreign language (Wray, 2000, pp.482-483). The learner might be puzzled by the fact that while regular expressions such as *largely speaking* may be broken down and used for generalizations of rules, syntactically or semantically irregular expressions such as *at large* cannot be subjected to this kind of analysis (ibid.).

In their longitudinal study of 11 – 13 years old children learning French as a foreign language, Myles, Hooper and Mitchell examined the acquisition of unanalyzed chunks in the classroom context (1998, pp.358-359). They concluded that formulaic language helped their subjects to communicate in French and also facilitated the production of linguistic output in the early stages of the instruction. As the children progressed in their foreign-language skills, the simple formulaic material provided by the instruction for conversation tasks were no longer complex enough to satisfy their growing communicative needs. This led the majority of learners to break the formulaic expressions down and revise them in such a way that would match their current needs. However, according to Myles, Hooper and Mitchell (ibid., p.359) the retention of the formulas in the

repertoire indicates that the formulas still served the function of a reference point from which learners could test their hypotheses about correct forms.

2.5.2 Formulaic language in the repertoire of EFL learners

Mastery of formulaic language has often been described as a skill rather difficult to attain for an EFL speaker, being one of the aspects of language use that sets non-natives apart from native speakers of English (Conklin and Schmitt, 2008, p.84). This sensitivity to idiomatic expressions and structures causing severe challenges even for advanced learners has been referred to as the native-like selection (Pawley and Syder, 1983). According to Pawley and Syder (ibid., p.215), language learners tend to generalize about the use of certain elements within expressions and falsely assume that they can be used in other contexts, often resulting in unidiomatic stretches of language (e.g. *John has a thigh-ache; I intend to teach that rascal some good lessons he will never forget*). The lack of this kind of pragmatic linguistic knowledge about preferred native-like expressions produces a knowledge gap where even advanced learners may select utterances that native speakers would not use (ibid., p.215).

Researching the phenomenon through the ELF context, Kecskes (2016, p.3) has pointed out that L2 speakers use more creative constructions compared to native speakers who have a larger number of prefabricated phrases available. Among the possible factors contributing to this discrepancy he points to factors such as lack of skills, preference for particular formulas, fluency of conversationalists and lack of common ground. The link between the use of formulaic expressions and nativelikeness has also been examined by Ortaçtepe (2013) who sought to find out how nativelike international students with Turkish background and American students were perceived by external raters. According to the results, the American students were perceived as more nativelike by the judges and they evidently used more extensively the inspected formulaic responses than the Turkish students who preferred analytically composed language, indicating non-native like linguistic behavior.

Despite these differences in the use of formulaic language by average native and non-native users, Wray (1999, p.213) has argued that the formulaic lexicon of a non-native speaker is not only a flawed adaptation of the one mastered by competent natives but rather a category of its own which deserves a closer examination. According to Schmitt (2010, p.142), the non-native use of formulaic language can be approached from three directions: how much non-natives use such language, how accurately they use it, and how well-founded their intuitions about formulaic language are. The

often-made claim that non-natives use less prefabricated language than natives is called into question by Schmitt (2010, p.143) who cites a number of studies which have reported that non-natives tend to use formulaic expressions they have in their repertoire and underuse certain kinds of formulas. Likewise, in a comparative study of spoken language learner and native corpora, De Cock, et al. (1998, p.78) found out that prefabricated language may occasionally occur in the speech of advanced learners even more frequently than in the speech of natives. However, the study concluded that non-natives sometimes prefer expressions that natives would not select. There are also differences in the frequency of formulaic expressions, the preferred syntactic patterns, and pragmatic functions for which those items are used by non-natives compared to natives (ibid., p.78). De Cock came to similar conclusions in a subsequent corpus-based study (2004) on spoken native and non-native recurrent sequences by stating that "advanced learners' use of frequently recurring sequences of words displays a complex picture of overuse, underuse, misuse of target language NS sequences and use of learner idiosyncratic sequences" (ibid., p.243).

Carey's corpus-informed study (2013) on the so-called organizing chunks in the ELF context adds another important dimension to the complex nature of formulaic expressions. Just as both native and non-native users of language may resort to grammatically ill-formed novel sentences, formulaic expressions may exhibit the same kind of variance. Thus, formulaic expressions such as *in my view* may have several other manifestations such as *in my sense*, *in my belief* and *in my thoughts* (Carey, 2013, p.215). According to the study, the less frequently used a formulaic expression is, the more prone it is to exhibit variance. Carey suggests that this could be the result of the lack of exposure to these expressions especially in the case of L2 users (ibid., p.225).

3 Material

To investigate the use of formulaic vague expressions by Finnish EFL learners, two Finnish EFL spoken language corpora were chosen as the target of analysis: the Hy-Talk corpus and the FUSE corpus (The Finnish Upper Secondary School Corpus of Spoken English). Both of the corpora were freely available for research purpose in transcribed form with relevant metadata provided along with the texts. These corpora consist of transcribed oral performances of students taking part in spoken language tests aimed at upper secondary school level. Each performance was assessed according to the CEFR framework of reference. The corpora were used to investigate to what extent and in what ways Finnish EFL learners employ formulaic vague expressions in their speech. Moreover, the assessment metadata contained in the Hy-Talk corpus was utilized to inspect how individual users operating at different skill levels avail of these expressions, and whether more frequent and multifaceted use of vague expressions correlates with higher oral proficiency exhibited in the test. Since the individual texts included in the FUSE corpus differ drastically in terms of their sizes, ranging from only 45 words to 1,311 words, and the transcripts contains only one part of the spoken language examination, only the Hy-Talk corpus was used for the purposes of answering the third research question. The details concerning these corpora are summarily discussed before turning the attention to the methods used in this research.

3.1 The FUSE corpus

FUSE (FUSE — The Finnish Upper Secondary School Corpus of Spoken English) is a web-based open-access corpus that includes recordings of Finnish upper secondary school students taking part in the oral examination of English taking place at the end of the vocational course 8 (Ehrnrooth, 2018). All of the recordings are available both in audio and transcribed text format on the website of the project. Only the text files were used for the purposes of this study. The earliest recordings included in the database date back to 2014. Although the corpus is quite small as yet, the aim of the corpus is to become a nationwide database to which any EFL teacher working in Finland and willing to cooperate with the project may upload examination recordings. The size of the original corpus is 23,263 words, containing transcripts of 28 pairs taking part in the oral examination. After preparing the texts for the purposes of this study, the details of which are explained in the methodology section below, the size of the corpus was reduced to 20,404 words.

3.1.1 The set of tasks in the FUSE corpus

The tasks transcribed for the corpus belong to the third part of the Spoken Examination of English for Finnish upper secondary school students (Ehrnrooth, 2018). In the current exam format, there are two types of tasks: structured dialogue tasks and open-ended conversation tasks. The instructions and some examples of the assignments are included in the appendices section (Appendix A).

3.2 The Hy-Talk corpus

Hy-Talk (The HY-Talk Corpus) was a 3-year research project on oral proficiency in foreign languages in compulsory basic education and general upper secondary education carried out in the University of Helsinki. The database was originally designed as a source of information for investigating whether the oral proficiency scales of the core curricula enabled the assessment of communicative oral proficiency of students in the general language education with sufficient reliability (Hilden, et al., 2010). Besides English, the database includes transcripts of examinees taking the test in Swedish and German. In total, the English part of the database includes transcribed recordings of 20 upper secondary school student pairs and eight lower secondary school student pairs taking the test. The data was obtained by recording the speech of pupils studying in two upper secondary schools and one comprehensive school, each of them located in the Helsinki metropolitan area.

The size of the original English part of the corpus is 55,402 words. After having concluded that lower secondary school students do not use vague expressions to the extent that would produce meaningful insights, the scope of this research was restricted to the transcripts of tests taken by upper secondary school pupils. After preparing the text files contained in the corpus for the purposes of the present study, the word count reduced to 39,569 words. The average length of the individual texts is 989 words, ranging from 570 words to 1670 words.

Each transcript include metadata such as school, language, recording duration, recording date, and possible notes. Each examinee was also designated with an identification number, making individual comparisons possible in the current research. All grades are based on the CEFR framework of reference. The performances of students in each task were graded individually according to five skill areas: (1) performance, (2) fluency, (3) pronunciation, (4) broadness of expression and (5) correctness. In addition to these, a mean grade based on the performance in the entire test has been given to each examinee. The assessments of each performance were marked in

the excel files contained in the complementary metadata. This study inspects only the correlation between overall proficiency grades and the use of formulaic vague expressions in the test.

3.2.1 The set of tasks in the Hy-Talk corpus

The set of tasks included in the Hy-Talk corpus includes a brief warm-up interview, one structured monologue task, three structured dialogue tasks and one extra open-ended dialogue task. The structured tasks involve a short presentation and imaginary communicative situations which are acted out by the examinees. The extra task involves a more open discussion about issues related to the phenomenon of reality TV. The transcripts of all of these tasks were included in the corpus analyzed in this study. The instruction hand-out and details concerning the tasks are included in the appendices section (Appendix B).

4 Methods

This study adopts a corpus-based approach to investigate to what extent Finnish EFL learners utilize formulaic vague expressions in their speech. The use of the expressions is also considered from a qualitative perspective. Thus, the identified vague expressions and their co-texts are inspected to investigate what functions these types of expressions serve in the speech of Finnish EFL learners. Lastly, the correlation between two variables, (1) received overall grades from a spoken language test and (2) the observed use of vague expressions, is inspected with statistical methods to answer the third research question. In this section, justification for the focus of the research is provided, after which the methods used for answering the research questions are explained more in detail.

4.1 The selected areas of focus

The decision to select the category of vague expressions as the target of analysis was first based on its relatively established definitions outlined by previous research. Significantly, the literature (e.g. De Cock, et al., 1998) suggested that the use of vagueness tags might be linked to the overall oral proficiency exhibited by the examinees. This enabled comparisons between the outcomes of the current research and prior studies. The notion that the texts included in the corpora do not represent natural language influenced the selection of the category as well, leading the focus to those types of formulaic expressions that were likely produced by the examinees in a more unprompted fashion. The utilization of vague expressions could be seen as a result of any speaker's attempt to address general needs pertinent to oral language, such as the need to mark messages as imprecise and willingness to attend to social concerns of interpersonal communication.

4.2 The corpus analysis software

The freeware corpus software AntConc (Anthony, 2018) was used to investigate the frequency of multiple-word units known as *n-grams* which are any combinations of two or more words that co-occur adjacently in some part of a text. The AntConc software has several useful functions for the investigation of recurrent vocabulary patterns in a text or a collection of texts. Particularly useful for the current research are the n-gram search function and the concordance tool. The search function allows users to search for multiple-word units according to their size (e.g. two-word units), their range (i.e. in how many texts they appear) and frequency (i.e. the number of occurrences). The concordance tool provides the co-text surrounding the expressions yielded by searches, making the

interpretation and analysis of the results easier. The program also allows the user to convert the concordance and n-gram search results into text files.

4.3 The preparation of the texts for analysis

The texts included in the FUSE corpus were originally in web page format, whereas the Hy-Talk texts were in binary file format (.doc). To make the file format consistent and readable for the corpus analysis tool used in this research, the texts were converted to the text file format (.txt).

Some extra preparation work had to be done manually so that the texts would not include any material that might skew the results. The AntConc software enables its users to perform searches on sequences of words, yielding results that ignore speaker turn signs and any symbols that are not alphabetic such as numerals and slashes. Additionally, it is possible to ignore a stretch of speech produced by a particular speaker by letting the program know that it should not take into consideration any words enclosed within a particular bracket pair indicating a specific speaker turn (e.g. *<S1> </S1>*).

Unfortunately, in the case of the Hy-Talk corpus, the transcription policy had been slightly inconsistent in terms of the way that the speaker turns were marked. In some of the files, the utterances made by the examiners were marked with the identifier *<S1> </S1>* while in other files the examiners' output was marked with, for example, the identifier *<S3> </S3>* or *<S3> </S3>* and *<S4> </S4>* if the test was carried out with a pair of examiners. Moreover, the transcriptions of the Hy-Talk contained some errors in the marking of turns, which made the program to interpret the language produced by examiners as language uttered by examinees. To solve this issue, separate text files for each of the examinees were created in which only the utterances produced by a particular examinee were included. This warded off the possibility of linguistic noise (e.g. comments made by the examiners) skewing the data, and also made individual comparisons more reliable and easier to process.

Since the comments made by an interlocutor in the immediate vicinity of vagueness tags could help to interpret the intended pragmatic functions of vagueness tags, original text files of the corpus were still kept in store in case there would be a need to inspect the larger context in which a particular vague expression occurs.

4.4 The criteria and the identification of formulaic expressions

In line with prior corpus-based research, predefined criteria were used to conduct n-gram searches on the items of interest. A formulaic expression had to consist of at least two words but be no more than five words of length to be included in the study. The identification of formulaic expressions relied firstly on the frequency information obtained from n-gram searches. Thus, any expression included in the first step of analysis had to surpass certain number of occurrences to be included in the study. For two-word n-grams, the frequency threshold was set to the level of four, while three-word and four-word n-grams received the frequency threshold level of three and five-word n-grams were inspected according to the threshold level of two occurrences. As the choice to concentrate only on a limited set of formulaic expressions had already limited the scope of this study, this rather inclusive approach ensured that also lower-frequency vagueness tags could be tracked down by n-gram searches.

The second step was to subject this set of potential vague expressions to manual analysis together with their co-texts to decide whether they meet the criteria set for the formulaic expressions of interest. Since the n-gram search results would presumably contain recurrent word combinations which might co-occur without a clear pragmatic function (e.g. Altenberg, 1998), these instances had to be separated from the ones that actually served the predefined functions. In other words, phrases without a clear pragmatic function of denoting vagueness were not counted.

Granted, the second part of the analysis is rather subjective, a concern raised also by De Cock, et al. (1998, p.75) in their similar work on the vagueness tags used by native and non-native speakers of English. However, since the definitions outlined for the expressions hinge mainly on the pragmatic functions for which these expressions are used in conversation, some closer manual inspection of the texts must be carried out to be sure that a given expression is indeed used for a distinctive pragmatic function.

One further essential choice had to be made regarding both the way single expressions are treated as separate entries and how their occurrences are counted. This stemmed from the fact that certain expressions form quite flexibly variant expressions that utilize some words of the larger and more frequent formulaic expression but omit others (e.g. Nattinger and DeCarrico, 1992). For example, the common vague expression *and stuff like that* may also be encountered in the abbreviated form *stuff like that*. In the case of vagueness tags, a decision was made to follow the principle set by prior research (e.g. Altenberg, 1998; De Cock, et al., 1998) and to treat these items as separate entries,

establishing separate frequency counts for each of these expressions. In other words, instances of *stuff like that* were not regarded in the frequency count of *and stuff like that*.

4.5 The correlation between the use of formulaic vague expressions and received grades

To answer the third research question, the numbers and types of formulaic vague expressions exhibited by each of the 40 examinees in the spoken language tests transcribed for the Hy-Talk corpus were first counted manually. The metadata included in the corpus files contained overall grades for each of the examinees, based on the mean grade for the performances in four spoken language tasks. The performances of the examinees had been rated by five judges individually. As the corpus contained warmup interviews (see Appendix B) that were not rated by the judges, a subcorpus excluding these sections was compiled so that the expressions of interest occurring there would not skew the results. The statistical significance of the correlation between both number as well as types of vague expressions used by each examinee and the received grades were then tested with Microsoft Excel.

5 Results and analysis

In this section, frequencies of vague expressions in the speech of Finnish learners are reported. The results of this study are also compared to a prior corpus-based research conducted by De Cock, et al. (1998) which includes a separate discussion of the use of vagueness tags in the speech of young adult British English natives and French young adult learners of English. The qualitative aspects related to the use of vague expressions are approached by observing whether the functions for which Finnish EFL learners use vague expressions differ from the functions identified by prior research (e.g. Overstreet and Yule, 1997; De Cock, et al., 1998). Lastly, the hypothesized correlation between the use of vague expressions and oral proficiency exhibited in a spoken language test is examined.

5.1 The most frequent potential formulaic vague expressions

As mentioned earlier, the first procedure in the identification of formulaic vague expressions involved an n-gram search for n-grams consisting of two to five words in the chosen corpora. To illustrate the preliminary data obtained at this stage, Table 1 shows all the potential vague expressions that reached the level of 15 occurrences as well as their ranges and frequency ranks compared to all n-grams of similar sizes.

N-grams	Frequency	Range	Rank
kind of	127	46	21
like that	80	40	46
or something	56	28	81
and that	54	29	86
something like	38	22	145
something like that	30	18	37
and stuff	29	18	207
and like	21	15	321
(i)s kind	20	11	356
that kind	20	8	358
and all	19	12	379
[i]s kind of	19	11	89
that kind of	19	8	91
or something like	18	15	100
it [i]s kind	16	10	123
or something like that	15	13	27
it [i]s kind of	15	10	26

Table 1. The most frequent potential vague expressions in the Hy-Talk/FUSE corpora.

The frequencies, ranges and ranks shown in Table 1 provided some preliminary evidence for the considerable presence of vague language in the corpora. For example, the two-word n-gram *kind of* is one of the most common two-word n-grams in the corpora. The makeup of some of the potential vague expressions bear resemblance to other items in the list, several of them including core words such as *stuff*, *something* and *kind* in them. It should be noted that all of the occurrences of the four-word n-gram *or something like that* are also included in the frequency count of the two-word n-gram *or something*. Thus, a manual analysis of each of these n-grams was necessary to ensure whether a potential vague expression was actually a type of its own or embedded within a lengthier vague expression.

5.2 Manual filtering of the potential vague expressions

The manual filtering process following the automatic n-gram search deserves a separate discussion. As mentioned earlier, AntConc's concordance tool allows its users to view the linguistic context, i.e. the co-text, of each item of interest. Closer inspection on the potential vague expressions revealed that the n-gram searches had yielded many instances which under careful scrutiny did not meet the criteria outlined for the expressions. However useful the n-gram search is for detecting recurrent patterns in the data, the search yields many instances which under careful examination do not meet the criteria outlined for expressions that have distinct pragmatic function of denoting vagueness. Consequently, phrases with the pragmatic function of denoting vagueness were identified by inspecting each potential vague expression and its co-text manually with the concordance tool and counting only those instances which fulfilled the pragmatic function. Thus, instances such as *and everything* in the extract "…it worked well and everything was great*"* (FUSE) were not counted in the final frequency counts since they did not meet the criteria.

The in-depth manual analysis of the phrases posed some other subjective choices that had to be made before the absolute frequency counts could be established. In a few instances, a vague expression was followed by material that had not been intelligible for the transcribers. Those passages were marked with the letters *xx* in the transcripts, as in "maybe some sausages and things (xx)" (Hy-Talk). These were categorized according to the part that had actually been heard by the transcribers, since it was impossible to know whether the intelligible part was in reality a part of a lengthier vagueness tag (e.g. *and things [like that]*). However, these instances were rare in the observed data. Repetitions of vague expressions by individuals in the same turn were removed from the final count since counting repetitions of identical phrases would have arguably skewed the results to some extent. To identify passages where hesitation features such as *erm*, pauses, stuttering

or other extralinguistic units occurred between the elements included in the vague expressions, wildcard search function was applied, and the instances were added to the final frequency count, disregarding the extraneous elements contained in the vagueness tags. The frequencies of items differing only in terms of number (e.g. *and that kind of thing/s*) or orthographic form (e.g. *kind of/kinda*) were conflated.

It should be pointed out that some of the choices made in this stage of in-depth analysis are inevitably subjective in nature. In some cases, it was very difficult to decide whether an expression had been retrieved by a speaker to fulfill the pragmatic function of vagueness. One of the consequences originating from the fragmented nature of speech is that sometimes phrase boundaries are obscured, making the analysis difficult. For example, in the extract *yeah there's yeah good hamburger places and something like that in er you can buy french fries* (Hy-Talk) it is difficult to ascertain where the phrase boundary separating the vague expression (i.e. *and something like that*) from the rest of the sentence exactly is. However, the definitions for vagueness tags and the two markers of imprecision *kind of* and *sort of* are nevertheless firmly established by earlier research (e.g. Overstreet and Yule, 1997; De Cock, et al., 1998; Altenberg, 1998; Biber, Conrad and Cortes, 2004), and the manual identification process of these expressions in this research was largely influenced by these prior works, somewhat toning down the subjective element present in the manual analysis.

5.3 The frequencies of the identified vague expressions

This section addresses the first research question of how frequently vague expressions are used by Finnish EFL upper secondary school learners in the selected corpora. The corpora are also compared with each other to inspect whether the frequencies of the expressions differ notably in these corpora. To draw more generalizable conclusions, the normalized frequencies of vague expressions are compared to a similar corpus-based research conducted by De Cock, et al. (1998). This enables the detection of such patterns of usage in the speech of Finnish upper secondary school learners that possibly diverge from the ones observed in the speech of French young adult EFL learners and young adult native speakers of British English.

5.3.1 The absolute frequencies of vague expressions in the chosen corpora

Table 2 below lists the combined absolute frequencies and ranges of vague expressions that were identified in the FUSE and the Hy-Talk corpora.

Vague expressions	Frequency	Range
Two-word		
kind of	67	28
or something	39	22
and stuff	17	8
and everything	13	6
or anything	8	5
and that	5	4
sort of	4	1
and things	2	2
or whatever	2	1
and something	1	1
like that	1	1
Three-word		
and so on	8	8
and like that	6	6
something like that	5	4
and all that	3	3
stuff like that	1	1
that kind of	1	1
and that kind	1	1
and that stuff	1	1
Four-word		
or something like that	15	13
and stuff like that	12	10
and that kind of	4	3
maybe something like that	2	2
and all that stuff	2	1
and things like that	2	1
and something like that	1	1
and everything like that	1	1
but something like that	1	1
nor anything like that	1	1
Five-word vague expressions		
and that kind of stuff	5	3
and that kind of things	2	2
and this kind of stuff	1	1
and all kinds of that	1	1
and this kind of things	1	1
Total	236	58

Table 2. The combined absolute frequencies of the vague expressions in the FUSE/Hy-Talk corpora.

Table 2 above reveals that only a handful of formulaic vague expressions accounts for most of the instances in the selected corpora. Even after eliminating non-vague instances of *kind of*, such as questions in the form of *what kind of*, this two-word n-gram was still the most frequent expression

used for denoting vagueness with the frequency of 67 occurrences and the range of 28. Each of the vague expression sets of varying length includes a few distinguishably frequent phrases, ranging from two-word expressions such as *or something* and three-word tags such as *and so on* to lengthier ones such as *or something like that*. As expected, expressions belonging to the lengthiest category of five-word units turned out to be rare, with *and that kind of stuff* occupying the position as the most common five-word vague expression.

In terms of their structure, most of the identified vagueness tags used by Finnish EFL learners are built upon core frames which enable flexibly the omission or insertion of optional items. Examples of such productive frames are the construction *and that/this kind of things/stuff* as well as *and/or/but/nor something/everything/anything like that* in which the underlined words are the only constant constituents in the frames. The markers of imprecision *kind of* and *sort of* do not exhibit such structural variability, except the ones that are part of vagueness tags such as *this kind of thing*.

The apparent mismatch between the frequency and range of items such as *kind of, or something, and everything* and *and stuff* suggests that the frequent use of vague expressions could partly be a phenomenon determined by idiosyncratic speech habits. For example, *kind of* was used 13 times by one speaker, accounting for nearly one fifth of the total number of occurrences, while *or something* occurred six times in the speech of one examinee. However, some of the more frequent vague expressions such as the four-word items *or something like that* as well as *and stuff like that* are more evenly distributed across examinees, demonstrating that the use of vague expressions is not only a matter of individual speech habits. The individual perspective is considered more closely in section 5.5 where the suspected connection between overall oral proficiency and the use of vague expressions is discussed.

5.3.2 The corpus-specific frequencies of vague expressions

In this section, the frequencies of vague expressions in the corpora are compared between each other to see whether there are differences in the distribution of vague expressions. Since the Hy-Talk corpus is larger than the FUSE corpus, the relative frequencies of phrases in the two corpora must be established according to standard normalization procedures so that the figures become comparable. Table 3 below reports the absolute frequencies and normalized frequencies[1] (henceforth abbreviated to F_n) of identified vague expressions in the examined corpora.

[1] Absolute frequencies were normalized per 10,000 words according to the convention followed in studies investigating smaller corpora. Thus, for instance, the normalized frequency of, say, the item *or something* in the Hy-talk corpus tells us that the item appears relatively in the data 7,08 times per 10,000 words, whereas the same item appears relatively less frequently, 5,39 times per 10,000 words, in the FUSE corpus.

Vague expressions	Hy-Talk	F_n	FUSE	F_n
kind of	54	13,65	13	6,37
or something	28	7,08	11	5,39
and stuff	12	3,03	5	2,45
and everything	12	3,03	1	0,49
or anything	8	2,02	0	0
and that	5	1,26	0	0
and things	2	0,51	0	0
and so on	3	0,76	5	2,45
and like that[2]	5	1,26	1	0,49
and all that[3]	3	0,76	0	0
that kind of	1	0,25	0	0
and that kind	1	0,25	0	0
and that stuff	1	0,25	0	0
something like that	4	1,01	1	0,49
stuff like that	1	0,25	0	0
or something like that	11	2,78	4	1,96
and stuff like that	7	1,77	5	2,45
and that kind of	4	1,01	0	0
and all that stuff	2	0,51	0	0
and something like that	1	0,25	0	0
maybe something like that	2	0,51	0	0
and everything like that	1	0,25	0	0
nor anything like that	1	0,25	0	0
and things like that	2	0,51	0	0
and that kind of stuff	4	1,01	1	0,49
and that kind of things	2	0,51	0	0
and all kinds of that	1	0,25	0	0
and this kind of things	1	0,25	0	0
or whatever	0	0,00	2	0,98
but something like that	0	0,00	1	0,49
like that	0	0,00	1	0,49
and this kind of stuff	0	0,00	1	0,49
sort of	0	0,00	4	1,96
Total	**179**	**45,24**	**56**	**27,45**

Table 3. The absolute and normalized frequencies of the vague expressions in the Hy-Talk corpus compared to the FUSE corpus

[2] There was one instance of *and also like that;* this was included in the frequency count of *and like that*.
[3] The data included one instance of *and all that's that.* The intervening part *that's* was interpreted as a slip and the instance was included in the frequency count of *and all that*.

As the total normalized frequencies reported in Table 3 show, vague expressions occur approximately 65% more frequently in the Hy-Talk corpus compared to the FUSE corpus. Many vague expressions occurring only a few times or once in the Hy-Talk corpus such as *and that kind of, and things like that* and *and all that* do not occur at all in the FUSE corpus. Since the FUSE corpus is smaller than the Hy-Talk corpus, the absence of such infrequent items could be explained by the size differences between the two corpora. Moreover, a few fairly common expressions such as *and stuff like that* and *and so on* occur more frequently in the FUSE corpus than in the Hy-Talk corpus. Still, some of the more frequent vague expressions such as *and everything* and *or anything* are conspicuously more common in the Hy-Talk corpus than in the FUSE corpus. The Hy-Talk corpus also includes 28 different types of vague expressions in total compared to 15 types occurring in the FUSE corpus, suggesting that these items are used more multifacetedly by the examinees taking the Hy-Talk test.

The attempt to explain discrepancies between the frequencies reported in Table 3 led to some interesting observations. There are a few crucial points to be made here in relation to the corpus data. As was previously mentioned, the FUSE corpus currently includes two types of tasks: open-ended and structured dialogues. Approximately 70% of the corpus consists of open-ended dialogues which have also produced lengthier transcripts on average than the ones elicited from structured dialogue tasks. After categorizing the individual texts included in the FUSE corpus into two sub corpora according to the dialogue type they contain and normalizing the frequencies of vague expressions according to the lengths of the texts, it can be concluded that vague expressions are used almost three times more frequently in the open-ended dialogues than in the structured dialogues[4].

The set of tasks in the Hy-Talk test included both structured and open-ended dialogues as well as a short warm-up interview and a monologue task. Still, the identified vague expressions are distributed across all of the included tasks, although the two more open-ended dialogues (task 3 and the extra task, see Appendix B for details) do include roughly half (N=90) of the identified vague expressions in the corpus. Considering that certain vague expressions have been found to be particularly emblematic feature of informal and congenial conversation (Overstreet and Yule, 1997, p.252; Aijmer, 2002, p.209), the high frequency of these units particularly in the open-ended dialogues is not that surprising. Although more profound consideration of the task-type effect on the use of these expressions lies beyond the scope of this work, future inquiries could gain invaluable

[4] The normalized frequency of all the vague expressions in the FUSE texts containing structured dialogues is 11,86 while the normalized frequency of all the vague expressions in the open-ended dialogues is 33,79 per 10,000 words.

insights by addressing in a more detailed manner how different task designs elicit vague expressions or other types of formulaic phrases.

Based on the frequency-based data examined during the course of this section, it may nevertheless be concluded with reasonably reliable certainty that formulaic vague expressions are present in the speech of Finnish EFL learners to a considerable extent. The fact that the items occur in both of the examined spoken language corpora adds further weight to these conclusions. The evidence above also confirms that Finnish EFL learners use mostly the same types of vague expressions identified by prior research on formulaic language (e.g. Erman and Warren, 2000; Biber, Conrad and Cortes, 2004; Altenberg, 1998; De Cock, et al., 1998). In the next section, some interesting quantitative comparisons are made with prior research to make the results presented in this section more informative.

5.3.3 The frequencies of vagueness tags in the speech of Finnish EFL learners compared to prior research

Comparing a learner spoken language corpus to a native corpus of similar size, De Cock, et al. (1998) reported a "highly significant underuse" of vagueness tags among 19-25-years-old learners compared to native speakers of British English of same age. The combined size of the FUSE corpus and the Hy-Talk corpus (59,973 words) is smaller than the native spoken language corpus (80,448 words) yet almost the same size as the non-native corpus (62,975 words) investigated in the aforementioned study. However, it should be pointed out at this point that the genre represented in the corpora investigated by De Cock, et al. (ibid.), consisting of informal interviews, is not entirely comparable to the spoken language tests transcribed for the corpora examined in the current study.

Table 4 below shows the absolute frequencies of various vagueness tags reported by De Cock, et al. (1998) and the combined absolute frequencies of the items analyzed in this research. Also the normalized frequencies[5] are provided in the adjacent columns to make the frequencies comparable between these corpora of different sizes. The abbreviation *NSC* refers to the native-speaker corpus and *NNSC* to the non-native speaker corpus investigated by De Cock, et al. (ibid.).

[5] The normalized frequencies were not reported in the research by De Cock, et al. (1998). These figures were added to the table by the author of this paper.

Vagueness tag	NSC	F_n	NNSC	F_n	Hy-Talk/FUSE	F_n
and all	0	0	0	0	0	0
and everything	21	2,61	4	0,64	13	2,17
and so on	2	0,25	18	2,86	8	1,33
and stuff	12	1,49	0	0	17	2,83
and stuff like that	15	1,86	0	0	12	2,00
and that	2	0,25	0	0	5	0,83
and that sort of thing	3	0,37	0	0	0	0
and things	31	3,85	1	0,16	2	0,33
and things like that	14	1,74	3	0,48	2	0,33
or anything	14	1,74	0	0	8	1,33
or something	30	3,73	4	0,64	39	6,50
or something like that	8	0,99	8	1,27	15	2,50
or whatever	11	1,37	1	0,16	2	0,33
something like that	1	0,12	4	0,64	5	0,83
sort of thing	5	0,62	0	0	0	0
stuff like that	4	0,50	0	0	1	0,17
that sort of thing	1	0,12	0	0	0	0
things like that	7	0,87	4	0,64	0	0
Total	181	22,50	47	7,46	129	21,51

Table 4. Absolute and normalized frequencies for the vagueness tags reported by De Cock, et al. (1998, p.77) and for the identified vagueness tags in the Hy-Talk/FUSE corpora.

The normalized total frequencies shown in Table 4 above indicate that Finnish EFL learners use these particular vagueness tags almost three times more frequently than French advanced EFL learners in the data examined by De Cock, et al. (1998). Significantly, the normalized total frequencies reported above also reveal that these particular vagueness tags occur in the speech of Finnish EFL learners almost to the same extent as in the speech of native speakers of British English (ibid.). Some of the vagueness tag types that were either entirely or almost absent in the speech of French EFL learners (ibid.) such as *and stuff*, *and stuff like that* and *or something* are even more frequent in the speech of Finnish EFL learners than in the output of native British English speakers (ibid). Surprisingly, vagueness tags such as *or something* and *or something like that* are even more frequent in the speech of Finnish pupils than in the output of British English speakers (ibid). The only item almost absent from the speech of natives but occurring in the non-native data sets is the tag *something like that*.

According to De Cock (2004, p.236), learners often favor vague expressions such as *and so on* and *et cetera* which are more typical for formal speech (Overstreet and Yule, 1997). In the case of French learners, this is partly confirmed by figures reported in Table 4 which shows that the phrase *and so on* is by far the most frequent type of vague expression used by French learners (De Cock, et al., 1998). While this expression is also represented in the FUSE and the Hy-Talk corpus, Finnish EFL learners seem to use more frequently vague expressions which are more typical for informal speech, including phrases such as *or something like that* and *and stuff like that*. The same pattern is observable in the preferences of natives (ibid.), adding further support for the notion that Finnish EFL learners prefer native-like ways to express vagueness.

Some of the vagueness tags appearing in the British English native corpus such as *that sort of thing* and *sort of thing* did not have any occurrences in the corpora analyzed in this study or in the non-native corpus examined by De Cock, et al. (1998). As natives are assumed to have a considerably larger stock of prefabricated phrases in general (e.g. Pawley and Syder, 1983), the more even distribution of occurrences between various items in the native corpus could be an indication of a more extensive stock of various vagueness tags among natives compared to non-natives, who might resort to a more limited set of phrases. However, this observation should be corroborated in subsequent studies utilizing directly comparable corpora since the setting and the genre might impact at least partly the way speakers utilize these items in speech. Nevertheless, most of the more frequent vagueness tags used by native speakers are evidently present in the speech of Finnish EFL learners at least to some extent. One aspect that the current methodology relying on frequency information cannot account for are approximate (or perhaps erroneous) forms that do not occur above the threshold level in the initial n-gram search.

5.4 Vague expressions and their observed pragmatic functions

In this section, the focus is shifted to a more qualitative consideration of how Finnish EFL learners use vague expressions in the observed data. Of key interest is whether the functions observable in the data differ somehow from the ones identified by earlier research on native spoken language. The identified functions for vagueness tags are discussed in their separate subsections below. Moreover, a separate subsection is dedicated to the functions identified in the analysis of the two markers of imprecision, *kind of* and *sort of*.

5.4.1 Establishing common ground and rapport

Prior research on vague expressions (e.g. Aijmer, 2002; Overstreet and Yule, 1997) has paid attention particularly to the important social functions these items serve in interpersonal communicative speech, significantly contributing to the informal ethos of conversations as well as strengthening common ground between speakers. This social aspect related to the use of vague language can also be discerned from the speech of Finnish EFL learners who evidently utilize vague language particularly in those tasks which have a more open-ended design, and which encourage the speakers to co-construct discourse in a creative manner. The following excerpts show some examples of the way speakers utilize vagueness tags to attend to social concerns pertinent to the current speech events:

(1) Speaker 1: let's just really you know have fun and (xx)
Speaker 2: (xx) yeah
Speaker 1: not counting calories or anything just seriously eat whatever we want
(Hy-Talk)

(2) in Holland there's what kind of freaks and that you know weirdos sick people
(Hy-Talk)

(3) Well, yeah I've seen your iPad and it has good games and stuff like that but I like to read my newspaper like . . . in a paper.
(FUSE)

(4) Yeah. Erm I also think that erm... er like free time. . . like all kinds of like going to movies and doing sports and that kind of stuff... especially with friends... and also alone sometimes... er like... er it really has a positive impact on your mind...
(FUSE)

In Example 1, the attempt to establish common ground is fairly evident by the strikingly informal word choices and adverbial elements. In this context, the use of the vagueness tag *or anything* is well in line with the overtly informal mode, making the communicative situation seem even more laid-back and co-operative. Adhering to the notion made by Overstreet and Yule about typical informal speech events (1997, p.255), vagueness tags used in these kinds of contexts often co-occur with the phrase *you know* as may be seen in Example 2. In this particular example, a similar kind of social function may be discerned as in Example 1. Moreover, the use of the formulaic vagueness tag *and that* is clearly utilized to attend the hearer as well, making the line of thought easier to follow for the hearer. The use of vague language contributes for its part to make the exchanges seem more natural, relaxed and informal. Example 3 shows how the tag *and stuff like that* is utilized by the

speaker to refer to shared common knowledge about the use of tablet computers, thus strengthening common ground between the speakers. Although the task itself is of a more closed-ended nature, dealing with an argumentation about the fate of the newspapers, the vagueness tag more closely associated with informal communication brings an air of relaxedness and informality to the communicative situation, somewhat softening the confrontational nature of the task. Often vagueness tags are used multiple times during one turn, as may be seen in Example 4 which is a part of discussion dealing with the open-ended topic of free-time opportunities. Again, the prolific use of vague language partly makes the exchange seem more communicative and informal.

5.4.2 Economy of processing

Corresponding with findings of previous research (See section 2.4.1), vagueness tags are also used by Finnish EFL learners to present information in a more succinct manner, facilitating the production of speech. In this kind of usage, the speaker first introduces one example of a category which may be more or less transparent from the context and completes the phrase (usually a noun or verb phrase) with a vagueness tag. This has the corollary effect of saving processing time for both the hearer and listener, corresponding with one of the key benefits of formulaic language (and vague expressions) established by earlier research. The use of vagueness tags for these functions makes the speech seem more economic and to the point, as illustrated by the following samples retrieved from the data:

(5) . . . i think you should listen to P-M-M-P i like it very much they have very funny lyrics and stuff

(Hy-Talk)

(6) well you you can take a little bit money if you want to buy something but we can erm buy your food for example ice cream and so on

(Hy-Talk)

(7) Speaker 1: Er... I think... it hasn't changed a lot

Speaker 2: Yeah yeah

Speaker 1: There has... been Angry Birds and stuff like that but

Speaker 2: Yeah

Speaker 1: that doesn't affect much

Speaker 1: Er like er i- it's because Estonians used to watch Finnish TV in the er times of

Speaker 2: Aah

Speaker 1: of CCCR and

Speaker 2: Yeah

Speaker 1: and stuff like that so

(FUSE)

(8) Yeah erm and yeah I totally agree with that er we have er very advanced educational system here in Finland... and er... of course we ha- we already know that because have studied for like... I don't know twelve years or something... er... and er... it's pretty sad to think that in like third world countries... not so many kids get to go- go to school even like elementary school... er when at the same time we have the opportunity to... study... till we're like adults... for free and er even when we go to university we don't have to pay for it... at least not yet... and er... e- everything that wou- on- only things we have to pay for when we study is like f- food and... living and this kind of stuff

(FUSE)

(9) it's it's a show . . . where they have to win the competition they have to do so many stuffs to show theirse- themselves and to make people like them by making some stupid things and this kind of things and everything so

(Hy-Talk)

As the above examples show, phrases such as *and stuff* and *and so on* are used as summarizers following an example or examples of a specific category. In these cases, vagueness tags are utilized to indicate that more information or examples of entity X could be provided for the hearer, but these are deliberately left out to save the effort. Thus, in Example 5, the tag *and stuff* implies that the speaker could provide more reasons to listen to the aforementioned band besides the entertaining lyrics the songs contain, but the speaker chooses not go into more details for the sake of economy and also because the speaker assumes the hearer is able to deduce the pragmatic function of the tag. Likewise, in Example 6, the hearer is encouraged to conclude that there are other compelling food items to buy in addition to ice cream. Consequently, these kinds of usages save processing effort from the perspective of both the speaker and the hearer given that the hearer understands the pragmatic function of the retrieved vagueness tag and is also able to deduce the semantic hyperonym category the preceding noun, verb or other word type is alluding to. Again, besides saving processing effort, the use of vagueness tags obviously necessitates some form of common ground between the speakers in order to fulfil the desired pragmatic effect (e.g. Overstreet and Yule, 1997, 255-256). When this understanding is achieved, the rapport between the speakers is further enhanced.

The saving of processing effort is similarly evident in Example 7 where the item *and stuff like that* occurs in a stretch of speech involving an open-ended discussion about the public image of Finland. At the same time, it could also be argued that the use of such prefabricated phrases makes the conversation seem more informal, affecting positively to the relaxed atmosphere of the

conversation. One recurrent phenomenon in the data is that speakers repeat either the vagueness tags they have uttered themselves, as may be seen in Example 7, or the ones uttered by their interlocutor. The vagueness tag repeated in Example 7 at the end of the extract by Speaker 1 constitutes a full turn and could be interpreted as a sort of filler by which the speaker manages to buy time to encode the next thought into words. Although the question is slightly off the mark outlined for this research, future works should examine more closely how much the use of vague expressions affects the use of interlocutor's use of those same expressions. The quantitative consideration of vagueness tags from the individual point of view already proved to some extent that individuals tend to cling to those expressions which they have been found effective for the current task.

Although the majority of the vagueness tags in the observed data occur in the more open-ended tasks where individual turns tend to be rather short, Finnish EFL learners use surprisingly many vagueness tags also in lengthier individual contributions. For instance, the monologue task (see Task 1 in the Appendix B) in the Hy-Talk corpus contains 33 instances of vague expressions, indicating that vagueness tags are by no means limited to short bursts of speech. Thus, vague expressions evidently form a significant prefabricated building block of discourse which enable Finnish EFL learners to make their speech production more fluent also in lengthier contributions. In Example 8, the speaker utilizes two tags in a lengthier turn, *or something* and *and this kind of stuff*. The use of such prefabricated chunks facilitates the production of speech from the speaker's perspective, relieving the speaker from the effort to give unnecessary details about a point that has already been made, giving more time for formulating the next thought into words as well as providing a way to distance oneself from the contents of utterances (Aijmer, 2002). Example 9 shows another illustration of how these items are often used in conjunction, enabling the speaker to think about what to say next.

5.4.3 Approximation

A further important pragmatic function of vague expressions relates to one of the universal features of language which is particularly emblematic for the spoken language mode where preparation time for the production of utterances is often limited. Thus, speakers have the need for linguistic devices by which they may express imprecision and distance themselves from the premises implied by utterances. To attend to these universal concerns in communication, vagueness tags are a particularly useful linguistic device, although not the only available option for speakers to carry out such pragmatic functions (e.g. Altenberg, 1998). As pointed out by prior research delving into the

functions these items perform (See section 2.4.1), vagueness tags are also used by Finnish EFL learners to hedge utterances and to indicate that the propositions are only approximations and somewhat imprecise in nature. By far, the most commonly used item by Finnish EFL learners to achieve this function is the phrase *or something* with its variant forms (e.g. *or something like that*). Examples include the following excerpts from the data:

(10) ...it's not a dream but it's more like a fee- er feeling, this sixth sense or something
(FUSE)

(11) it’s probably about i don’t know 29 or something
(Hy-Talk)

(12) er big brother house it must be a big sho- shock or something like that for the er when he the baby is young boy
(Hy-Talk)

(13) Speaker 1: for the food and other so it will cost about 35 euros i think
Speaker 2: yeah something like that
(Hy-Talk)

In Example 10, the phrase *or something* is utilized by the speaker to indicate that the description given about the feeling is not entirely accurate but there might be more to it than can be put into words straight away. The last noun phrase in the excerpt is preceded by slight hesitation about the word choice (i.e. *fee- er feeling*) which nevertheless is circumvented by applying vague language to give a more imprecise description of the phenomenon (i.e. *this sixth sense or something*). In Example 11, the phrase *or something* is combined with the phrase *I don’t know* denoting epistemic uncertainty, further underlining the effect of the vagueness tag. Indeed, vagueness tags may occasionally be used for the sincere purpose of indicating a lack of knowledge about the truth conditions expressed by utterances. This does not directly relate to any uncertainties about specific linguistic choices but about the contents of the claims themselves, reflecting the inaccurate nature of spoken language in general (e.g. Altenberg, 1998). In Example 12, the vagueness tag is preceded as well as followed by hesitation markers and slight stuttering, suggesting that the vagueness tag is, in addition to indicating approximation, used by the speaker to buy time for the formulation of the next utterance.

Identifying a distinct function for formulaic items is not always straightforward. Biber, Conrad and Cortes (2004) concluded in their study on common lexical bundles that many of these multiple-word items may have more than one pragmatic function in discourse. As was already noted above,

many vagueness tags in the data do not in fact exhibit only one distinct function but two or even several. Example 13 shows how some vague expressions such as *something like that* may occasionally serve a secondary function of response to a previous comment or a question, as in "yeah something like that" (Hy-Talk) or the expression *like that* in the response "Yeah like that" (FUSE), but these kinds of usages were relatively infrequent in the data.

5.4.4 Vagueness tags in conjunction with other markers of imprecision and further general patterns

The above excerpts already showed some examples of how these expressions co-occur with other formulaic devices such as epistemic tags in the form of *I don't know*. Although this research focuses on a very specific set of formulaic expressions, it should be pointed out that vagueness tags are indeed often accompanied with other devices expressing vagueness or impreciseness such as the adverbs *maybe* and *like*, as the following extracts obtained from the data show:

(14) i used to like it like series like survivors or something like that
(Hy-Talk)

(15) it's seven erm maybe nine something like that...
(Hy-Talk)

(16) ...and only thing I don't like to share is like blankets and stuff
(FUSE)

(17) er probably i'm gonna buy like tickets so they're gonna pay er cost about 20 euros or something and maybe an ice cream or something and games and stuff
(Hy-Talk)

Above examples illustrate the fact that formulaic vagueness tags form only one yet important node in the composition of imprecise language by which speakers may express themselves in an informal and less detailed way, reflective of the general features of spoken discourse. For their part, the use of tags evidently contributes positively to the overall flow of the communication and production of foreign language speech.

What is striking in the above examples is that in most cases stalling features such as stuttering and the presence of the hesitation signal *er* are absent in the immediate vicinity of vagueness tags (cf. Boers, et al., 2006). In addition to the important pragmatic functions they carry out, this could be seen as a further argument for the case that these items are indeed processed by speakers in a formulaic fashion, even though in some rare cases vagueness tags could be interpreted as repetitive

linguistic behavior rather than formulaic retrieval of prefabricated elements. For instance, in Example 17, the hesitation markers are present in the beginning of the turn, but they subside as the speaker retrieves the assumingly prefabricated vagueness tags at the end of the turn. Arguably, formulaic vagueness tags, together with other prefabricated material and vague language, construct "zones of safety" (Boers, et al., 2006, pp.246-247) for the speakers. In these zones, speakers may operate with smaller likelihood of slipping beyond the boundaries of their current proficiency levels (ibid.).

5.4.5 Markers of imprecision *kind of* and *sort of*

These two important markers of imprecision in speech, *kind of* and *sort of*, occur in slightly different surroundings at a clause level than vagueness tags and are not generally treated as vagueness tags in the literature, justifying the dedication of a separate subsection for the consideration of these phrases. The item *kind of* was already found to be very frequently present in the speech of Finnish EFL learners, being the most common vague item in the observed data. Overall, these markers of imprecision are used by Finnish EFL learners for very similar functions as vagueness tags, namely, to hedge information, express approximate thoughts and ideas in an imprecise manner, enhance the fluency of speech and add an informal element to exchanges. The following examples illustrate the usages of these phrases by the examinees:

(18) . . . if you're in a- some country where like some people are really poor and some people are quite rich... then maybe the rich people kind of like rule... and erm have more power than the poor people...

(FUSE)

(19) Speaker 1: my favourite er reality TV that i have watch is this dancing with the stars

Speaker 2: ok well i thought it was kind of boring why did you like it

(Hy-Talk)

(20) Speaker 1: Yeah just like a comedy, kind of but

Speaker 2: Mm, kind of comedy.

(FUSE)

(21) well it was kind of boring and there was some fancy explosions and stuff but i am not kind of action guy

(Hy-Talk)

Example 18 demonstrates how the item *kind of* is used in a very similar way as various vagueness tags discussed earlier. In these kinds of usages, the marker of imprecision presents the information

as an approximation and thus enables speakers to distance themselves from the ideas they communicate, bearing resemblance to one of the common functions identified in the case of vagueness tags. In this example, the phrase is yet again used in conjunction with other important elements of vague language, i.e. the words *like* and *maybe*. In Example 19, the speaker hedges the declarative statement expressing objection to the interlocutor's earlier comment by adding the phrase *kind of* to the utterance. This affects positively to social concerns since the marker of imprecision makes the comment seem more polite and softens the impact of the dissenting view. In Example 20, the phrase is used as an approximator by Speaker 1 very much in the same way as vagueness tags in some of the examples cited earlier. In fact, the same passage could contain the vagueness tag *or something like that* instead of the marker of imprecision *kind of* and still retain the same function. The phrase is picked up and repeated in the subsequent turn by the interlocutor as a way to approve of the proposition made by the Speaker 1, once again adhering to the social pragmatic function these expressions often have. Example 21 shows how two important features of vague language, vagueness tags and markers of imprecision are used in conjunction to enhance the vagueness of the message.

The inspection of these two markers imprecision in the speech of Finnish EFL learners offers some interesting insights when compared to earlier research. In addition to the list of vagueness tags shown in Table 4 above (See section 5.3.3 for details), De Cock, et al. (1998) reported a considerable underuse of these two markers of imprecision in the speech of learners compared to native speakers. Although the frequencies of these items were unfiltered in the aforementioned study (i.e. including also non-vague usages), making direct comparisons between their frequencies impossible, Table 2 (See section 5.3.1 for details) does suggest the underuse of the phrase *sort of* by Finnish EFL learners, with only four instances by one examinee. However, the phrase *kind of* is used even more frequently by the examinees compared to French learners or even native speakers of British English (De Cock, et al., 1998).

De Cock, et al. (1998, p.78) also reported that the marker of imprecision *kind of* is mostly preceded by verbs in the speech of natives whereas in the speech of non-natives they are often accompanied by noun phrases. Interestingly, the data investigated in this study indicate that Finnish EFL learners tend to use the phrase as a modifier for adjectives as Examples 19 and 21 show, accounting for roughly half of the instances in the corpora, while the other half modifies either nouns and noun phrases or verbs. Contrary to the observation made by De Cock, et al. (ibid.), according to which the phrase is more often used by learners to mend the lack of foreign language skills, there is only one such instance in the Finnish corpora where the phrase is used for this particular purpose. As

with the use of vagueness tags by the examinees, the use of markers of imprecision does not stem from an overt lack of linguistic knowledge but rather intentional use of prefabricated phrases that help speakers to contribute to the ongoing social discourse and communicate their ideas more fluently.

To conclude the qualitative observations, both vagueness tags and markers of imprecision occupy mainly the same kinds of functions in the speech of Finnish EFL learners for which such expressions are used by natives according to earlier literature (e.g. Erman and Warren, 2000; Aijmer, 2002; Overstreet and Yule, 1997): to make the exchange more informal and interpersonal, to denote strategic vagueness, and enhance the overall flow of discourse with both the speaker and hearer in mind. This suggests that the frequency of such phrases in the speech of Finnish EFL learners is not explained merely by idiosyncratic factors or lack of linguistic knowledge. On the opposite, they seem to form an integral albeit small element in the building of overall discourse and meaning especially in the tasks eliciting more natural linguistic behavior. Thus, they can be seen as one minor yet important detail in an individual's pragmatic knowledge and communicative competence. The next section probes more closely into the hypothesized connection between the use of vague expressions and oral proficiency.

5.5 The correlation between the use of vague expressions and exhibited oral proficiency

This section investigates the individual perspective related to the use of vague expressions by Finnish EFL learners. Thus, an answer is provided for the third research question probing the correlation between the use of vague expressions and oral proficiency exhibited in the Hy-Talk spoken language test. To answer the question, the correlation between two variables were inspected: the overall grades examinees had received for their performances and the observed use of formulaic vague expressions by individuals, taking into consideration both the exhibited number and types of expressions in their separate subsections. Since the marker of imprecision *kind of* was cited by De Cock, et al. (1998) and De Cock (2004) as one particular feature of vague language that learners tend to underuse, this category of vague language was included in the analysis as well. As pointed out earlier, the FUSE corpus was not used for the purposes of answering the third research question.

5.5.1 The correlation between received overall grades and frequencies of vague expressions

Figure 1 below shows the absolute frequencies of vague expressions used by individual examinees in the Hy-Talk sub corpus[6] and the overall grades they received from the test.

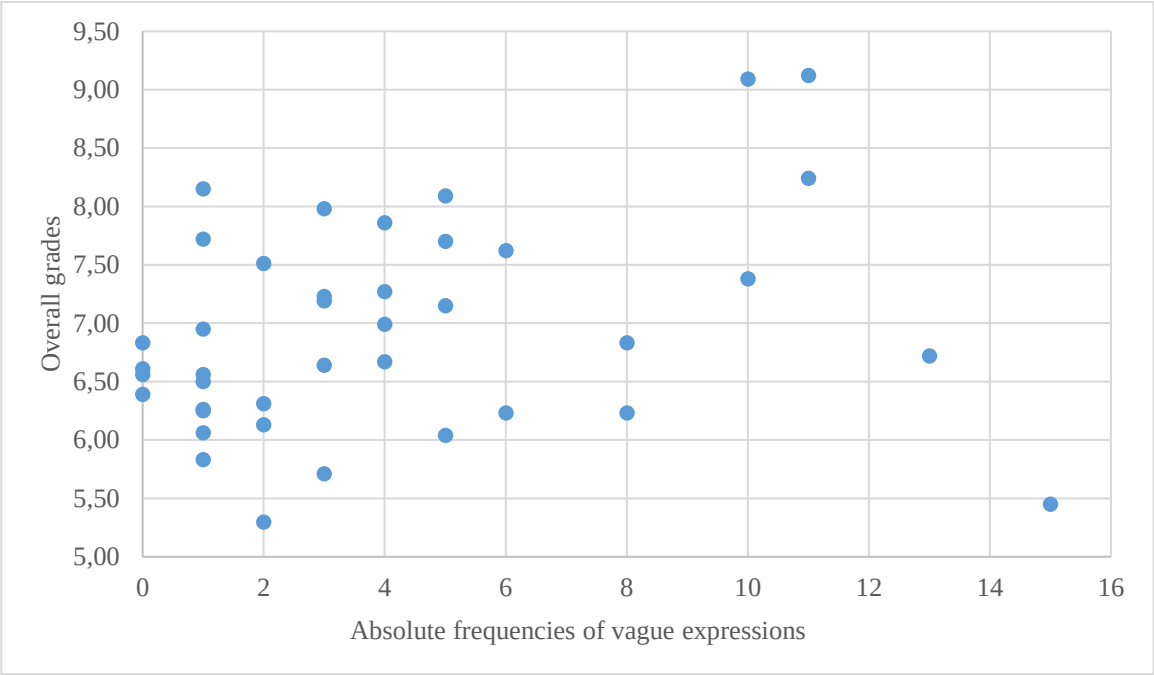


Figure 1. The overall grades and the absolute frequencies of vague expressions per individual

As Figure 1 indicates, most individuals out of the pool of 40 examinees are situated somewhere along the scale of 1-5 in terms of the numbers of formulaic vague expressions utilized in the spoken language test, the average absolute frequency amounting to 4,15 expressions per performance. The top-three performers use considerably more frequently vague expressions than examinees on average, each of them utilizing 10,5 such phrases in the test. Figure 1 above shows that there is considerable dispersion in the total number of vague expressions in the pool of 40 examinees, ranging from zero to 15 usages by one individual. In total, the data includes four individuals who do not exhibit any use of such phrases in the test.

The mean overall grade among the examinees is 6,93. Figure 1 indicates that some less well performed individuals used considerably more frequently vague expressions compared to some examinees who received better overall grades from the test. For instance, the individual who has

[6] As mentioned earlier, the sections containing the warmup interviews were removed from the corpus for the purposes of inspecting the correlation between oral language proficiency and the use of vague expressions.

received the second lowest overall grade (5,45) exhibits the most frequent use of vague expressions with 15 tokens. With the overall grade of 6,72, the second most prolific user of vague expressions is also situated slightly below the average score.

As the sizes of the individual transcripts of performances varied substantially, the absolute frequencies shown in Figure 1 were normalized per 1,000 words to minimize the effect of text lengths to the results. The squared correlation coefficient was inspected to investigate the correlation between these two variables, i.e. the grade and the normalized frequency of vague expressions. The trendline for the correlation pattern is observable in Figure 2 below:

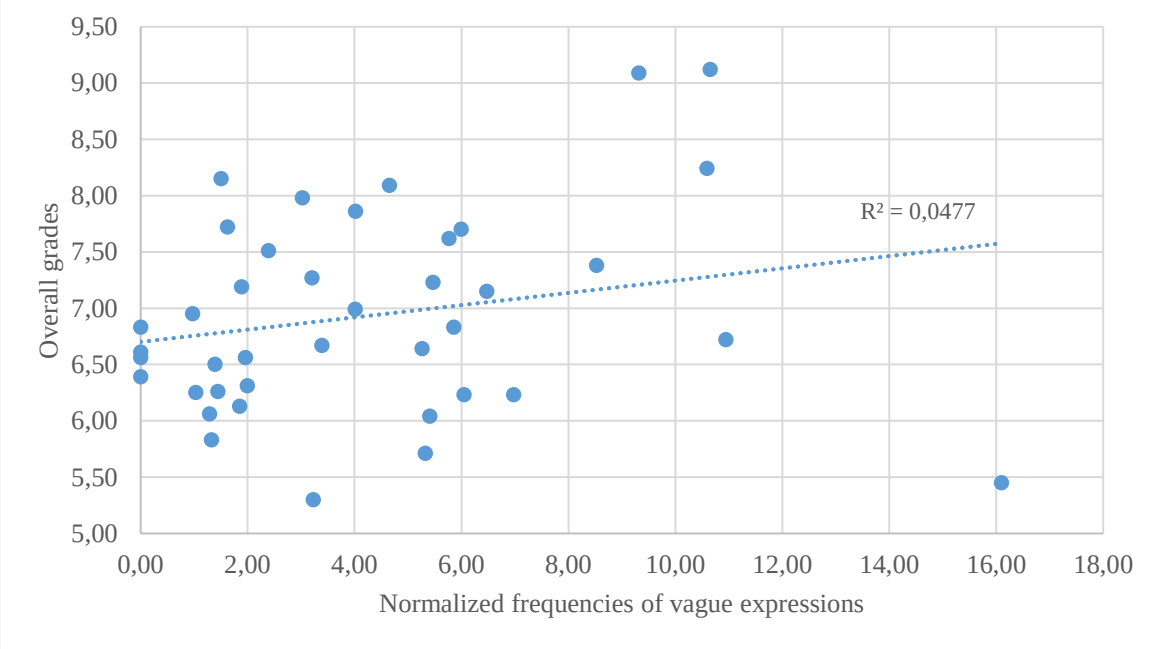


Figure 2. The overall grades of examinees and the normalized frequencies of vague expressions

The trendline shown in Figure 2 indicates that higher grade correlates only marginally with more frequent use of vague expressions, explaining 4,77 % of the variance. The mean normalized frequency of vague expressions occurring in the top-10 performances is 5,7. This is slightly higher than the mean normalized frequency of 4,3 which corresponds to the ascending trendline presented in Figure 2. However, the group of top-ten performers includes two individuals who use only one formulaic vague expression in their entire performance. Moreover, the mean normalized frequency of vague expressions for the top 10 lowest performances is approximately 4,9 expressions, slightly above the mean frequency of all the examinees. To conclude the investigation into the connection between more frequent use of vague expressions and higher perceived oral proficiency, the statistical significance of the result was tested. The analysis revealed that there is not a significant

positive relationship between the number of vague expressions used in the test and the overall grade the examinees received from the test, $r(38) = .1757$, $p > .05$.

5.5.2 The correlation between received overall grades and types of vague expressions

Figures 1 and 2 presented above show only the absolute and normalized numbers of vague expressions used by the individuals regardless of whether they constitute only one or several types of items. For example, closer examination on the most prolific user of vague language revealed that this particular individual used the same marker of imprecision *kind of* 13 times in the data. Since a large word stock is, in general, considered a sign of language proficiency, the consideration of the number of different types used by individuals was also included in the analysis to see whether there is a correlative element present in the number of different expressions utilized by examinees and the grades they received from the test.

Figure 3 below shows the number of different types[7] of formulaic vague expressions used by examinees in the test and the grades they received for their performances.

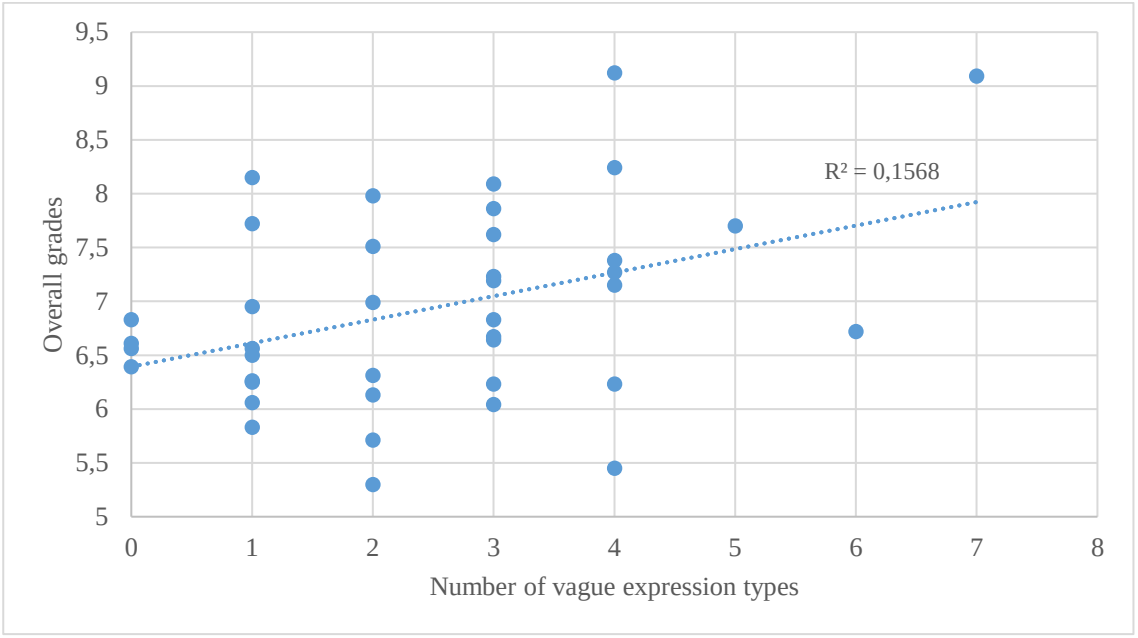


Figure 3. The overall grades of examinees and vague expression types

On average, the examinees use approximately 2,5 types of vague expressions in the test. Again, the top 10 performers exhibit somewhat larger stock of different types of phrases, with the average of 3,3 different types of expressions per performance while the top-10 lowest performers exhibited the use of 2,3 expressions per individual. As may be observed in Figure 3, testing the link between

[7] Again, vague expressions such as *stuff like that* and *and stuff like that* are treated as separate types of expressions.

these two variables produced a visible trendline with the R-squared value of 0,1568. The statistical significance of the result was tested. There is a significant positive correlation between the two variables, $r(38) = .011$, $p < .05$. Thus, examinees who received higher grades exhibited more multifaceted use of vague expressions in the Hy-Talk test. Although the relatively small P-value does suggest that the more proficient speakers did not exhibit more multifaceted use of formulaic vague expressions merely by chance, the test does not directly provide evidence for any causality. Furthermore, an inquiry into the word stock of these individuals – regardless of whether we are dealing with one-word or multiple-word units – would likely correlate with higher grades as well. Still, the claim that more proficient speakers have a larger stock of formulaic vague expressions is perfectly in line with what we know about formulaic language and its connection to greater proficiency and fluency. This result should be corroborated with further studies utilizing larger data sets.

6 Discussion

6.1 Formulaic vague expressions in the speech of Finnish EFL learners

Formulaic language has, in general, been cited as one of the major challenges for EFL learners yearning to acquire near native-like competence (e.g. Wray, 2002; Pawley and Syder, 1983). One of the explanations for this has been that non-natives lack the access to the large native reservoir of prefabricated phrases which are effortlessly applied by natives in recurrent situationally defined linguistic needs. Instead of resorting to the idiom principle (Sinclair, 1991), facilitating the appropriateness and fluency of language production, they cling more likely to analytic speech production strategies (e.g. Conklin and Schmitt, 2012) which makes their language sound more unorthodox and less suitable for the context.

The present study, addressing the topic particularly from the perspective of Finnish EFL learners, has presented evidence that runs somewhat contrary to this dichotomy between native and non-native use of formulaic language. The numerical data obtained from the analysis of two spoken-language corpora has revealed that Finnish EFL learners make extensive use of at least one specific type of formulaic language, i.e. vague expressions, in a set of spoken language tasks containing both structured and open-ended tasks. Comparisons to frequencies reported in the study by De Cock, et al. (1998) revealed that vagueness tags are by a large margin more common in the examined Finnish EFL data than in the French EFL spoken language data. Most significantly, Finnish EFL learners use formulaic vagueness tags almost to the extent observed in the speech of young native speakers of British English (ibid.). Still, as mentioned earlier, these observations must be dealt with some caution because the corpora investigated by De Cock, et al. (1998) and the ones viewed here are not directly comparable.

The qualitative consideration of the functions these expressions fulfil in the speech of Finnish learners adds further support to the conclusions made about their frequency compared to native use. In most cases, vague expressions are not used by Finnish EFL learners to mend problems resulting from lack of linguistic proficiency as reported by De Cock, et al. (1998, p.78) in the case of French EFL speakers. Instead, the functions for which they are used correspond more closely with the functions reported by earlier research on native data (e.g. Overstreet and Yule, 1997; Altenberg, 1998; Erman and Warren, 2000), offering commonplace advantages associated with formulaic language in general such as attending to social dimensions of speech events, offering processing short-cuts, allowing speakers more time to formulate their subsequent utterances and manipulating

information (Wray, 1999, pp.473-477). When one considers both the frequency of vague expressions as well as their multifaceted functions, the high presence of these items in the examined data might actually be an indication of high proficiency among Finnish learners of English.

These observations regarding the salient presence of formulaic vague expressions in the speech of Finnish EFL learners leads to an interesting question: why does this particular group of learners use such language so extensively compared to French learners? Perhaps explaining the low frequency of such language in certain contexts, one of the suggestions made in the literature (De Cock, et al., 1998, p.78; Overstreet and Yule, 1997, p.250) is that such language has sometimes been somewhat paradoxically stigmatized both in the language class and native speech, even though the use of such expressions has a solid foothold in natural everyday communication. It is more than likely that times have brought changes into how these kinds of phrases are perceived compared to the days of the late 1990s, and such language has, perhaps rightfully so, been admitted to the language class as an essential part of the spoken language discourse. Because different languages might also have different norms regarding the way impreciseness is expressed, the varying L1 backgrounds of learners might also affect the way how certain kinds of linguistic devices are used in a particular linguistic environment.

It could also be that the use of these expressions is still not a result of explicit instruction, but students bring their own linguistic influences into the class and use them creatively with their peers to enhance the social aspects of spoken communication. But if so, from which sources do learners acquire these expressions if the access to such language use has not been part of the official curricula until now? One of the observations made in the qualitative analysis of the data might provide partial answer to the question. The analysis revealed that Finnish EFL learners occasionally recycle formulaic vague expressions used both by themselves and by their interlocutors. Indeed, formulaic expressions seem to perform important social routines such as creating rapport between speakers in addition to offering a shortcut for addressing commonplace needs and desires in conversation, with repetitive linguistic behavior only enhancing the effect. This is consistent with the remarks made by Aijmer (2002) about the social importance of vague language. A certainly viable learning route for formulaic language, as for language is general, is mere exposure to such elements, whether it be through various media or by hearing such expressions used by peers. This is in line with the observation made by Kecskes (2016), according to which linguistic communities are inclined to begin to generate their own preferred ways of expressing themselves. Thus, the most effective way to acquire formulaic language may simply be exposure to such language in a repeated fashion. The current study did not cross examine the pairs to find out whether the pairing had an

effect on the frequency of formulaic items, but this would certainly be an issue worth investigating in future works.

Since the use of formulaic language and vague expressions have been linked to oral proficiency in earlier research (e.g. De Cock, et al., 1998; Boers, et al., 2006), the current research sought to find out whether oral proficiency correlates with more frequent and multifaceted use of vague expressions in the selected data. The data did preliminarily suggest that both the number and types of vague expressions exhibited by learners correlate with higher grades in a spoken language test. However, statistical analysis did not establish significant link between the number of vague expressions used in the test and oral proficiency. It is likely that individuals, usually having one or a few types of vague expressions in their repertoire, tend to overuse those phrases as was seen in the case of one individual who received the lowest grade of all in the test. This element of repetitive use of formulaic elements by non-natives corresponds with the observations made by De Cock (2004). In the case of different types and perceived oral proficiency, statistical analysis confirmed the significance of the correlation. This is not surprising, considering that larger phrasal reservoir is a typical feature of proficiency (e.g. Pawley and Syder, 1983). The test examining correlation between these variables should be repeated in further studies, factoring in a more versatile types of formulaic expressions in consideration of their effect on perceived oral language proficiency. Preferably, the analysis would also include other skill constructs such as pronunciation, fluency and appropriateness of expression.

6.2 Limitations and suggestions for further topics

Even though the observations made about the frequency of vague expressions do offer evidence about the salient presence of formulaic language in the speech of Finnish EFL learners, it must be acknowledged that the study has several limitations. First, the chosen method might have a considerable impact on the results. It may be claimed that the selected corpus-based method and adherence to certain predefined criteria does yield more trustworthy results than purely subjective analysis of the data. However, the current methodology applied in the study is still unable to address one of the most fundamental questions related to the study of prefabricated language: how we can ascertain which stretches of spoken language are produced by speakers by way of formulaic processing and which are not. After all, as pointed out by Erman and Warren (2000, p.33) the formulaicity of an expression might be dependent on individual and situational factors. Moreover, the detection of various pragmatic functions and the categorization of vague expressions lean partly on subjective interpretation of spoken language data and is admittedly fertile ground for

misinterpretations. For these reasons, the results presented in this study should be dealt with some caution.

Secondly, the current study has approached the field of formulaic language only from a narrow point of view. Knowledge about the use of vague expressions, and of formulaic language in general, would certainly avail of a wider perspective on the phenomenon. This could be achieved, for example, by addressing both qualitatively and quantitively how the formulaic items considered in this study co-occur with other categories of formulaic expressions such as discourse markers and hesitation markers in the speech of Finnish EFL learners. Particularly the connection between perceived oral language proficiency and use of formulaic expressions should be investigated by including a more inclusive set of formulaic language. A further limitation in the current study relates to the data consisting of transcribed oral test recordings. Since the corpora were not directly designed for the current purposes, the nature of the tasks included in the transcripts makes it impossible to make more generalizable claims about the way Finnish EFL learners use English in authentic contexts. Thus, a more selective approach to the inclusion of data would be needed in further inquiries. Moreover, the theme could be approached through a L1-sensitive comparative approach that would address why a specific L1 learner group uses certain kinds of formulaic expressions in a different way than learners with a different L1 background.

7 Conclusion

This corpus-based research has probed the use of formulaic language in the context of Finnish upper secondary school EFL learners, focusing specifically on one category of formulaic language, vague expressions. The first research question examined the extent to which Finnish learners utilize vague expressions. The analysis of two Finnish spoken language corpora, the Hy-Talk corpus and the FUSE corpus, revealed that vague expressions occur frequently in the speech of Finnish EFL learners. The set of vague expressions in the examined data corresponded mostly with the set of vague expressions identified by earlier research concerning both native and non-native spoken language. A comparison to a similar study also revealed that Finnish learners use vague expressions markedly more frequently than young adult French learners and almost to the same extent as native British young adults, although the set of vagueness tags represented in the Finnish corpora was slightly narrower than the one favored by natives.

The second research question sought to find out the functions for which these expressions are used by Finnish learners by inspecting qualitatively the co-texts of the vague expressions. The analysis revealed the same kinds of functions identified by earlier research on native data. Finnish speakers used these expressions mainly to attend to social aspects related to informal speech situations. Moreover, these expressions helped the speakers to express themselves in a more fluent and succinct manner, making the interpretation of utterances easier also for the hearer. It was concluded that the detected functions seem to support the view made about the frequency of these items in the speech of Finnish EFL learners even compared to natives. Thus, the use of vague expressions by Finnish EFL learners could be seen as a sign of proficiency indicating near native-like awareness of pragmatics and typical speech behavior in informal contexts.

Since the use of vagueness tags and markers of imprecision, and formulaic language in general, has been linked to linguistic proficiency by earlier research, the third research question sought to inspect the connection between the use of vagueness tags in the spoken language test contained in the Hy-Talk corpus and the grades examinees received from the test. It was concluded that higher number of vague expressions used in the test does not correlate with higher grades to a statistically significant extent. Even though the top performers evidently exhibited more frequent use of vagueness tags than the examinees in average, some pupils from the group receiving the lowest grades were also prolific users of vague language. However, the inspection of the number of types of vague expressions used in the test yielded a statistically significant relationship. Thus, based on the results, the individuals who used a wider set of formulaic vague expressions performed better in

the spoken language test. Although this confirmed a correlation between these two variables, it was concluded that future studies should include a wider set of different types of formulaic expressions to obtain more reliable results about the connection between the use of formulaic items and higher foreign language proficiency.

The results gained from the current research will hopefully lead to further inquiries into the roles different aspects of formulaic language play in the speech of Finnish EFL learners. One of the key observations made in the current study, also echoed by the existing literature, is that the benefits of formulaic language lie particularly in the important pragmatic roles they play in the spoken language, enabling speakers to formulate their thoughts more fluently and naturally. The spoken language has, in general, gained more and more attention during recent decades. The current study has added another contribution to the body of research that can also potentially offer applicable new pedagogic insights into the ways formulaic language, a crucial aspect of spoken language, can increasingly be considered in the foreign language teaching practices.

8 References

Primary sources

The HY-Talk Corpus, English subset. Department of Modern Languages/English and Department of Applied Sciences of Education, University of Helsinki, 2007-2009.

FUSE — The Finnish Upper Secondary School Corpus of Spoken English. Compiled by L. Ehrnrooth, 2018 -. [online] Available at: <https://fusecorpus.eu> [Accessed 09 August 2019].

Secondary sources

Aijmer, K., 2002. *English Discourse Particles: Evidence from a corpus*. Philadelphia: John Benjamins Publishing Company. Available through: Helsinki University Library website <http://www.helsinki.fi/kirjasto/en/home/> [Accessed 09 August 2019].

Aijmer, K., 1996. *Conversational routines in English: Convention and creativity*. London: Longman.

Altenberg, B., 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In: A. Cowie, ed. 2001. *Phraseology: Theory, analysis and applications.* Oxford: Oxford University Press. pp.101-122.

Altenberg, B. and Eeg-Olofsson, M., 1990. Phraseology in spoken English. In: J. Aarts and W. Meijs, eds. 1990. *Theory and Practice in Corpus Linguistics*. Amsterdam: Rodopi. pp.1-26.

Anthony, L., 2018. *AntConc* (Version 3.4.4w). [Computer Software]. Tokyo, Japan: Waseda University. Available at: <http://www.laurenceanthony.net/software> [Accessed 09 August 2019]

Biber, D. and Barbieri, F., 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes,* [e-journal] 26(3), pp.263-286. https://doi.org/10.1016/j.esp.2006.08.003.

Biber, D., Conrad, S., and Cortes, V., 2004. If you look at...: lexical bundles in university teaching and textbooks. *Applied Linguistics,* [e-journal] 25(3), pp.371–405. https://doi-org.libproxy.helsinki.fi/10.1093/applin/25.3.371.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E., 1999. *The Longman grammar of spoken and written English*. London: Longman.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., and Demecheleer, M., 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research,* [e-journal] 10(3), pp.245-261. http://dx.doi.org.libproxy.helsinki.fi/10.1191/1362168806lr195oa.

Cambridge Dictionary, 2019. *Vague expressions – English Grammar Today – Cambridge Dictionary*. [online] Cambridge: University of Cambridge Press. Available at: < https://dictionary.cambridge.org/grammar/british-grammar/useful-phrases/vague-expressions> [Accessed 09 August 2019].

Carey, R., 2013. On the other side: formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca*, [e-journal] 2(2), pp.207-228. https://doi.org/10.1515/jelf-2013-0013.

Conklin, K. and Schmitt, N., 2012. The processing of formulaic language. *Annual Review of Applied Linguistics*, [e-journal] 32, pp.45-61. https://doi.org/10.1017/S0267190512000074.

Conklin, K. and Schmitt, N., 2008. Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, [e-journal] 29(1), pp.72 – 89. https://doi-org.libproxy.helsinki.fi/10.1093/applin/amm022.

De Cock, S., 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, 2(1), pp.225-246. Available at: <https://dial.uclouvain.be/downloader/downloader.php?pid=boreal:75157&datastream=PDF_01> [Accessed 09 August 2019].

De Cock, S., Granger, S., Leech, G. and McEnery, T., 1998. An automated approach to the phrasicon of EFL learners. In: S. Granger, ed. 1998. *Learner English on computer*. London, New York: Addison Wesley Longman. pp.67–79.

Ehrnrooth, L., 2018. *Tasks used in the examination — FUSE*. [online] Available at: <https://fusecorpus.eu/explore/tasks-used-in-the-examination/> [Accessed 09 August 2019].

Ellis, N.C., 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, [e-journal] 32, pp.17-44. https://doi.org/10.1017/S0267190512000025.

Erman, B. and Warren B., 2000. The idiom principle and the open choice principle. *Text*, [pdf] 20(1), pp.29-62. Available at: <https://lextutor.ca/n_gram/erman_warren_2000.pdf> [Accessed: 09 August 2019].

Hilden, R., Lehti-Eklund, H., Mauranen, A., Vesalainen, M., Sihvonen, P. and Havu, E., 2010. *Hy-Talk (Research Project on Oral Proficiency in languages in compulsory basic education and general upper secondary education) - University of Helsinki Research Portal - University of Helsinki*. [online] Available at: <https://tuhat.helsinki.fi/portal/en/projects/hytalk-research-pro(ea2d9f47-04c0-4b18-9a68-f0d7ca1ed7a9).html> [Accessed 09 August 2019].

Howarth, P., 1998. Phraseology and second language proficiency. *Applied Linguistics*, [e-journal] 19(1), pp.24-44. https://doi-org.libproxy.helsinki.fi/10.1093/applin/19.1.24.

Kecskes, I., 2016. Deliberate creativity and formulaic language use. In: K. Allan, I. Kecskes and A. Capone, eds. 2016. *Pragmemes and Theories of Language Use*. Cham Switzerland: Springer. pp.3-20. Available at: ResearchGate <https://www.researchgate.net/> [Accessed 09 August 2019].

Kecskes, I., 2007. Formulaic language in English lingua franca. In: I. Kecskes and L.R. Horn, eds. 2007. *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects*. Berlin/Boston: De Gruyter, Inc. pp.191–218. [online] Available through: Helsinki University Library website <http://www.helsinki.fi/kirjasto/en/home/> [Accessed 09 August 2019].

Martinez. R. and Schmitt, N., 2012. A Phrasal Expressions List. *Applied Linguistics,* [e-journal] 33(3), pp.299 –320. https://doi-org.libproxy.helsinki.fi/10.1093/applin/ams010.

Moon, R., 1998. Frequencies and forms of phrasal lexemes in English. In: A. Cowie, ed. 2001. *Phraseology: Theory, analysis and applications.* Oxford: Oxford University Press. pp.79–100.

Myles, F., Hooper, J. and Mitchell, R., 1998. Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, [e-journal] 48(3), pp.323–363. https://doi-org.libproxy.helsinki.fi/10.1111/0023-8333.00045.

Nattinger, J.R. and DeCarrico, J.S., 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nikula, T., 1996. *Pragmatic Force Modifiers. A study in Interlanguage Pragmatics*. Jyväskylä: University of Jyväskylä.

Ortaçtepe, D., 2013. Formulaic language and conceptual socialization: The Route to becoming nativelike in L2. *System*, [e-journal] 41(3), pp.852-865. https://doi.org/10.1016/j.system.2013.08.006.

Overstreet, M. and Yule, G., 1997. On being explicit and stuff in contemporary American English. *Journal of English Linguistics,* [pdf] 25(3), pp.250-258. Available at: CiteSeerX <http://citeseerx.ist.psu.edu/index> [Accessed 09 August 2019].

Pawley, A. and Syder, F. H., 2000. The one clause at a time hypothesis. In: H. Riggenbach, ed. 2000. *Perspectives on Fluency*. Ann Arbor: University of Michigan. Ch.10. Available at: ResearchGate <https://www.researchgate.net/> [Accessed 09 August 2019].

Pawley, A. and Syder, F.H., 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: J. C. Richards and R.W. Schmidt, eds. 1984. *Language and Communication*. New York: Longman. pp.191-226. Available at: ResearchGate <https://www.researchgate.net/> [Accessed 09 August 2019].

Schmitt, N., 2010. Formulaic Language. In: N. Schmitt, ed. 2010. *Researching Vocabulary. Research and Practice in Applied Linguistics.* London: Palgrave Macmillan. pp.117-146. [online] Available through: Helsinki University Library website <http://www.helsinki.fi/kirjasto/en/home/> [Accessed 09 August 2019].

Schmitt, N., 2004. Preface. In: N. Schmitt, ed. 2004. *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: John Benjamins Publishing Co (Language Learning and Language Teaching). pp. viii-ix. [online] Available through: Helsinki University Library website <http://www.helsinki.fi/kirjasto/en/home/> [Accessed 09 August 2019].

Sinclair, J., 1991. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Wood, D., 2006. Uses and Functions of Formulaic Sequences in Second Language Speech: An Exploration of the Foundations of Fluency. *The Canadian Modern Language Review*, [e-journal] 63(1), pp-13-33. https://doi.org/10.1353/cml.2006.0051.

Wray, A., 1999. Formulaic language in learners and native speakers. *Language Teaching*, [e-journal] 32(4), pp.213-231. https://doi.org/10.1017/S0261444800014154.

Wray, A. and Perkins, M.R., 2000. The functions of formulaic language: an integrated model. *Language and Communication,* [e-journal] 20(1), pp.1–28. https://doi.org/10.1016/S0271-5309(99)00015-4.

Wray, A., 2000. Formulaic sequences in second language teaching: principles and practice. *Applied Linguistics*, [e-journal] 21(4), pp.463–489. https://doi-org.libproxy.helsinki.fi/10.1093/applin/21.4.463.

Wray, A., 2002. *Formulaic Language and the Lexicon*. [e-book] Cambridge: Cambridge University Press. Available through: Helsinki University Library website <http://www.helsinki.fi/kirjasto/en/home/> [Accessed 09 August 2019].

9 Appendices

Appendix A

Tasks used in the examination (Ehrnrooth, L. 2018)

Below you can see a list of task types and instructions that were used in the exams that have been recorded and added to FUSE (C = conversation).

Please note that this list represents just some of the tasks in **the third part** of the Spoken Examination of English for Finnish upper secondary school students. However, there are only two task types in part 3 of the current exam format: **structured dialogue and mind map supporting the conversation**. Every upper secondary school English teacher who teaches the vocational course 8 in English and arranges the final exam for the course, downloads a set of tasks from the Finnish National Agency for Education exam database and decides which exam package fits his/her students.

Task in C1 – C3 (structured dialogue)

NEWSPAPERS
STUDENT A
You think that newspaper are old-fashioned and that eventually nobody will read them. Instead, people will read their news on the Internet, like you do. You do not read newspapers at all.
STUDENT B
You think that newspapers will survive the age of the Internet. You like to read your daily newspaper and believe in the future of newspapers as well.

COVER AT LEAST THE FOLLOWING TOPICS:
– Starting the dialogue about newspapers
– Student A: expressing your opinion about the future of newspapers
– Student B: expressing your opinion about the future of newspapers
– Continuing the conversation and arguing for your own point of view
– Ending the conversation

Task in C4 (c = conversation) (structured dialogue)

EATING OUT
STUDENT A
You are a customer in a cafe. You have a big appetite. The waiter/waitress comes to serve you. Order your food and drink.

STUDENT B
You are a waiter/waitress in a cafe. You are serving a customer. Take his/her order and answer any questions.

COVER AT LEAST THE FOLLOWING TOPICS:
– The soup of the day and the other types of starter available
– Recommended main courses and desserts
– The types of drinks available
– The loudness of the background music
– Whether the meal was satisfactory
– Payment by credit card

Task in C6, C11 – C15 (c = conversation) (open-ended conversation task)

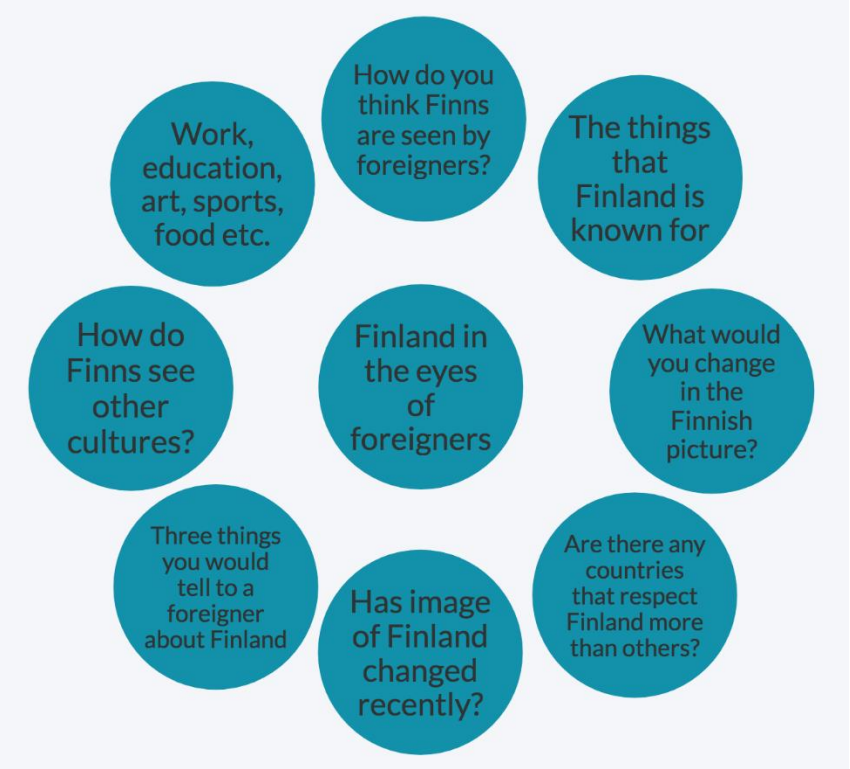
Look at the mind map. Together with your partner, discuss what affects young people and their lives in various parts of the world today. Share your knowledge and opinions on as many aspects as you can in the time available. Make sure that during the discussion you both comment on what the other person says.

Task in C16 (c = conversation) (open-ended conversation task)

Look at the mind map. Together with your partner, discuss the topic. Share your knowledge and opinions on as many aspects as you can in the time available. Make sure that during the discussion you both comment on what the other person says.

Task in C20 (c = conversation) (open-ended conversation task)

Look at the mind map. Together with your partner, discuss the topic. Share your knowledge and opinions on as many aspects as you can in the time available. Make sure that during the discussion you both comment on what the other person says.



Task in C22 – C25, C27 (c = conversation) (structured dialogue)

Read the situation, then act it out together with your partner.

AN UNSUITABLE PRESENT

STUDENT A
You give a friend/relative a birthday present. You have tried hard to choose a suitable present.

STUDENT B
A friend/relative gives you a birthday present. You were hoping for something different than what she/he gives you.

COVER AT LEAST THE FOLLOWING TOPICS:

– Starting the dialogue
– Giving (Student A) and receiving (Student B) the present
– Politely expressing disappointment (Student B)
– Reacting to disappointment, offering alternative (Student A)
– Reacting to alternative idea (Student B)
– Organize alternative
– Ending the dialogue

Appendix B

PUHETEHTÄVÄT [HY-TALK]

Valmistelu 20 min + suoritus 20 min

Suorita seuraavat puhetehtävät parisi kanssa. Tutustukaa niihin ensin 20 minuutin ajan. Älkää käyttäkö apuvälineitä (sanakirjoja tms.). Muistiinpanoja voit tehdä, mutta niitä ei saa lukea suoritustilanteessa. Kukin tehtävä vie korkeintaan viisi minuuttia.

Aluksi käydään lyhyt vapaa keskustelu ”syntyperäisen” puhujan kanssa. (a brief open-ended warm-up interview)

Tehtävä 1. Esittelyvideo (structured monologue task)

Saat kesävieraaksi etäisen sukulaisnuoren (nimeltään Nico tai Anna), jonka perhe on muuttanut kohdekieliseen maahan kauan sitten, eivätkä lapset enää osaa suomea. Toimitat hänelle ensin lyhyen videokatkelman, jossa esittelet perheesi ja itsesi (sen, mitä sanot, ei tarvitse olla totta).

- Tervehdi.
- Esittele perheesi ja itsesi (nimet, iät, mitä kieliä kukin puhuu, mistä pitää tai mitä harrastaa).
- Kysy, mitä kieliä Nico/Anna puhuu ja paria muuta asiaa.
- Kerro, missä asutte ja millainen asunto teillä on.
- Kerro koulustasi.
- Kerro, miten vietät vapaa-aikaasi.
- Mainitse, ketkä ovat parhaat ystäväsi.
- Kerro, mitä teit heidän kanssaan viime kesänä.
- Lupaa jotakin Nicolle/Annalle, kun hän tulee Suomeen.
- Päätä esityksesi kohteliaasti.

Tehtävä 2. Arkitilanteita

Keskustele parisi kanssa mahdollisimman luontevasti. llmaise vuorosanojen asiasisältö kohdekielellä. Älä käännä, vaan yritä saada itsesi ymmärretyksi omin sanoin. Jos et tiedä jotain, älä juutu vaikeaan kohtaan vaan jatka eteenpäin ja puhu mahdollisimman paljon.

Nico/Anna viipyy luonasi kuukauden, jonka aikana käytte seuraavat kaksi keskustelua (numerot 2.1 ja 2.2). Vaihtakaa vuoroja niin, että kumpikin teistä on toisessa tilanteessa oma itsensä **(S=sinä)** ja toisessa vieraan **(Nico/Anna)** roolissa. Sopikaa roolijako ennen kuin alatte puhua.

2.1. Majoittuminen (structured dialogue task)

N/A: Kommentoi kohteliaasti huonetta, jonka olet saanut käyttöösi.

S: Kerro kuka siinä yleensä asuu ja missä tämä henkilö nyt on.

N/A: Kysy, mihin voit laittaa tavarasi.

S: Vastaa, että kaapissa on tilaa vaatteille ja että peseytymisvälineet voi viedä kylpyhuoneeseen.

N/A: Kysy, mihin aikaan perheessä herätään aamulla.

S: Vastaa ja kerro muutenkin päiväohjelmasta kesäaikaan.

N/A: Ojenna ja esittele kaksi tuliaista, jotka olet tuonut perheelle. Kerro myös, miksi valitsit ne.

S: Kiittele tuliaisista ja kerro, mitä aiotte tehdä niillä.

2.2 Keskustelua matkalla elokuvista kotiin (structured dialogue)

S: Kysy, mitä vieraasi piti elokuvasta (mainitse elokuvan nimi).

N/A: Kerro mielipiteesi ja tiedustele toisen mielipidettä elokuvasta.

S: Vastaa kysymykseen ja kuvaile tunnetilaasi elokuvan jälkeen.

N/A: Vertaa elokuvaa johonkin toiseen näkemääsi elokuvaan ja perustele näkemyksesi.

S: Mainitse, mikä muu elokuva on tehnyt sinuun vaikutuksen ja miksi.

N/A: Kerro mielipiteesi elokuvan musiikista ja kysy jotain suomalaisesta musiikista.

S: Vastaa kysymykseen ja suosittele toiselle jotain suomalaista musiikkia.

N/A: Äkkiä huomaat jotakin, joka yllättää sinut (mainitse mitä) ja kehotat toista kiirehtimään.

S: Reagoi tilanteeseen rauhoittavasti.

Tehtävä 3. Retkipäivästä sopiminen (structured dialogue task)

Suunnittelette yhdessä retkeä johonkin suosittuun paikkaan kotiseudullasi.

Sopikaa yhdessä seuraavista asioista:

- mihin retki tehdään, mihin aikaan ja mistä lähdetään
- keitä lähtee mukaan
- miten pitkä matka on ja miten se tehdään (kävellen/bussilla/pyörillä)
- mitä kumpikin haluaa tehdä ja nähdä
- missä syödään ja mitä
- miten paljon rahaa otetaan mukaan ja mihin sitä kuluu
- milloin palataan takaisin
- mitä pitää muistaa / mitä ei saa unohtaa

(extra open-ended dialogue task)

Read this and then go on to discuss reality TV with your pair.

Globetrotting controversy - BB scandals from around the world!

Sex on screen may have been okay in Holland in the first ever show, but it caused an outrage in Portugal, and the offending couple were thrown out of the house before an official investigation began (they subsequently got married).

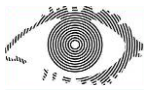
Other countries have even weaker stomachs when it comes to on-screen behaviour. In Thailand the government considered shutting down the first series when two contestants became friendly; holding hands, hugging and cuddling on the sofa.

Viewers in the Netherlands accepted sex, but even they have limits. Authorities were not prepared to allow a pregnant housemate to give birth on screen last year.

Meanwhile, it was in Denmark that the first contestant actually became pregnant while on the show, not something that would have been allowed in the Middle East version of BB - which despite separate living quarters for men and women was abandoned due to mass protests in Bahrain at 'offensive' content.

Britain has pushed a few boundaries of its own – they've had the first gay winner (Brian Dowling) and the first transsexual winner (Nadia Almada). Although nudity and racism-related drama of the Celebrity BB earlier this year have caused some controversy in the press, they've yet to erupt in public protest, as seen in Germany, France and Greece.

TOP 5 BB-scandals



1. **The first pregnancy that began in the BB-house** (Denmark)
   Revealed only after the mother was evicted from the BB-house.

2. **The birth of a BB-baby live on TV** (Holland)
   Only the mother's face was shown on camera

Adapted from:

http://www.tiscali.co.uk/events/2006/big-brother-7/features/bb-scandal.htm

ARVO 10/07 (A youth publication of Central Organisation of Finnish Trade Unions)