Pattern, form and function of code-switching in a Dutch-English online community

BA Thesis English Language and Culture, Utrecht University

Tessa van der Heide

5894220

Supervisor: dr. Nynke de Haas

Second reader: dr. Koen Sebregts

November, 2019

10154 words

Abstract

Much is still unknown about code-switching, particularly about written code-switching online. This thesis investigates the English-Dutch bilingual online community at r/theNetherlands, a subforum of the online discussion forum Reddit, which had not yet been investigated. The focus is on the pattern, form and functions of code-switching. English-Dutch bilingual code-switching patterns had yet to be investigated; this thesis found that Dutch-English is the most dominant pattern of code-switching for this community, though motivations for this choice are still to be studied. There had been very few studies comparing intersentential and intrasentential code-switching, though it has been found that intersentential code-switching appears to be the dominant form of code-switching of written discourse online. This study found no conclusive evidence for preference of either form of code-switching, possibly indicating that online written code-switching is a combination of written and oral discourse. Code-switching can be used to fulfil several functions within the discourse. This thesis found that code-switching was most frequently used to quote, to fulfil a lexical need and to add emphasis. A logistic regression analysis was applied to the data in order to investigate interactions between pattern, form and function. The results showed that there was a significant association between pattern and form, and pattern and function, though there was no significant association between form and function. This indicates that there are more factors associated with code-switching pattern, for which further investigation is necessary.

Keywords: code-switching, intersentential, intrasentential, Internet, code-switching pattern, code-switching function, code-switching form, Reddit, English-Dutch, Dutch-English

Table of contents

Introduction

Code-switching, a fluid alternation between two languages (Rabinovich, Sultani & Stevenson, 2019), is a language contact phenomenon that has been used by individuals speaking multiple languages. Fluid is defined by Myers-Scotton (1993) as "[...] in the same conversation, within the same conversational turn, or even within the same sentence of this turn" (p. vii). The usage of two or more languages, bilingualism, has been studied extensively over the past decades. Nevertheless, the field cannot agree on the definition of bilingualism. Bullock and Toribio (2009) discuss multiple definitions: those who have been exposed to two languages from birth or early childhood are bilingual, or those who acquire their first language and then a second language, or those who are equally fluent in both languages, or those who are able to communicate in any way in multiple languages. This thesis will follow Grosjean's (2010) definition of a bilingual: "Bilinguals are those who use two or more languages (or dialects) in their everyday lives" (p. 4). By following this definition, emphasis is placed on the importance of language use and code-switching is part of a bilingual's language use. This thesis intends to provide insights into how code-switching is influenced by its pattern, its form and the functions it can fulfil in online discourse by studying code-switching in a bilingual Dutch-English online community, because much is still unknown about code-switching, as well as the fact that English-Dutch language mixing and this online community has yet to be studied.

1. Theoretical background

1.1 Language contact phenomena

Bilingualism, the usage of two or more languages (Grosjean, 2010), results from language contact, the instances in which multiple languages meet and can mix with each other. Code-switching is a language contact phenomenon where a speaker speaks one language and then

switches to another one for a word or phrase (intrasentential code-switching), or sentence (intersentential code-switching), or when a speaker switches languages for an extended period (Grosjean, 2010; Winford, 2002). In order to define code-switching, it must be distinguished from other language contact phenomena such as lexical borrowing, usage of unassimilated loan words, loan translations (calques), diglossia and translanguaging.

Lexical borrowing, or loan words, usually involves a word from one language being adopted and accepted into the other language. Unassimilated loan words only occur in bilingual speech; the bilingual transfers a lexical item from one language to another language of which the item is not an established part (Bullock & Toribio, 2009). Code-switching is occasionally difficult to distinguish from borrowing, as a lexical item that is in the process of being adopted by the other language could be present in both monolingual and bilingual speech, whereas code-switching is only present in bilingual speech.

Calques are defined by Backus and Dorleijn (2009) as "words or phrases that are reproduced as literal translations from one language into another" (p. 75) and are similar to lexical borrowing in the sense that they transfer a lexical item from one language to another. However, calques adapt the item to the new language while retaining a structure similar to the original language. This is different from code-switching, where two languages co-exist, without any lexical items being established as part of another language.

Diglossia occurs in a community where a specific language is associated with a particular social function. The selected language is socially imposed; this stands in contrast with code-switching, where the speaker is freely able to switch between languages (Bullock & Toribio, 2009). Blom and Gumperz (1972, as cited in Nilep, 2006) named code-switching as a representation of a change in social setting, as is the case in diglossic communities, situational switching; code-switching within a single social setting was named metaphorical or conversational setting. Situational switching relates to the concept of translanguaging,

which “refers [...] to the complex language practices of plurilingual individuals and communities [...]” (García & Wei, 2014, p. 20), emphasising the usage of two or more languages in one context and employing languages for different purposes. Translanguaging, however, does not mandate one language over the other as diglossia does, and is mostly used in pedagogical settings; code-switching is one of those complex language practices that can be part of translanguaging. Gumperz (1982, as cited in Nilep, 2006) acknowledged the difficulties in establishing whether language choices are situational or metaphorical, as there are too many interactions between linguistic form, setting, activities, topics and participants.

1.2 Motivations for code-switching

Code-switching is done by bilinguals at different levels of proficiency and with different levels of usage in various language contact situations, which means that their code-switching patterns can differ greatly. As stated before, code-switching can happen in the same sentence (intrasentential) or between sentences (intersentential). An example of intrasentential code-switching in English and Spanish (Poplack, 1980, p. 594) is given in (1) and an example of intersentential code-switching in Swahili-English is given in (2) (Myers-Scotton, 1993, as cited in Bullock & Toribio, 2009, p. 3):

(1) *Spanish-English*

Sometimes I’ll start a sentence in Spanish ***y termino en español***.

‘Sometimes I’ll start a sentence in Spanish and I finish in Spanish’

(2) *Swahili-English*

That’s too much. ***Sina pesa***.

‘That’s too much. I don’t have much money’

Both methods of code-switching require advanced knowledge of grammar for both languages, since code-switching requires the production of a fully grammatical clause in both languages, though it could be argued that intrasentential code-switching is grammatically more complex, since two grammars have to be acceptably mixed into one utterance (Bullock & Toribio, 2009). Bilingual speakers can differ in their level of fluency per language; Poplack (1980) found that non-fluent bilinguals favoured intersentential code-switching, whereas fluent bilinguals favoured intrasentential code-switching, though Berk-Seligson (1986) found evidence against the assumption that the form of code-switching is correlated with the degree of bilingualism. There has been little research comparing the frequency of intersentential code-switching with the frequency of intrasentential code-switching. Koban (2012) interviewed twenty first- and second- generation Turkish-English bilinguals from New York City and Li, Yu and Fung (2012) created a Mandarin-English code-switching corpus from transcriptions of conversational meeting speech data, project meeting speech data and student interviews, as well as online news data. Only non-written discourse was used for analysis, as Li, Yu and Fung note that code-switching in online news data is not representative of written code-switching, since it is too different in style from spoken and written Mandarin. Both studies found a preference for intrasentential code-switching.

The question of why and where bilinguals code-switch has been a topic of research for the past decades. Initial research primarily focused on linguistic constraints rather than extra-linguistic factors, such as Poplack (1980), who argued for the equivalence constraint on code-switching, a constraint by which code-switching may only occur at points in the discourse where it does not violate any of the syntactic rules for the two used languages. Berk-Seligson (1986) discussed the three general linguistic constraints that emerged from previous research: the equivalence of structure constraint, the size-of-constituent constraint, and the free morpheme constraint. The equivalence of structure constraint was proposed by

Poplack (1980): “According to this simple constraint, a switch is inhibited from occurring within a constituent generated by a rule from one language which is not shared by the other” (p. 586). Poplack studied Spanish-English bilinguals and suggested that the size of the code-switched constituent correlates with the bilingual ability of the speaker and the frequency of code-switching, meaning that fluent bilinguals tended to code-switch more frequently, preferred to use intrasentential code-switching and code-switched smaller constituents, whereas non-fluent bilinguals code-switched less frequently, preferred to use intersentential code-switching and code-switched larger constituents (Poplack, 1980). The free morpheme constraint states that that code-switching is not acceptable between a free morpheme (a morpheme that can stand alone as its own word) and bound morpheme (a morpheme such as a suffix or prefix, such as *de-*, *in-*, *-able* and *-ing*). The debate on the universality of these constraints is still ongoing (Berk-Seligson, 1986; Bhat, Choudhury & Bali, 2016; Redouane, 2005). Since then, research has started to focus on social factors, such as context, identity and interaction, and their influence on code-switching, starting with the study of code-switching though the model of situational switching and conversational switching by Blom and Gumperz (1972, as cited in Nilep, 2006).

Most researchers base their studies on the motivations behind code-switching on the premise “that [code-switching] is a conscious choice on the part of the speaker, used to mark quotations, emphasis, realignment of speech roles, reiteration, and elaboration” (Bullock & Toribio, 2009, p. 10). Grosjean (1982) suggested multiple reasons as to why a speaker would code-switch, such as emphasising an individuals’ group identity, specifically directing their conversation to a specific participant, adding emphasis to what has been said, or talking about past events. Auer (1984, as cited in Wei, 1998) proposed a theory which viewed a social situation as an interactively achieved phenomenon. Auer used the terminology and analytic framework of ethnomethodology and conversation analysis to analyse code-switching; he

focused on how bilinguals come to their choice of language on the assumption that every choice made by a participant influences the subsequent choices of other participants. Stroud (1998) argued against Auer's (1984, as cited in Wei, 1998) approach of only using conversation analysis to analyse code-switching, suggesting that "(...) conversational code-switching is so heavily implicated in social life that it cannot really be understood apart from an understanding of social phenomena" (p. 322), meaning that both language use and social action should be analysed, rather than just language use as conversation analysis does. Wei (1998), on the other hand, argued in favour of the conversation analysis approach, as it examines the types of interactions involving code-switching, rather than relying on "intuitive categories as a basis for the description of code-switching […] [w]hile such models may be convenient for those working on community where a rigid diglossia obtains, their methodological validity and applicability are questionable" (p. 309).

Another theory that attempts to explain the social motivations behind code-switching is the markedness theory of code-switching by Myers-Scotton (1993, as cited in Wei, 1998), which suggests that every interaction is based on a marked and unmarked rights-and-obligations set and that speakers know which linguistic realisation is unmarked and which one is marked; this leads to the assumption that speakers who choose to use a marked linguistic realisation, such as code-switching, do so deliberately. However, Bullock and Toribio (2009) pointed out the following:

> [N]ot all language alternations in bilingual speech do signal a particular communicative intent or purpose; for many bilinguals, [code-switching] merely represents another way of speaking; that is, some bilinguals code-switch simply because they can and oftentimes may not be aware that they have done so. (p. 11)

In conclusion, though the question of the motivations, social and/or linguistic, for bilinguals to code-switch is certainly very interesting, no consensus has been reached yet;. the relationship between the function and form of code-switching remains unclear.

1.3 Oral and written discourse

A brief discussion of the difference between oral and written discourse is necessary to be able to make a distinction between oral code-switching and written code-switching. Olson (2006) described oral discourse as a complex set of procedures that are at play in order to produce a common understanding. Examples of these procedures are building linguistic structures through grammar and form, as well as stress and intonation, which are prosodic features exclusive to oral discourse, and the interpretation of body language. Social properties, such as the physical space the conversation is in, shared background knowledge, and information about the identity of the participants also influence the conversation. Written discourse only uses part of the linguistic procedures used in oral discourse; it especially lacks context, as writing can be open to interpretation as time passes, which is less of an issue with oral discourse. Because of this, written discourse requires more effort on the writer's part, in order to maintain the intended interpretation:

> Simply put, writing, like quoted speech, invites the distinction between the speaker's intended meaning and the sentence's meaning. [...] Writers must invest considerable effort in making the linguistic properties of the written form capture or sustain, so far as possible, the meaning [they] intended. (Olson, 2006, p. 138)

This assumption applies mostly to written monologues, such as a novel or article, and to a lesser extent to written conversations, especially online. Online forums, where these written conversations take place, often have the ability to edit or add to posts, meaning that writers can sustain or clarify their intended meaning after posting, which is not possible in

monologues or offline written discourse. The theory that writing costs more effort was also put forth by Cayer and Sacks (1979), who described writing as "a more formal, less natural endeavour, one which involves the development of a sensitivity to requirements unique to the written mode" (p. 121). They argued that, as the writer lacks a known and specific audience, something that is usually present with oral language, written language imposes more linguistic demands, causing the semantic and syntactic complexity to increase. Cayer and Sacks also mentioned surface structures: words such as *right*, *well*, *yes* and phrases such as *I guess*, *I feel*, *I think*, which occur frequently in oral language, but are usually not present in written language. This suggests that oral and written discourse are different. Redeker (1984) elaborated on these differences between spoken and written language, focusing on to what extent the scales (a range between high and low) of involvement, detachment, fragmentation and integration differ when comparing oral and written discourse. Though involvement (such as self-references, mentions of mental processes and use of colloquial expressions) and detachment (such as the usage of the passive voice, past perfect and literary expressions) are both regarding "the communicator's perceptions of the situation and [their] attitude towards the message and the recipient(s)" (p. 44), they are measured by the presence and absence of different features as they reflect different cognitive factors. Text fragmentation (such as one-verb clauses without coordinating or subordinating conjunction) and integration (such as adjectives, nominalisations, and complement clauses) are the other two scales. All four scales are displayed in both speech and written text, though Redeker found that involvement features occur less frequently, and detachment features occur more frequently in written discourse; text fragmentation features occur more frequently, and integration features occur less frequently in spoken discourse. She concluded that there were extensive differences in the level of integration and fragmentation, and that these could reliably be used to discriminate between spoken and written discourse. Levels of involvement proved to be a less

reliable discriminator and the level of detachment proved too unreliable. In conclusion, it can be said that oral discourse and written discourse both have their own characteristics and, more specifically, that written discourse requires more planning and effort than oral discourse, and that this is visible through the presence and absence of these features.

1.4. Written code-switching

Most research on code-switching discussed up to this point focused on code-switching in oral discourse. However, oral discourse and written discourse are different. This leads to the question whether code-switching would be governed by different principles and motivations in written discourse. Written discourse appears to be more deliberate, which could lead to the conclusion that code-switching in written discourse is also more deliberate. So far, these has been no consensus whether code-switching differs depending on genre and discourse style, either oral or written. McClure (2001) concluded that the form and function of code-switching are different across genres and across different modes of communication, such as text and speech. Sebba, Mahootian and Jonsson (2012) noted that "online chat and text-messaging share many of their features with spoken conversation" (p. 7). Gardner-Chloros and Weston (2015) found that the functions of code-switching in written discourse at least partly overlap with the functions of code-switching in oral discourse: "The conventions and constraints of speaking and writing may be different, but the broad semiotic consequences of setting up contrasts by alternating languages are common to both" (p. 189). Code-switching might be used in a similar manner to spoken discourse on informal written platforms, such as computer-mediated communication. Sebba, Mahootian and Jonsson (2012) emphasised the importance of the study of written code-switching as a its own phenomenon, separate from oral code-switching. They emphasise the lack of both a coherent framework to contextualise

code-switching, and the lack of an "independent theoretically informed field of 'written multilingual discourse studies'" (p. 2).

1.5 Methodologies

Various methodologies have been traditionally used to study code-switching, such as introspective observation, observation of bilingual children, group testing, and guided conversations (as cited in Weinreich, 1968). Sebba (2012) argued in favour of a new research approach, situating the study of written multilingual discourse in a field that encompasses the semiotics of all mixed-language texts, analysing written multilingual discourse within a literacy framework, and taking contextual elements, both visual and spatial, into account as contextualisation cues. More recent studies also used informal interviews (Ortega, 2008) and text corpus analysis, such as a study done by Pahta and Nurmi (2011) on historical code-switching.

Additionally, researchers have started to recognise the value of the Internet as a source of linguistic data from various genres and contexts. Crystal (2006) argues that the Internet has led to language change, as online communities nurture innovation and creativity, which support the development of internet varieties, such as Instant Messaging (IM) language, which typically use abbreviations and contain typing errors, as well as lack capitalisation and full stops. The usage of computer science in the humanities led to the development of the field of Digital Humanities, defined as "the applications of computing to research and teaching within subjects that are loosely defined as 'the humanities'" (Hockey, 2004, Introduction, para. 2), making the tools of computer science, such as computational statistical analysis and online corpora, available and accessible for researchers in the Humanities; this enables researchers to study language use on the Internet. A study done by Piperski, Belikov, Kopylov, Selegey, and Sharoff (2013) investigated linguistic variation in

present-day Russian that is available on the Web. They illustrate that one of the benefits of an online corpus is the large amount of data available on present-day language. This sentiment is echoed by Minocha, Reddy and Kilgarriff (2013), who put forward a method to continually update a corpus on the English language by crawling social media feeds to ensure that current language use is always represented.

1.6 The Internet and code-switching

The Internet has also been used to study code-switching. Dorleijn and Nortier (2008) discussed code-switching on the Internet, noting that written discourse in computer-mediated communication, such as written dialogues on the Internet, has a tendency be much more informal and less reflected upon than written discourse in general. Online written discourse also frequently contains features that are usually associated with oral discourse. They emphasised that the same level of consciousness required for written discourse is still present with code-switching on the Internet, as it is still part of written discourse. A study done by Androutsopoulos (2006, as cited in Dorleijn & Nortier, 2008) on a Persian-German internet forum found that "[i]ntersentential [code-switching] [...] is the main form of language contact in internet forums; intra-sentential [code-switching] [...] occurs only sporadically and is caused by local changes in the discussions" (p. 134). The study of code-switching on the Internet might help shed light upon the stylistic uses of code-switching (Hinrichs, 2006; Jaworska, 2014) as well as explain the role of code-switching in constructing online identities (Leppänen, 2012; Themistocleous, 2015). Dorleijn and Nortier (2008) noted that it remains to be investigated to which degree there is an overlap in usage of code-switching in oral discourse and informal written discourse online. They also suggested studying language pairs in different social contexts in order to investigate social perceptions of code-switching and whether these perceptions change depending on context.

Social media is one of those contexts. Caparas and Gustilo (2017) investigated code-switching on Facebook by analysing Facebook posts by Filipino English-Tagalog bilinguals, who used Tagalog and English, and also used the regional languages Chavacano, Cebuano or Tausug. They focused on the form and functions of code-switching, and compared the standards for oral code-switching to written code-switching. They categorised code-switching on the basis of an analytical framework (Saville-Troike, 1986; Hoffman, 1991, as cited in Caparas & Gustilo, 2017) by identifying code-switching functioning as content-specific discourse (a preference to use a specific language depending on the topic), emphatic, quoting (famous expressions, proverbs or sayings), interjection (inserting sentence fillers or connectors), repetition (repeating the same content in a different language in order to amplify the message), clarification (translation), an expression of group identity (using community-specific terms), limiting the audience, a stand in for a word that lacks a lexical equivalent (such as jargon) or strengthening or softening a command (the speaker establishes a position of power by showing off the ability to switch languages, or softens the command by switching to a shared (non-dominant) language). They found that the primary functions of code-switching in English-Tagalog were that the speaker felt their lexicon in their language was lacking and that another language would be better suited to convey their message. They felt that their original list of functions was insufficient and added other functions: spontaneously expressing ideas (using common expressions and formulaic language), retaining native terminology (maintaining native concepts, for example *Misa de Gallo* 'Night Mass'), expressing disappointment (indirectly, by using a different language) and promoting relationships (using terms of endearment and greetings). They also found that there was a preference for Tagalog-English ('Taglish') code-switching over all other language combinations; this is due fact that Taglish can act as a bridge, to connect Filipino speakers with multiple languages, acting as a unifying language available to all participants, including

those not from a Tagalog-speaking region; they suggest that the role of the English language on the Internet cannot be understated. They also found that code-switching was more frequently intrasentential than intersentential. They suggest that "online communication is a quintessential place for code-switching" (p. 357), as the Internet encourages language contact, language alternation and code-switching due to the dominant position of the English language online, which is used in combination with the speaker's native language and possible regional languages. This shows that the extensive data available on social media can have various linguistic applications, such as investigating linguistic variation and analysing the use of code-switching.

1.7 Reddit and CS

Reddit is another social media platform that can offer valuable data for linguistic research. In 2015, a data set on the entirety of the online forum Reddit was published as a project called Pushshift.io (Baumgartner, 2014). Reddit is a social news platform where individuals interact with each other in interest-based communities called subreddits by creating posts (submissions) and responding with comments. The corpus, made available to the public to share with the academic community, included over 1.7 billion submissions and comments.

This data set has been used in various research papers (Massanari, 2015; Guestrin, 2016), such as a study done by Saleem, Dillon, Benesch and Ruths (2016), who used Reddit for the development of a tool to detect hateful speech without the need for manual annotation. Rabinovich, Sultani, and Stevenson (2019) were the first to use this Reddit data set to investigate online written code-switching, as this remains understudied, in order to compare it to oral code-switching. They analysed code-switching and its relation to linguistic proficiency, linguistic style and content in English-Tagalog, English-Greek, English-Romanian, English-Indonesian and English-Russian communities. General language

proficiency, across all languages, was estimated by calculating several lexical and grammatical measurements that are usually used for language proficiency assessment: type-token ratio (the number of unique words divided by the total number of words), lexical density (the number of content words, so excluding function words, divided by the total number of words), average age of acquisition of tokens, average word concreteness (concreteness was based on a list by Brysbaert et al., 2014, as cited in Rabinovich, Sultani, & Stevenson, 2019) and mean word length. They found that highly proficient code-switchers had a lower score for lexical density but scored higher in grammatical complexity and support Bullock and Toribio's hypothesis: "while bridging lexical deficiencies, it may require advanced grammatical capabilities in order to construct mixed sentences without distorting the 'grammaticality' of the target utterance" (p. 8). They also found that Reddit posts in general tended to be very informal (as measured by the presence of informality markers, such as exclamation marks, contractions and colloquialisms, such as *lol*, *dude*, *like*), and that code-switched Reddit posts were more informal than monolingual posts; this is in line with what is expected for oral discourse. Additionally, code-switched posts often discuss relationships and family. They concluded that there are topical and stylistic distinctions between code-switched and monolingual communication in Tagalog, Greek, Romanian, Indonesian and Russian on Reddit. The English-Dutch bilingual community on Reddit, however, has yet to be explored.

The aforementioned studies show that Pushshift can be a useful source of linguistic data, but it comes with its own flaws. Gaffney and Matias (2018) discuss the gaps in Pushshift's data set, pointing out the fact that data is missing from the data set because of incomplete data gathering, which is due to Reddit's infrastructure, and the ability of users to delete their submissions.

1.8 Identification of research niche

Although much research has been carried out on spoken code-switching and code-switching in general, much is still unclear about the functions and forms of written code-switching, especially online. Additionally, the Dutch-English community remains to be investigated. This thesis will attempt to bring its contribution to the growing body of research on code-switching. Following the definition that code-switching is a fluid alternation between two languages, this thesis intends to investigate the patterns, types (intrasentential and intersentential) and functions of code-switching in a Dutch-English online community on Reddit, as the patterns of code-switching remains to be investigated for Dutch-English, as well as little research into comparison of code-switching forms. This study will follow Caparas and Gustilo's (2017) approach of categorisation and frequency analysis, in order to determine the functions, patterning of code-switching, and to determine what the relative frequency distribution of code-switching form is. This thesis will present an exploration of the Dutch-English bilingual community, as present on r/theNetherlands on Reddit.

2. Research aims

2.1 Research questions

How is written code-switching, defined as a fluid alternation between two languages in written discourse, used in the Dutch-English online community at /r/theNetherlands, and how does this compare to what is known about oral code-switching?

I. What is the distribution for code-switching forms (intrasentential and intersentential)?

II. What is the distribution of code-switching pattern? Is English-Dutch or Dutch-English switching more frequent? Does this differ depending on form (intrasentential or intersentential)?

III. What are the functions, as defined by Caparas and Gustilo (2017), of code-switching in this community?

IV. Are there any interactions between the form, pattern, and function of code-switching, and if so, what are they? Are the functions different for intrasentential code-switching than for intersentential code-switching? Are the functions different for Dutch-English or English-Dutch? Is there a difference in functions per form for Dutch-English and English-Dutch?

2.2 Hypotheses

The issue of code-switching in online discourse has not been adequately addressed. For r/theNetherlands, there will most likely be a preference for intersentential code-switching; this is because most members of this community will have acquired English as a second language, and therefore might not be as proficient in English as in Dutch, and intersentential code-switching is less grammatically complex than intrasentential code-switching. In addition to that, it has been found that intersentential code-switching tends to be the main form of code-switching in Internet forums (Androutsopoulos, 2006, as cited in Dorleijn & Nortier, 2008). Though a tendency towards intrasentential code-switching has also been noted, this has only been studied for oral code-switching (Koban, 2012; Li, Yu & Fung, 2012).

There has been no research yet into the patterning of code-switching for English and Dutch, or whether the patterning differs depending on code-switching form. The dominant pattern could be from English into Dutch, as Reddit predominantly uses the English language, and English has been noted to be the dominant language on the Internet (Caparas & Gustilo, 2017). The pattern could also be from Dutch into English, as the dominant language on r/theNetherlands in particular is Dutch, and most members have Dutch as their first

language and English as their second language, and therefore might not be as proficient in English as in Dutch.

Caparas and Gustilo (2017) found that speakers most frequently code-switched because they felt there was a lexical deficiency in the vocabulary of one language, but code-switching was also used because the speaker felt that the other language was able to clarify the content of the message better and to add emphasis to their message. There has been no evidence that this would not be the case for English and Dutch, so code-switching is hypothesised to function similarly to the functions for English-Tagalog found by Caparas and Gustilo. As intrasentential code-switching and intersentential code-switching differ slightly in grammatical complexity, it could be that these two types are used for different purposes; it is also possible that the code-switching pattern is influenced by the functions of code-switching, though Caparas and Gustilo (2017) did not look into either of these possibilities.

As code-switching is a complex language contact phenomenon, this thesis will investigate if there are any interactions between the form, intersentential or intrasentential code-switching, pattern, English-Dutch or Dutch-English, and the functions, as established by Caparas and Gustilo. This has not been investigated before in the context of Dutch-English code-switching.

3. Methodology

3.1 Materials

A data set on r/theNetherlands that was collected by Rabinovich, Sultani and Stevenson (2019) will be used. They collected every post available at the time of collection from all country-specific communities with Pushshift, a project by Baumgartner (2014) designed to search Reddit content. They stored the username of the author, the subreddit name, the date of posting, the textual content of the post, and metadata regarding context (Reddit's

identifiers for the conversational chain in which the post is visible). The raw data set had 6.88 million posts from 71 different country-specific subreddits. The data set on r/theNetherlands contained 3112 data points before cleaning (see Procedure); for the Dutch language, only r/theNetherlands was used, as it was the only country-specific subreddit for the Netherlands. There are a few city-specific subreddits, such as r/Amsterdam (40.000 subscribers), r/Rotterdam (4000 subscribers), and r/theHague (1700 subscribers), but they are fairly small and inactive when compared to r/theNetherlands. There used to be another country-specific subreddit for the Netherlands: r/netherlands. However, this was deemed an unreliable source, as it was taken over by American moderators (community leaders) and closed in 2014, which led to all active members of r/netherlands to migrate to r/theNetherlands, which currently has 268.000 subscribers.

Polyglot (Chen & Skiena, 2014), a probabilistic tool to identify the languages present in a text, was used to identify only those posts which contained English and another language. Posts shorter than five tokens or containing weblinks were removed. Rabinovich, Sultani and Stevenson (2019) defined code-switching as "a fluid alternation between two languages in an author's own words" (p. 3) and filtered their data set to enforce this definition. The reply-to sections (parts of another Reddit users' post that were included in the post), and quotes were removed in order to prevent them from being counted as code-switching. Named entities were also filtered out, based on the Named-Entity Recogniser (NER) in spaCy (Honnibal & Montani, 2015). Posts that were most likely to be translations, based on several features typical to r/translator, which is a community for translation requests, were also removed. Then, the Polyglot tool (Chen & Skiena, 2014) was applied again and posts that were still identified as containing two languages were used to create the data set. In order to preserve contextual information, named entities and quotes were re-inserted into the final data set.

The precision of the identification of code-switching was tested by manual annotation of a random sample of 500 posts from the compiled data set for English with Tagalog, Greek, Romanian, Indonesian, Russian, Spanish and French, which was taken to be representative of the data set as a whole. The annotation had an agreement rate for manual annotation between annotators of 83.4% across all posts correctly identifying code-switching, which Rabinovich, Sultani and Stevenson deemed sufficient. The automatic detection of English-Tagalog and English-Indonesian correctly identified an actual code-switched post (true positive) 99% of the time; this was 87% for English-Romanian and English-Greek, and 85% for English-Russian. English-Spanish detection had a precision of 70% and the detection of English-French code-switching was extremely low due to a high extend of shared lexical items.

The data set on r/theNetherlands contains comments and submissions collected from r/theNetherlands were posted on Reddit between March 2014 and September 2019 and had been selected with the filters above, but had not yet been manually annotated in order to evaluate the precision of the detection of code-switching.

3.2 Participants

The Dutch online community on Reddit, known as /r/TheNetherlands, is part of the English-language online community at Reddit.com, therefore creating an Dutch-English online community. In this community, members can create submissions and comments in English, Dutch, or any other language they desire.

Demographic information on r/theNetherlands is available via the community's census survey, which is a voluntary anonymous survey repeated every year since 2015. The survey results from 2019 were used in this study; 2719 responses have been recorded, and the results are available at http://bit.do/censussurvey. This includes the distribution of age groups, gender, sexual orientation, marital status, nationality, province, language proficiency,

connection to the Netherlands, political and religious affiliation, work and education, and subreddit preferences. This census survey from 2019 is taken to be more representative of the community than a new survey specifically for users of code-switching on a smaller scale, due to the number of responses (2719). It is also not considered feasible to take surveys from 2015 to 2019 into account due to the lack of access to the original data, or approach users who posted code-switched posts directly due to time constraints.

The age distribution of r/theNetherlands, according to the census survey, is as follows: 1.3% (34) of the participants are between 13 and 15 years old, 10.2% (276) between 16 and 18 years old, 54,8% of the participants (1495) are between 19 and 25 years old, 17.3% (470) between 26 and 29 years old, 13.7% (372) between 30 and 41 years old, 1.9% (52) between 42 and 50 years old, and 0.8% (21) are over 51 years old. The overwhelming majority (82.9%, 2255) is male, 15.4% is female, and 1.7% is other. Almost all respondents are of Dutch nationality (94.8%, 2577), with a slightly smaller portion also being native speakers of Dutch (92.8%, 2517). The distribution of the highest level of education completed is 8.8% vocational (240), 33.4% secondary (908), 35.2% tertiary (958), 18.3% master's or doctorate (497), and 4.3% other (116). It is important to note that this is about the community as a whole, rather than specific to participants.

3.3 Procedure

The data set on r/theNetherlands had not been properly prepared yet for processing. The raw data set contained 3112 data points, with every data point containing English and another language. There were 3042 points with a combination of English and Dutch. The remaining 70 data points were a combination of English with German, Japanese, Greek, Afrikaans, Hungarian, Swedish, Danish, Latin, Spanish, Chinese, Hebrew, Russian, French, Indonesian, and Thai; these data points have been removed, as they were deemed irrelevant for the

present study. Then, a random sample of 500 data points was taken and manually checked for the presence of code-switching. This sample was used for further analysis.

The data set was manually annotated on whether the data point contained intrasentential or intersentential code-switching, or both, and whether the code-switch was from English to Dutch, or Dutch to English; the numbers of occurrences of both types of code-switching were then calculated, with data points that contained both adding to both counts, as well as calculating the number of the occurrences of English-Dutch and Dutch-English code-switching. If a post contained multiple code-switches, as in Dutch to English and back to Dutch, the first code-switching pattern (Dutch-English in this example) was used to code the post, as any possible following code-switches can only occur in the context of that pattern. These counts were tested on goodness of fit with a chi-squared test. The corpus was split into two subcorpora, Dutch-English and English-Dutch, and analysed on the number of occurrences of intrasentential and intersentential code-switching; the distribution of forms per pattern was tested for significance.

The list of possible functions of code-switching used by Caparas and Gustilo (2017) was then used to attribute one or more functions to each data point in the corpus, via manual annotation by one annotator. This list includes code-switching for content-specific discourse, being emphatic, quoting, repetition, clarification, expressing group identity, limiting the audience, filling in to fulfil a lexical need, strengthening and softening a command, as well as spontaneously expressing ideas, retaining native terminology, expressing disappointment, and promoting relationships. The function of spontaneously expressing ideas was renamed to formulaic language, as it is more representative of the actual use of code-switching, which, in this case, was to use common expressions and formulaic language. The number of occurrences of the functions were then calculated; these counts were used to create a list of the most frequent functions of code-switching. The data set was split into intersentential

code-switching and intrasentential code-switching, which were then analysed on function frequencies, in order to create a sorted list of the functions of code-switching per code-switching type; a chi-squared test was then applied in order to test whether the distribution of functions differed significantly between intersentential and intrasentential code-switching. The subcorpora of English-Dutch and Dutch-English code-switching were also analysed on the number of occurrences of functions, as well as tested for significant differences in distribution with a chi-squared test. Lastly, the English-Dutch and Dutch-English subcorpora were split further into intrasentential and intersentential code-switching and then analysed on the number of occurrences of functions, and also tested for significant differences in distribution with chi-squared tests.

In order to investigate interactions between pattern, form and function, a binary logistic regression analysis was applied to the data. Pattern, form and function are all categorical variables and had to be dummy-coded in order to conduct analysis. Pattern had two levels (levels: Dutch-English and English-Dutch), as did form (levels: Intersentential and Intrasentential). Function initially had 13 levels (levels: Audience, Quote, Lexical Need, Emphasis, Clarification, Interjection, Identity, Disappointment, Native terms, Repetition, Formulaic, Topic, Soften/Strengthen), but only the levels with 30 or more occurrences (levels: Clarification, Emphasis, Lexical Need, Quote, Formulaic, Topic) were taken into account for the logistic regression analysis in order to prevent skewing of the results. The logistic regression analysis was binary, with pattern as the dependent variable and form and function as predictors. The predictors were added at the same time. Form and function were tested as possible factors influencing pattern, as well as testing for interaction between form and function. An alpha level of 0.05 was used to determine significance.

4. Results

4.1 Form preference

Figure 1 shows the distribution of the relative frequencies of the forms of code-switching, either intersentential or intrasentential, with the absolute frequencies added. The distribution of the code-switching forms of intrasentential (51%) and intersentential (49%) code-switching appears fairly equal. This was tested for significance, with a p-value of .05. According to a χ^2 goodness of fit test, there is no significant difference in the distribution of code-switching form ($\chi^2(1, N = 516) = 0.12$, $p = .72$).

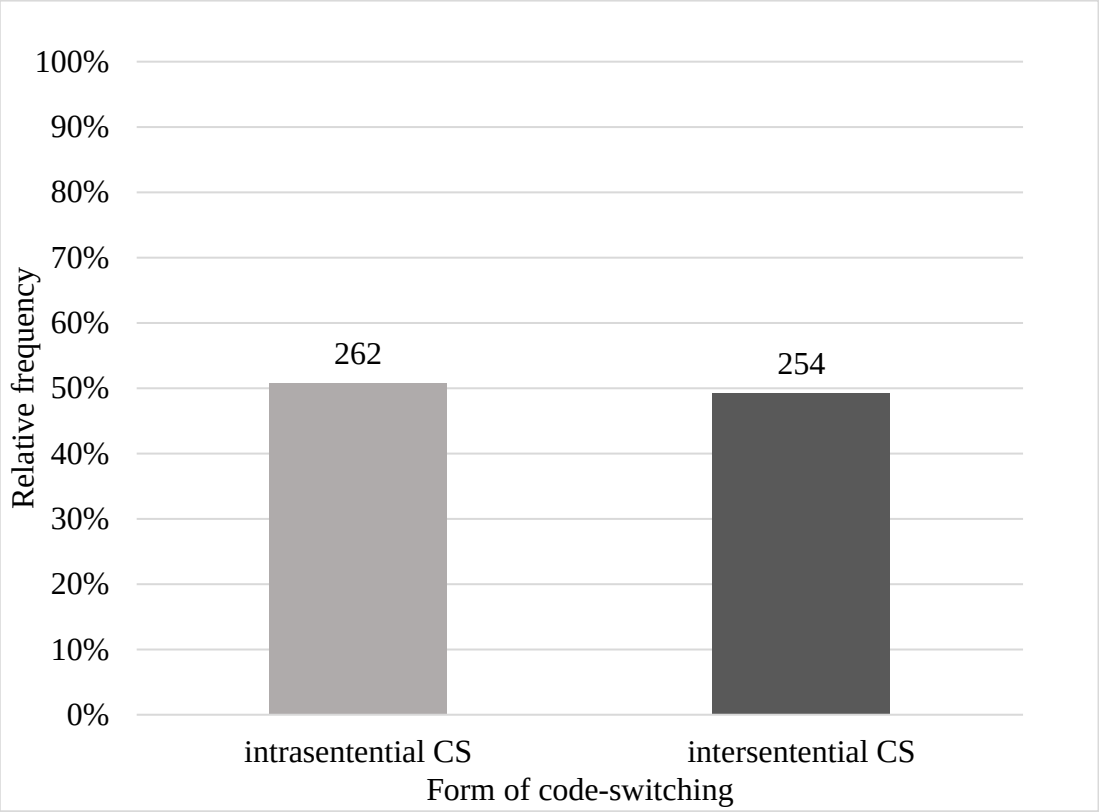

Figure 1. Distribution of relative frequencies of code-switching form (intersentential or intrasentential) with absolute frequencies (N = 516)

4.2 Pattern preference

Figure 2 shows the distribution of the relative frequencies of the patterns of code-switching, either Dutch-English or English-Dutch, with the absolute frequencies. The sentence in (3) is

an example of Dutch-English code-switching, and the sentence in (4) of English-Dutch code-switching.

(3) *Dutch-English*

Ze gewoon gras laten eten helpt ook, maar dat kunnen we niet in voldoende hoeveelheden produceren t.o.v. soja en tarwe - ***sooooo, maybe we should eat less meat***

'Just letting them eat grass also helps, but we can't produce enough of it compared to soy and wheat - sooooo, maybe we should eat less meat'

(4) *English-Dutch*

Nevertheless, ***hele knappe prestatie****!*

'Nevertheless, a great achievement!'

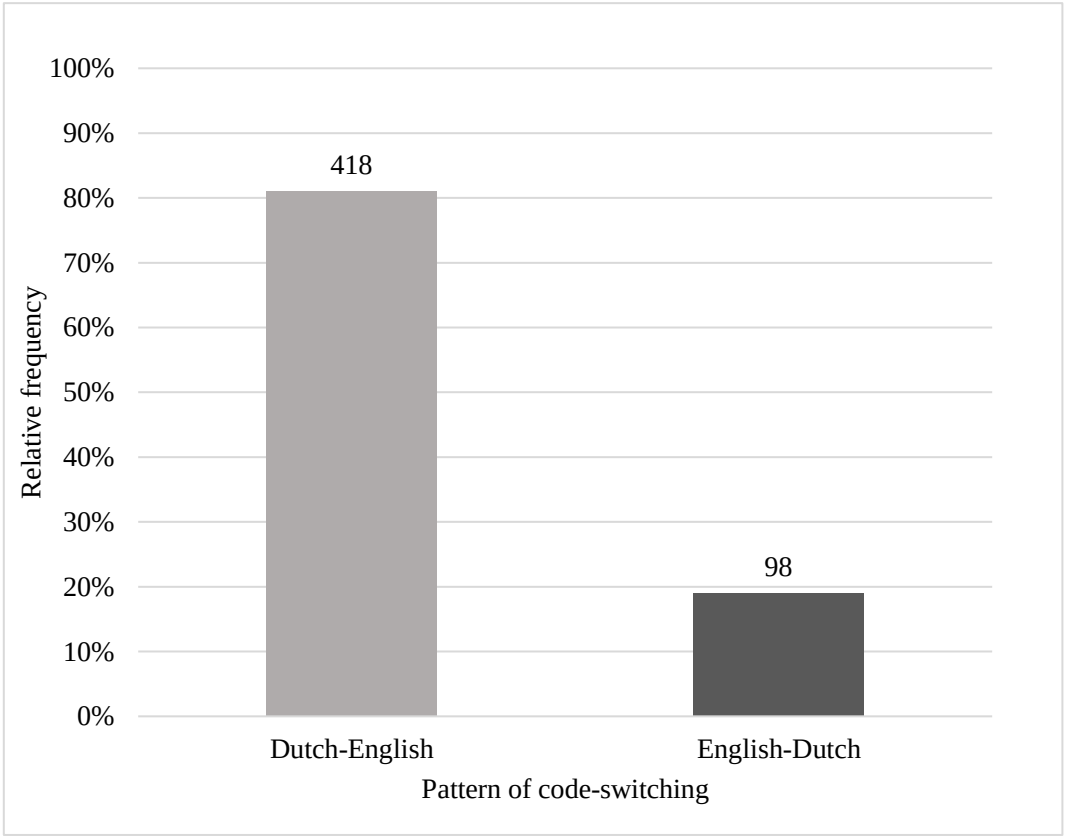

Figure 2. Distribution of relative frequencies of code-switching pattern (Dutch-English or English-Dutch) with frequencies (N = 516)

As shown in figure 2, there is a clear preference in the patterning of code-switching: Dutch-English code-switching (81%) occurs more frequently in the sample data than English-Dutch code-switching (19%). This was tested for significance. According to a χ^2 goodness of fit test, there is a significant difference in the distribution of code-switching pattern ($\chi^2(1, N = 516) = 198.45$, $p < .0001$).

Figures 3 and 4 show the distribution of the forms of code-switching per code-switching pattern (Dutch-English and English-Dutch). For Dutch-English code-switching, intrasentential code-switching (53%) appears to be the dominant form whereas for the English-Dutch pattern of code-switching, intersentential code-switching (59%) appears to be the dominant form; both distributions were tested for significance.

test, there is a significant difference in the distribution of code-switching pattern ($\chi^2(1, N = 516) = 198.45$, $p < .0001$).

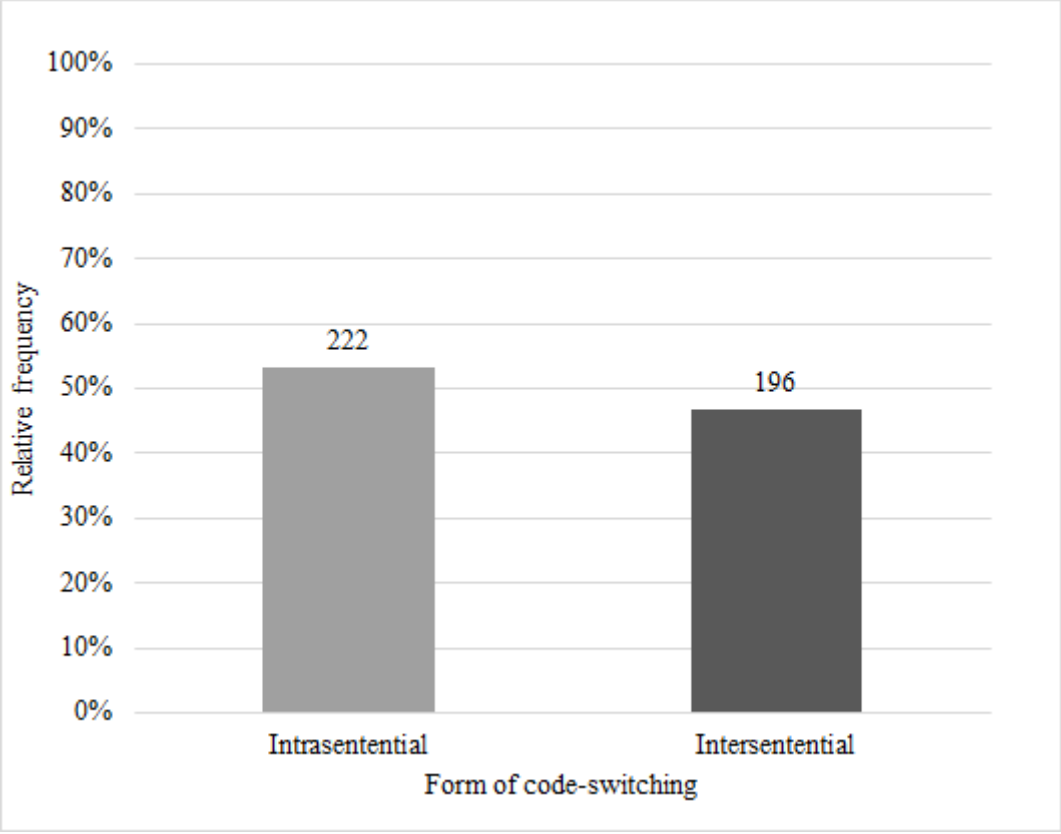

Figure 3. Distribution of relative frequencies of code-switching forms for the pattern Dutch-English with frequencies (N = 418)

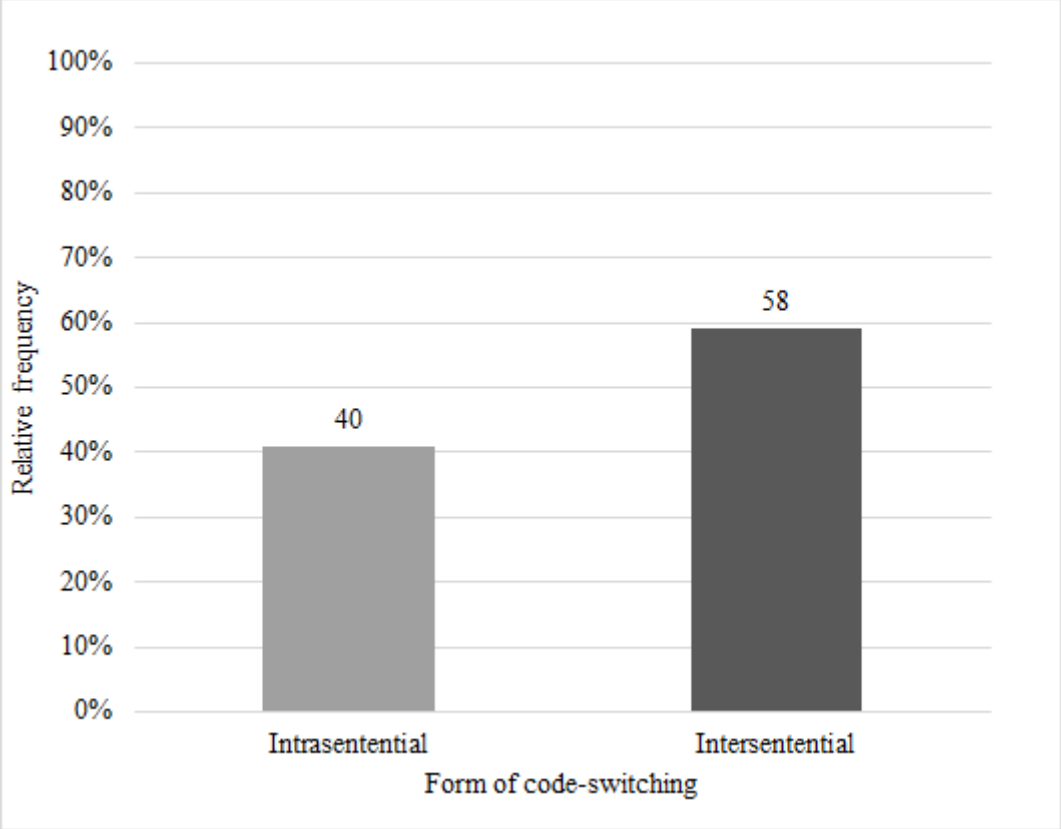

Figure 4. Distribution of relative frequencies of code-switching forms for the pattern English-Dutch with frequencies (N = 98)

According to a χ^2 goodness of fit test, there is no significant difference in the distribution of code-switching form for Dutch-English code-switching ($\chi^2(1, N = 516) = 1.6, p = 0.20$). The distribution of code-switching form for English-Dutch code-switching was also tested with aa χ^2 goodness of fit test, but there is no significant difference in the distribution of code-switching form ($\chi^2(1, N = 516) = 3.306, p = .07$).

4.3 Functions

The functions have been defined by Caparas and Gustilo (2017) and are exemplified with data from the sample in (5) to (16).

(5) *Quote: quoting a famous expression, saying or proverb*

Be the change you want to see in the world. ***Ga je bij Prinses Irene of bij de Jagers?***

'Be the change you want to see in the world. Are you joining Princess Irene or the Hunters?'

(6) *Lexical need: gap in the lexicon*

"Mag ik mijn mening soms niet meer uiten" is in ieder geval de domste ***comeback*** *die er is*

'"Am I not allowed to voice my opinion" is, in any case, the worst comeback in existence'

(7) *Emphasis: adding extra stress to the message*

Een vriendin van me hield ook op en zij stalde al de tabaksproducten die ze nog had uit in een rookhok. ***As a matter of fact, I didn't resist at all.***

'A friend of mine also quit and she put every tobacco product she still had in a smoking area. As a matter of fact, I didn't resist at all'

(8) *Clarification: clarifying the message, such as translation*

I'm pretty sure exposure is something they won't want right now. ***Exposure is ook de term voor blootstelling aan de elementen, wat in dit geval niet echt handig zou zijn.***

'I'm pretty sure exposure is something they won't want right now. Exposure is also the term of exposure to the elements, which, in this case, wouldn't be very beneficial'

(9) *Topic: a preference for one language when discussing a particular topic*

I don't think that penalty that won you guys the game was justified but there were plenty of fouls before that. Anyway.... ***Goed gedaan Nederland! Nu maar hope dat wij de leeuw niet in z'n hempie laten staan tegen de VS.***

'I don't think that penalty that won you guys the game was justified but there were plenty of fouls before that. Anyway.... Great job, the Netherlands! Now let's hope that we don't make the lion look foolish when he's up against the US'

(10) *Formulaic: using common expressions and formulaic language*

Je bedoelt die wereldvreemde extreem linkse lui die geen tegenspraak dulden en elke dissident zonder pardon permabannen? ***Thanks but no thanks****.*

'You mean those out of touch extreme leftist people who don't accept any contradictory messages and permanently ban every dissident without hesitation? Thanks but no thanks.'

(11) *Interjection: sentence fillers or connectors*

Iets andere discussie, maar ***"while we're at it"****, zou men de leeftijds-gebonden minimum loon ook mogen afschaffen.*

'Slightly different topic, but while we're at it, we should get rid of the minimum wage being determined by age.'

(12) *Identity: establishing identity and using community-specific terms*

You're telling me that it's different? Try growing up black in The Netherlands and come back crying. *Ik ben nederlands vriend, kom me niet uitleggen hoe het zit. Je weet niet waar je over praat.*

'You're telling me that it's different? Try growing up black in The Netherlands and come back crying. I am Dutch, pal, don't explain to me how it works. You don't know what you are talking about'

(13) *Audience: limiting the intended audience*

At least it's a welcome alternative to the popular "cheese-eating junkie dutch people who live in mills" stereotype. ***Tussen ons Nederlanders, ik zou een stuk blijer zijn als we "Goldmembers" werden genoemd ipv "tatas", omdat brugwuppen weer eens zo'n verzonnen kutwoord nodig hebben om populair te doen.***

'At least it's a welcome alternative to the popular "cheese-eating junkie dutch people who live in mills" stereotype. For the Dutch people, I'd be a lot happier to be called "Goldmembers", rather than "tatas", just because seventh-graders need a shit made up word to feel cool'

(14) *Native terms: retaining native terminology*

Oh I guess I wasn't clear. It's not about the ***aanvullende beurs*** *being* ***beschikbaar*** *or not.*

'Oh I guess I wasn't clear. It's not about the supplementary grant being available or not.'

(15) *Repetition: clarify or amplify the message*

Hadden we een afgelegen herberg geregeld. In ***the middle of fucking nowhere*** *voor zover dat mogelijk is in Japan. [...] Terwijl ik mijn schoenen uit doe zie ik in mijn ooghoeken iets vel gekleurd. Een plushe oranje paar klompen met een rood-wit-blauw vlaggetje op de hak.* ***TLDR: there's no escape****!*

'We had made sleeping arrangements at a remote inn. In the middle of fucking nowhere, as far as that is possible in Japan. While I was taking off my shoes, I saw something in the corner of my eye, something brightly coloured. A pair of orange, pluche clogs, with a red-white-blue flag on the heel. TLDR: [Too Long, Didn't Read] there's no escape!'

(16) *Strengthen/Soften (S/S*): adjusting the social standing of the speaker*

Het verhaal van *stiff Dutch attitude is:it takes long for a dutch person to warm up to ya..but once they let you in, you re set for life!*

The story of stiff Dutch attitude is: it takes long for a Dutch person to warm up to ya.. but once they let you in, you're set for life!

(17) *Disappointment: expressing disappointment subtly.*

Wij hebben zelf 7 jaar gewacht, onder andere vanwege studies maar ook omdat het nogal een grote beslissing is. ***It sucks, but it's worth it.***

'We waited 7 years ourselves, among other things because of studies, but also because it's such a big decision. It sucks, but it's worth it.'

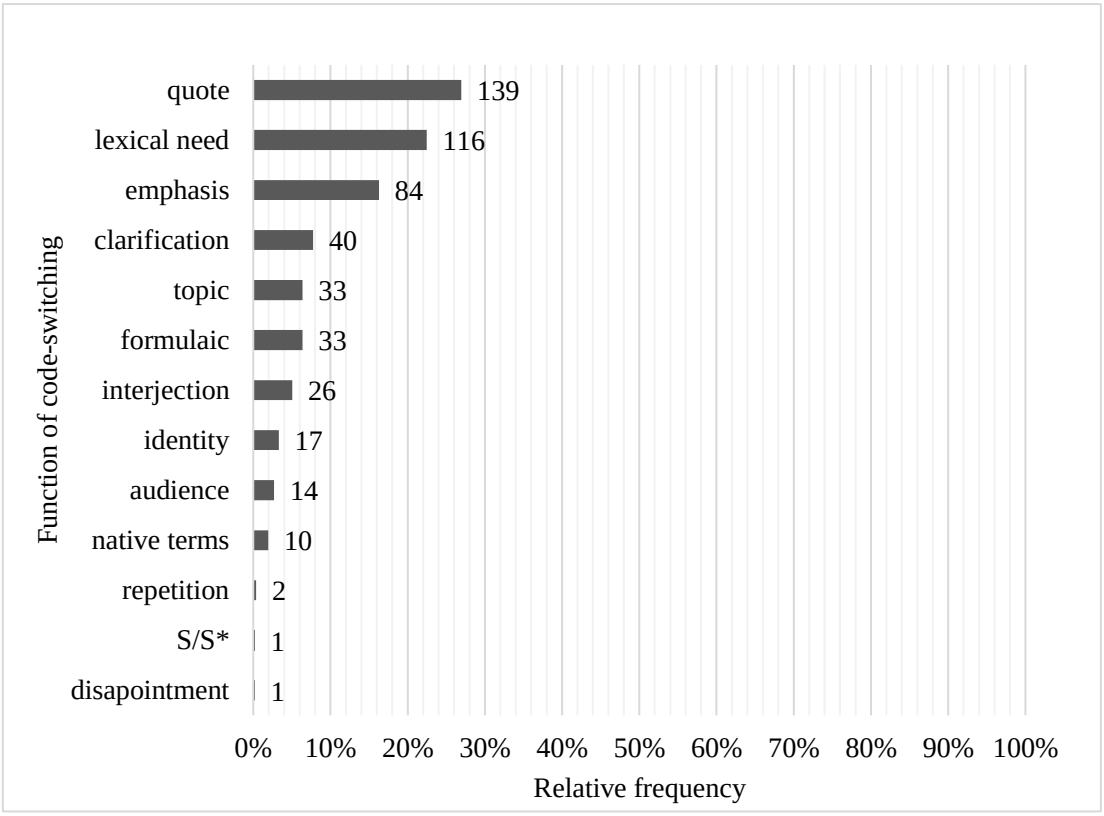

Figure 5. Distribution of relative frequencies of functions of code-switching, with absolute frequencies (N = 516)

Figure 5 displays the distribution of the functions of code-switching, as present in the sample, and shows that code-switching occurs most frequently due to a quote, phrase or saying being repeated (27% of the time), due to the speaker feeling the lexical need to code-switch (22% of the time) and due to the speaker wanting to add emphasis to their message (16% of the time). The differences between the five most frequent functions were tested for significance, as other functions did not have enough occurrences. According to a χ^2 goodness of fit test, there is a significant difference in the distribution of code-switching function ($\chi^2(1, N = 412) = 104.04$, $p < .05$), meaning that these five functions are not equally distributed.

4.4 Interaction between pattern, form and function

Figures 6 and 7 show the distribution of functions of code-switching per code-switching form (intrasentential or intersentential). Intrasentential code-switching appears to most frequently occur due to the speaker feeling the lexical need to code-switch (38% of the time), whereas

intersentential code-switching appears to most frequently fulfil a quoting function (41% of the time).

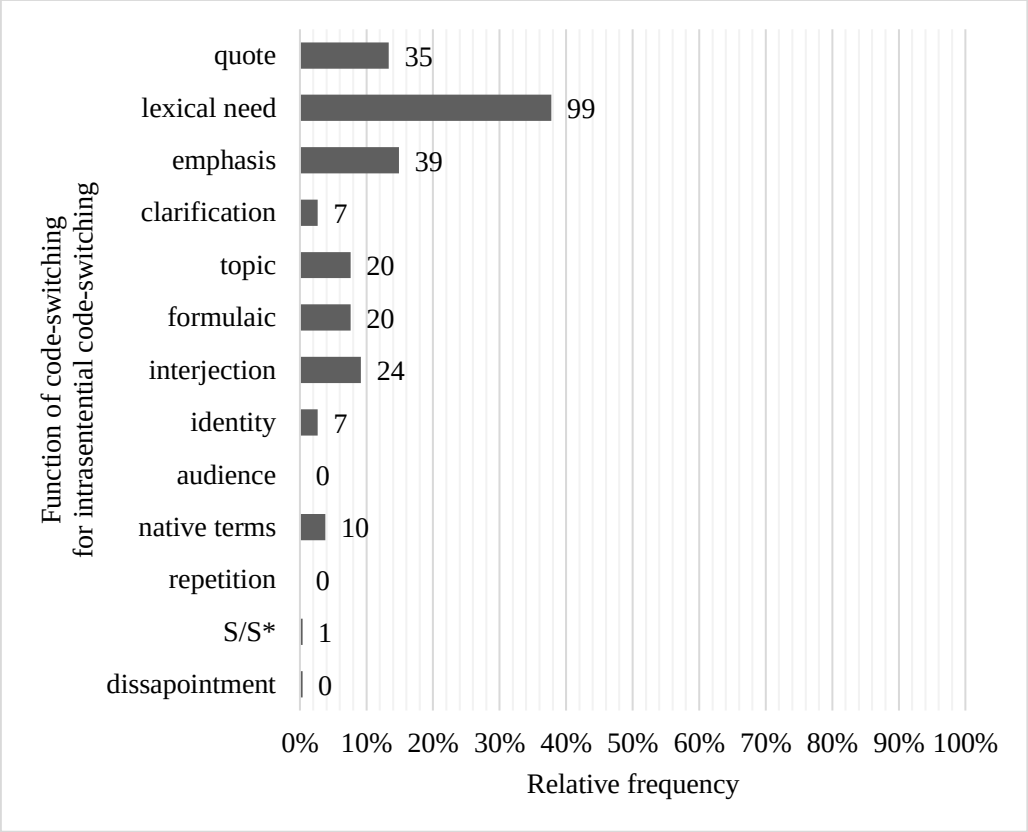

Figure 6. Distribution of relative frequencies of functions of code-switching for intrasentential code-switching, with absolute frequencies (N = 262)

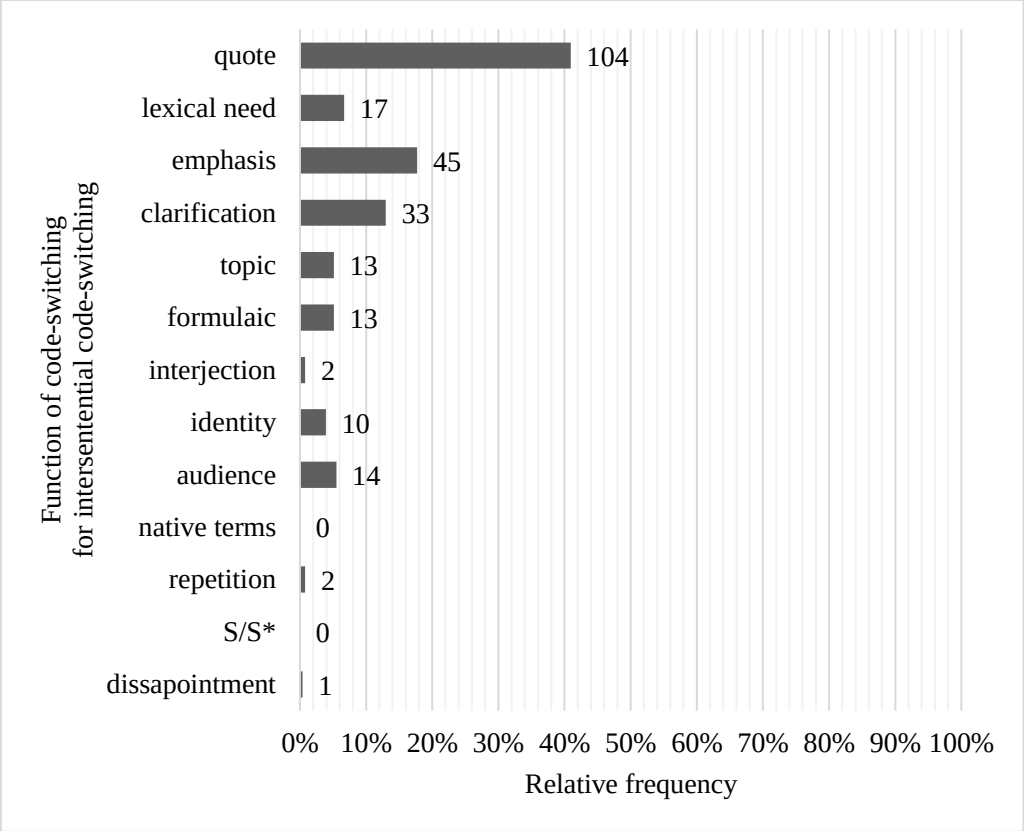

Figure 7. The distribution of the relative frequencies of the functions of code-switching for intersentential code-switching, with absolute frequencies (N = 254)

The differences in distribution for intrasentential and intersentential code-switching were tested for significance, by taking the functions quote, lexical need, emphasis, clarification and topic, as they appear most frequently for both forms of code-switching. According to a χ^2 test, there is a significant difference in the distribution of code-switching function ($\chi^2(1, N = 412) = 110.78$, $p < .00001$).

Figures 8 and 9 show the distribution of functions of code-switching per code-switching pattern (Dutch-English or English-Dutch). For Dutch-English code-switching, the most frequent function appears to be a sentence, phrase or saying being quoted (29% of the time) which is closely followed by the function lexical need (26% of the time). For English-Dutch code-switching, on the other hand, the most frequent function appears to be to add emphasis (26% of the time), which is closely followed by the function of quoting (22%) .

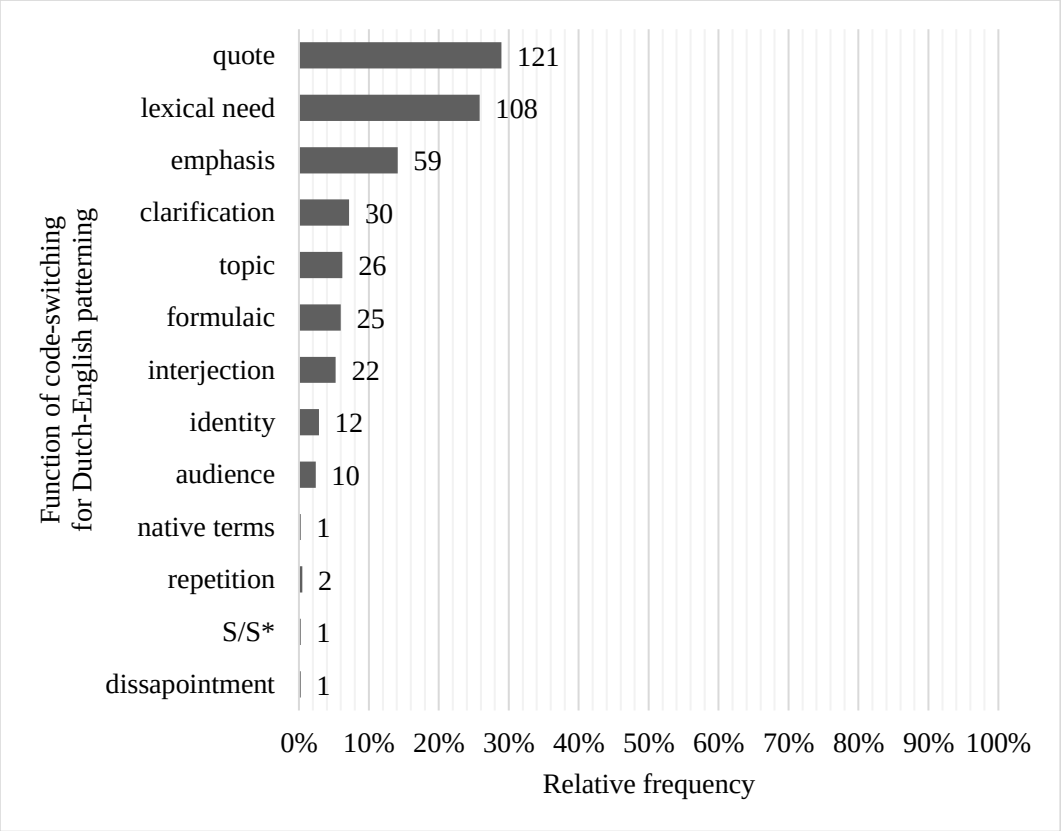

Figure 8. Distribution of relative frequencies of functions of code-switching for Dutch-English patterning, with absolute frequencies (N = 418)

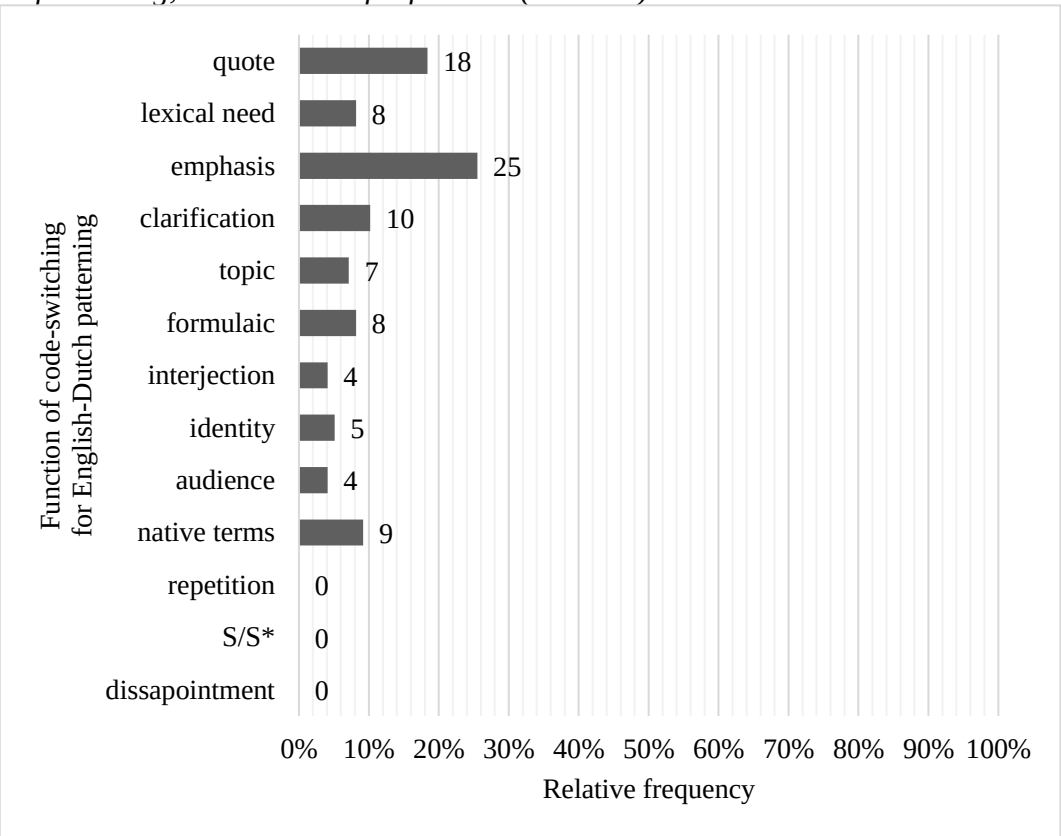

Figure 9. Distribution of relative frequencies of functions of code-switching for English-Dutch patterning, with absolute frequencies (N = 98)

According to a χ^2 test, there is a significant difference in the distribution of code-switching function per pattern ($\chi^2(12, N = 516) = 66.08, p < .001$).

Figures 10 and 11 show the distribution of functions for Dutch-English patterning, separated per code-switching form (intrasentential and intersentential) and figure 12 and 13 show the distribution of functions for English-Dutch patterning per code-switching form. For Dutch-English patterning, the most frequent function for intrasentential code-switching is to fulfil a lexical need (42% of the time), whereas for intersentential code-switching the most frequent function is to quote (46%). For English-Dutch patterning on the other hand, adding emphasis is the most frequent function for both forms (23% for intrasentential and 28% for intersentential code-switching); though for intrasentential code-switching, code-switching to retain native terms occurs equally frequently as adding emphasis (23%). According to a χ^2 test, there is a significant difference in the distribution of code-switching function per form for Dutch-English patterning ($\chi^2(12, N = 418) = 135.65, p < .001$). There was also a significant difference in the distribution of code-switching functions per form found for English-Dutch patterning, according to a χ^2 test ($\chi^2(9, N = 98) = 32.99, p < .001$). This indicates that there might be an interaction between form and pattern, as well as an interaction between function and pattern. This is further investigated in a logistic regression analysis.

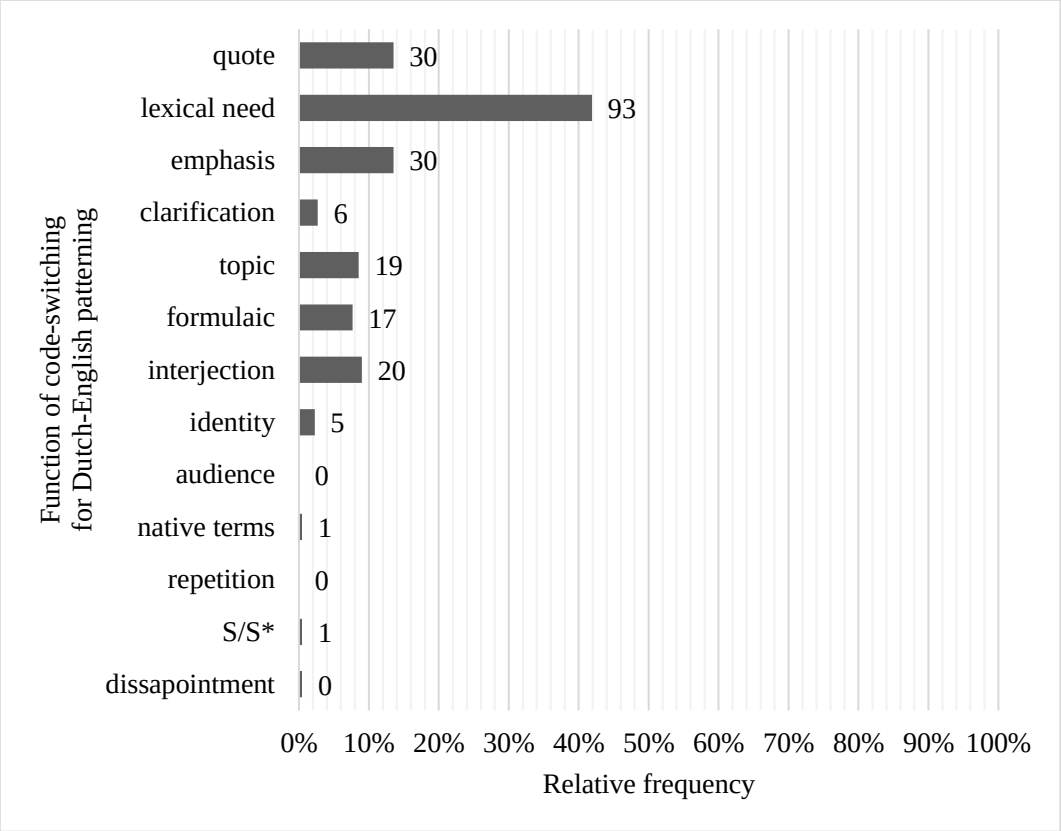

Figure 10. Distribution of relative frequencies of functions of code-switching for Dutch-English patterning for intrasentential code-switching, with absolute frequencies (N = 418).

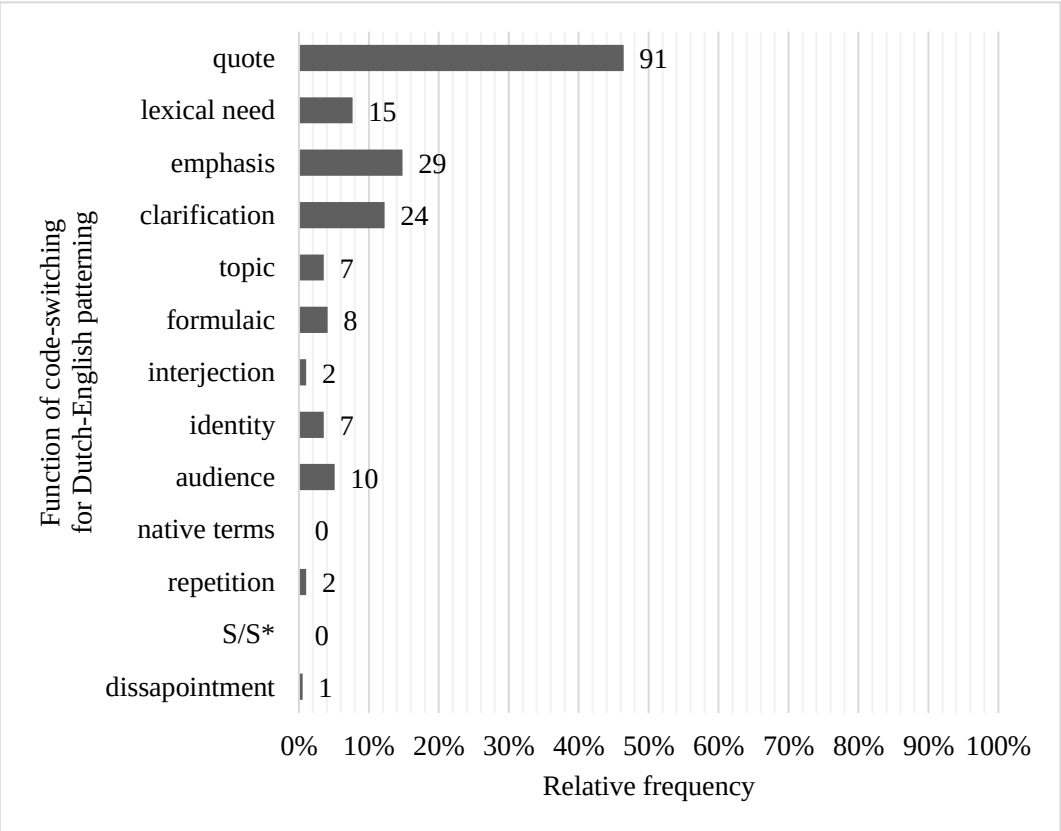

Figure 11. Distribution of relative frequencies of functions of code-switching for Dutch-English patterning for intersentential code-switching, with absolute frequencies (N = 418)

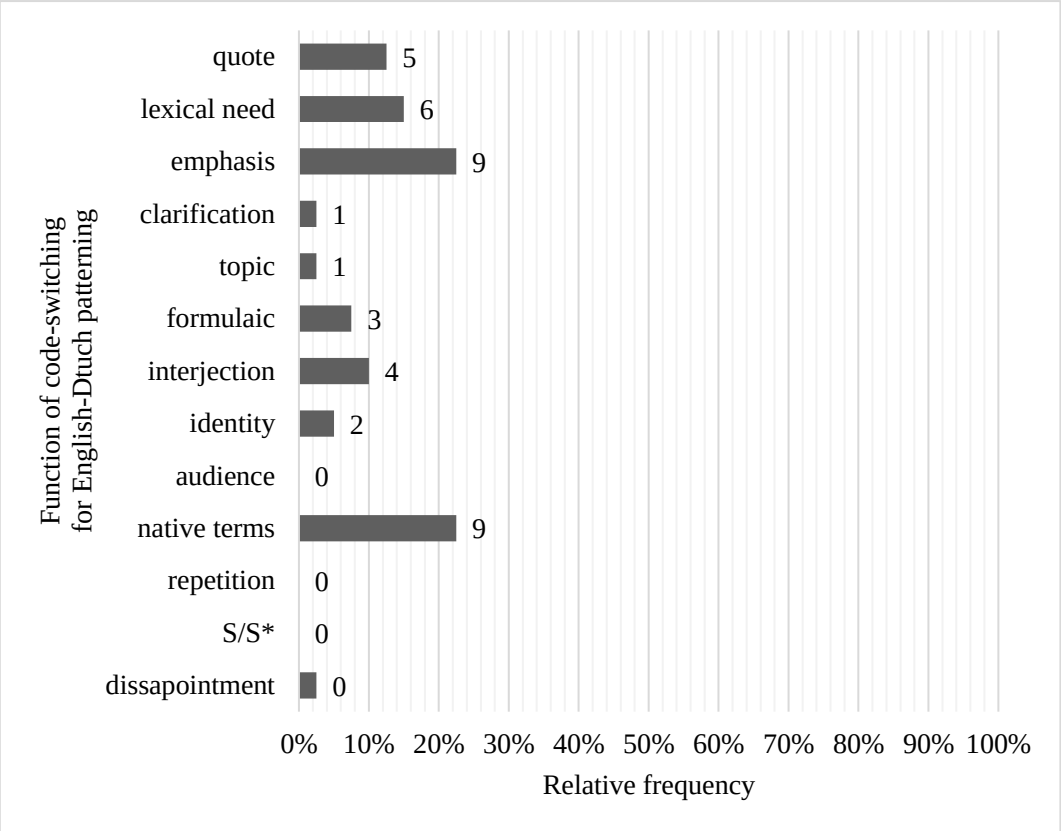

Figure 12. Distribution of relative frequencies of functions of code-switching for English-Dutch patterning for intrasentential code-switching, with absolute frequencies (N = 98)

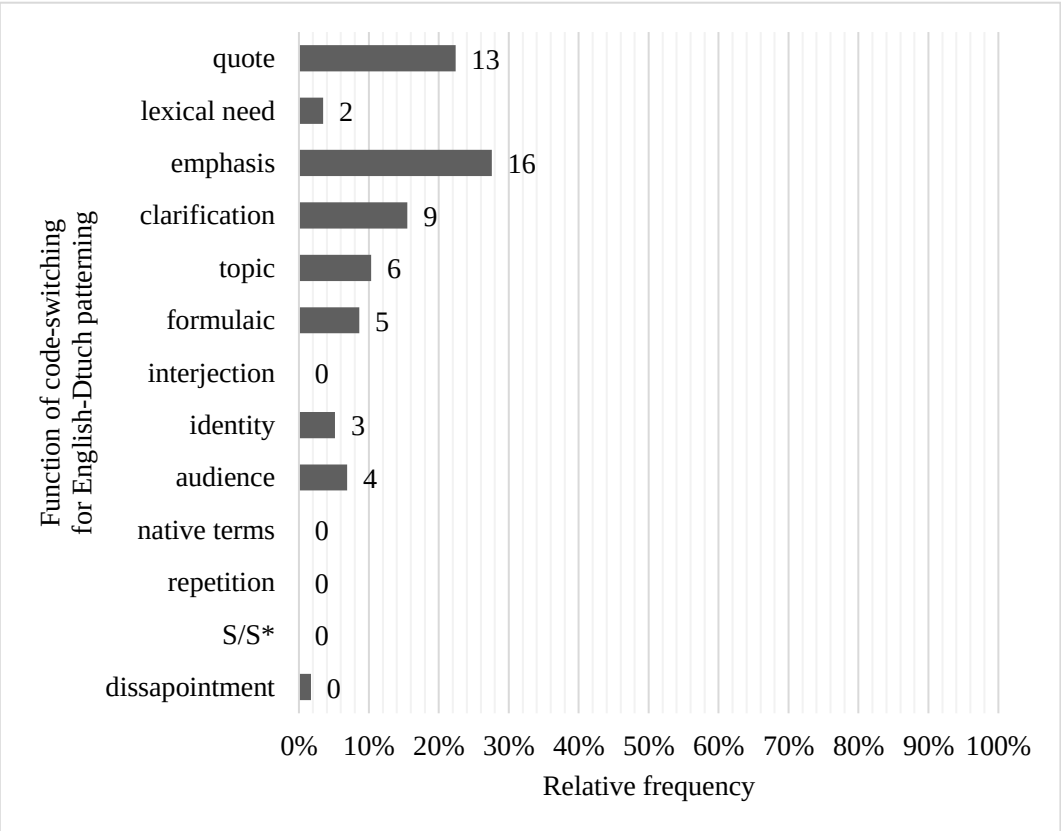

Figure 13. Distribution of relative frequencies of functions of code-switching for English-Dutch patterning for intersentential code-switching, with absolute frequencies (N = 98)

In order to further investigate interaction between code-switching pattern, function and form, logistic regression analysis was applied to the data. Pattern was taken to be the dependent variable (levels: Dutch-English, English-Dutch) with form (levels: Intrasentential and Intersentential) and function (levels: Topic, Clarification, Emphasis, Lexical Need, Quote, Formulaic) as predictors. The levels for function were chosen based on their frequency; the function had to occur at least 30 times in order to be considered, to ensure the results were not skewed. A simultaneous binary logistic regression analysis was applied to the data and its results are displayed in table 1 and plotted in figure 14.

The baseline model without any predictors added had an accuracy of 81.6%; it always guessed that the pattern would be Dutch-English, as that is the most frequent pattern present in the data-set. Then, the predictors of function and form were added, as well as an interaction between function and form as predictors, all added simultaneously. The model with the added predictors was tested with a χ^2 test by comparing the goodness of fit of the model with predictors to the goodness of fit of the baseline model, to test whether adding the predictors improved the model fit. The χ^2 test was significant ($\chi^2(11, N = 477) = 45.625$, $p < .001$), meaning that the model with predictors fits the data significantly better than a model with just the intercept (Dutch-English, in this case). Nagelkerke's R2, which was found to be 0.148, suggests that the model explains roughly 15% of the variation in the outcome. Table 1 shows the regression coefficient (B), the Wald statistic (to test the statistical significance) and the Odds Ratio (Exp (B)) for each variable category. There seems to be a significant overall effect of form (Wald = 5.72, df = 1, $p < .05$) and function (Wald = 12.95, df = 5, $p < .05$) on pattern, but no significant interaction between function and form (Wald = 6.58, df = 5, $p = 0.254$). There was a significant interaction between the form of intersentential code-switching and the function of quoting (Wald = 5.447, df = 1, $p < .05$), but this was the only interaction between form and function.

Table 1

Results of the Simultaneous Logistic Regression analysis on Pattern by Form and Function with the best fit and the least predictors

Predictor	*B*	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
Intersentential	2.790	1.167	5.716	1	.017*	16.286
Function			12.953	5	.024*****	
Clarification	.865	1.476	.344	1	.558	2.375
Emphasis	1.983	1.077	3.393	1	.065	7.265
Lexical	.193	1.109	.030	1	.862	1.213
Quote	1.240	1.118	1.230	1	.267	3.455
Formulaic	1.210	1.202	1.013	1	.314	3.353
Form * Function			6.582	5	.254	
Inter * Clarification	-1.772	1.624	1.190	1	.275	.170
Inter * Emphasis	-2.234	1.243	3.233	1	.072	.107
Inter * Lexical	-2.118	1.450	2.135	1	.144	.120
Inter * Quote	-2.990	1.281	5.447	1	.020*	.050
Inter * Formulaic	-1.526	1.442	1.119	1	.290	.217
Constant	-2.944	1.026	8.236	1	.004	.053

Notes. Nagelkerke's R2 = 0.148. * = $p < .05$

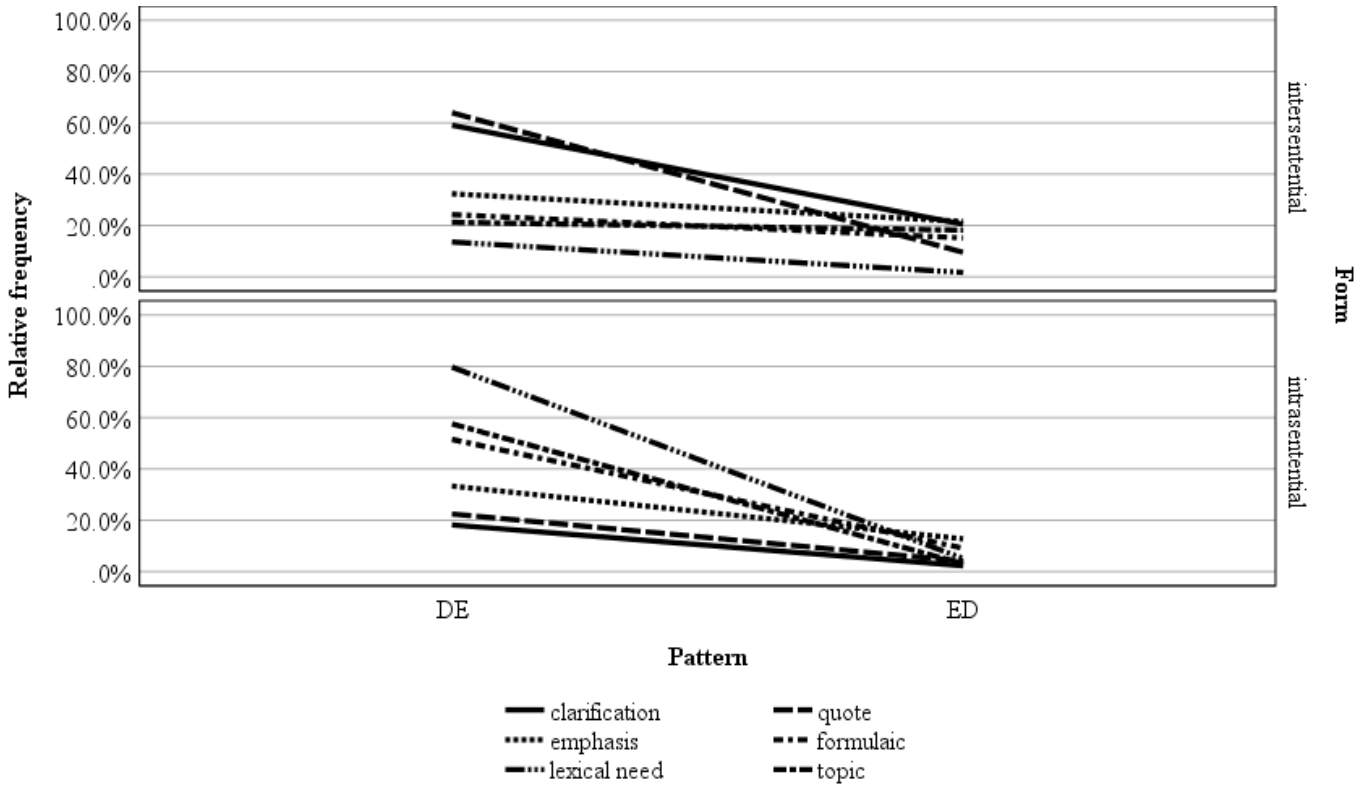

Figure 14. Interaction between code-switching pattern, form and function as used in logistic regression analysis.

Figure 14 shows that for Dutch-English patterning, quoting is the most likely function for intersentential code-switching and fulfilling lexical need is the most likely function for intrasentential code-switching. For English-Dutch patterning on the other hand, the most likely function for intersentential code-switching is to clarify, and for intrasentential code-switching is to add emphasis.

5. Discussion

5.1 Preference for form

The hypothesis for the first research question was that there was a preference for intersentential code-switching. This is due to the acquisition of English as a second language, the hypothesis that intersentential code-switching being less grammatically complex, in addition to the observation that intersentential code-switching tends to be the main form of

code-switching in Internet forums (Androutsopoulos, 2006, as cited in Dorleijn & Nortier, 2008), while intrasentential code-switching was noted to be the main form for oral code-switching (Koban, 2012; Li, Yu and Fung, 2012). Analysis found no significant differences in the distribution of code-switching form, meaning that there is no clear preference for intrasentential or intersentential code-switching. This could indicate that posts from r/theNetherlands are a hybrid between online written discourse and spoken discourse, as online discourse prefers intersentential code-switching and spoken discourse prefers intrasentential code-switching. This may also be due to the sample size, as only 500 out of the 3042 code-switched posts on r/theNetherlands have been analysed, which means that the patterns visible in the sample may not be representative of the data set as a whole.

5.2 Preference for pattern

There had been no research into the patterning of Dutch-English code-switching, meaning that the second research question was exploratory of nature, investigating both options of English-Dutch and Dutch-English code-switching, in addition to eventual interactions with code-switching form. Analysis found that Dutch-English was the dominant code-switching pattern visible in the sample as a whole. Additionally, for intrasentential code-switching, a preference was found for Dutch-English, whereas for intrasentential code-switching a preference for English-Dutch was found, though these were not found to be significant.

5.3 Preferences for function

The functions of code-switching present in the data set were also investigated. The list of possible functions was taken from Caparas and Gustilo (2017) and functions for Dutch-English code-switching were hypothesised to function similarly to what they found. Contrary to Caparas and Gustilo's (2017) findings, the most frequent functions of code-switching

being to fulfil lexical need, to add clarification, and to add emphasis, this study found that the most frequent functions of code-switching for Dutch-English code-switching were to repeat a quote, phrase or saying, to fulfil lexical need and to add emphasis. This implies that Dutch-English bilingual speakers prefer to use English to quote someone, more so than English-Filipino bilinguals do.

5.4 Interactions between pattern, form and function

There was a significant difference for functions for intrasentential code-switching when compared to intersentential code-switching; intrasentential code-switching was most frequently used to fulfil lexical need, add emphasis and to quote a phrase or saying, whereas intersentential code-switching was most frequently used to quote a phrase or saying, add emphasis and add clarification. Intrasentential code-switching is slightly more grammatically complex than intersentential code-switching and it appears that the forms are used for different purposes, though more research into this is necessary to draw meaningful conclusions from this. This indicated a possible interaction effect, though the logistic regression analysis did not reflect this.

There was also a significant difference found for the functions of Dutch-English code-switching when compared to English-Dutch code-switching; when looking at Dutch-English code-switching regardless of code-switching form, the most frequent functions are to quote, to fulfil lexical need and to add emphasis, whereas adding emphasis, quoting and adding clarification are the most frequent functions of English-Dutch. It appears that different language patterns fulfil different functions, though the relationship between function and code-switching pattern has not yet been investigated and could therefore be causal in either direction, or could depend on other factors not investigated in this study; more research into this is necessary to conclude causality for certain.

Splitting the Dutch-English pattern and English-Dutch pattern into intersentential and intrasentential code-switching allowed further investigation of the distribution of code-switching functions. A significant difference was found in the distributions of functions when comparing intersentential and intrasentential code-switching for Dutch-English, and for English-Dutch. This supports the hypothesis that different code-switching forms may fulfil different functions, as well as the idea that code-switching patterns influence or are influenced by code-switching functions.

The interactions between pattern, form and function in the Dutch-English community were tested for significance. Binary logistic regression analysis was applied to the data in order to explore these interactions. Form (Intrasentential and Intersentential) and function were found to be significantly associated with pattern (Dutch-English and English-Dutch), and can explain almost 15% of the variance of code-switching pattern in the dataset. However, it is important to note that they cannot predict the outcome for individual code-switching posts very well. This indicates that though form and function influence code-switching pattern, they do not determine it. Code-switching can vary per speaker for various reasons, and these results reflect the fact that code-switching patterns are not only explained by form and function, but might also be explained by other predictors. More research is necessary in order to identify other factors that influence English-Dutch code-switching, or code-switching in general.

6. Conclusion

This thesis intended to investigate code-switching form, patterns and functions in an online Dutch-English bilingual community, as well as possible interactions between these three characteristics, as this bilingual community had yet to be explored and very little research on written code-switching had been done.

There was no clear preference of code-switching form present in the data set, even though online discourse had been found to show a preference for intersentential code-switching and spoken discourse appears to prefer intrasentential code-switching. A clear preference for the Dutch-English code-switching pattern was visible, though it remains to be investigated why this is the case. Following Caparas and Gustilo's (2017) list of code-switching functions, the most frequent functions of code-switching on r/theNetherlands were to quote, to fulfil a lexical need and to add emphasis; though it is important to keep in mind that it is also possible that speakers code-switched because it was easier for them, or had no particular function in mind when writing their posts. There were interactions between pattern and form and pattern and function present, which explained part of the variance of patterning visible, though there was no interaction between form and function; other factors might be more predictive of code-switching pattern.

This study had several limitations. It is unclear how representative the used sample is of the r/theNetherlands community as a whole, meaning that the patterns visible in the data set do not necessarily apply to the population. Another limitation was the fact that only one annotator was used to code the data. Ideally, multiple annotators would have been used to assign code-switching pattern, form and function, ensuring annotator agreement in order to prevent bias and ensure correct annotation. Furthermore, the list of code-switching functions used could use more clarification and specification to enable more robust analysis. Some posts could have been coded into multiple categories, as the categories had fuzzy boundaries; using multiple annotators and more specific categories would have improved the analysis by disambiguating categories. Additionally, no measures were taken to exclude loanwords from the data set, meaning it is possible that both older and newer loanwords present in the sample were incorrectly identified as code-switching.

Further research could use the whole subreddit to investigate code-switching, which was beyond the scope of this thesis. Further research could investigate the reasoning behind code-switching patterns (Dutch-English and English-Dutch) in order to explain the patterns that were found in the data set. Further investigation of code-switching form (Intrasentential and intersentential) is also necessary as this thesis hypothesised that different code-switching forms could be used for different purposes, though no conclusive evidence for this hypothesis was found. More research into the relationship between code-switching function and code-switching pattern is also necessary, as causality is yet to be determined, as well as other possible predictors influencing function and/or pattern.

A different direction of research could be to investigate other English bilingual communities on Reddit. Reddit contains many English bilingual communities, other than Dutch-English, such as English-Spanish, English-Turkish, and English-Arabic that have yet to be investigated; this could further support the idea of English as the dominant language online. It would also be possible to compare Dutch-English online code-switching to Dutch-English spoken code-switching, in order to investigate whether the platform, the Internet, influences the frequency, form, pattern or function of code-switching.

7. References

Androutsopoulos, J. (2013). Code-switching in computer-mediated communication. In S. C. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 659–686). Berlin, Germany: Mouton de Gruyter. Retrieved from https://www.redalyc.org/revista

Baumgartner, J. (2014). *Pushshift* [Computer software]. Retrieved from redditsearch.io

Berk-Seligson, S. (1986). Linguistic constraints on intrasentential code-switching: A study of Spanish/Hebrew bilingualism. *Language in society*, *15*, 313-348. doi:10.1017/S0047404500011799

Bhat, G., Choudhury, M., & Bali, K. (2016). Grammatical constraints on intra-sentential code-switching: From theories to working models. *arXiv preprint arXiv:1612.04538*.

Caparas, P., & Gustilo, L. (2017). Communicative aspects of multilingual code switching in computer-mediated communication. *Indonesian Journal of Applied Linguistics*, *7*, 349-359. doi:10.17509/ijal.v7i2.8137

Cayer, R., & Sacks, R. (1979). Oral and written discourse of basic writers: Similarities and differences. *Research in the Teaching of English*, *13*, 121-128. Retrieved from https://www.jstor.org/stable/40170748

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. In

Crystal, D. (2006). *Language and the Internet*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511487002

del Pilar Agustín-Llach, M. (2017). The impact of bilingualism on the acquisition of an additional language: Evidence from lexical knowledge, lexical fluency, and (lexical) cross-linguistic influence. *International Journal of Bilingualism*, *23*, 888-900. doi:10.1177/1367006917728818

Dorleijn, M., & Nortier, J. (2008). Code-switching and the Internet. In B. Bullock, A. Toribio (Eds.), *The Cambridge handbook of linguistic code-switching* (pp. 127-141). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511576331

Gaffney, D., & Matias, J. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*, *13*(7), 1-13. doi:10.1371/journal.pone.0200162.

García, O., & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. London, England: Palgrave Macmillan. doi:10.1057/9781137385765

Gardner-Chloros, P., & Weston, D. (2015). Code-switching and multilingualism in literature. *Language and Literature, 24*, 182-193. doi:10.1177/0963947015585065

Grosjean, F. (1982). *Life with two languages: An introduction to bilingualism*. Cambridge, MA: Harvard University Press.

Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press. doi:10.4159/9780674056459

Hinrichs L. (2006). *Code-switching on the web*. Amsterdam, the Netherlands: John Benjamins.

Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A companion to digital humanities.* Oxford, England: Blackwell. Retrieved from digitalhumanities.org/companion/

Honnibal, M., & Montani, I. (2015). spaCy [computer software: Python library]. Berlin, Germany: ExplosionAI. Retrieved from spacy.io

Jaworska, S. (2014) Playful language alternation in an online discussion forum: the example of digital code plays. *Journal of Pragmatics, 71,* 56--68. doi:10.1016/j.pragma.2014.07.009

K. Toutanova, H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 383-389). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/v1/P14-2063

Koban, D. (2013). Intra-sentential and inter-sentential code-switching in Turkish-English bilinguals in New York City, US. *Procedia-Social and Behavioral Sciences*, *70*, 1174-1179. doi:10.1016/j.sbspro.2013.01.173

Leppänen, S. (2012). Linguistic and generic hybridity in web writing. In M. Sebba., S. Mahootian, & C. Jonsson (Eds.), *Language mixing and code-switching in writing: Approaches to mixed-language written discourse* (pp. 233-254). London: Routledge. doi:10.4324/9780203136133

Li, Y., Yu, Y., & Fung, P. (2012). A Mandarin-English code-switching corpus. *Proceedings of the Language Resources Evaluation Conference, Istanbul,* 2515-2519. Retrieved from http://www.lrec-conf.org/

Massanari A. (2015). *Participatory culture, community, and play: Learning from Reddit*. New York, NY: Peter Lang. doi:10.3726/978-1-4539-1501-1

McClure, E. (2001). Oral and written Assyrian-English codeswitching. In W. Winter (Ed.), *Codeswitching worldwide II* (pp. 157–191). Berlin, Germany: Mouton de Gruyter.

Milroy, L. (1980). *Language and social networks*. Baltimore, MD: University Park Press.

Minocha, A., Reddy, S., & Kilgarriff, A. (2013). *Feed corpus: An ever growing up-to-date Corpus*. Paper presented at Corpus Linguistics 2013: The 8th Web as Corpus Workshop, Lancaster University, UK. Retrieved from https://www.aclweb.org/

Myers-Scotton, C. (1993). *Social motivations for codeswitching: Evidence from Africa*. Oxford, England: Clarendon Press.

Nilep, C. (2006). "Code switching" in sociocultural linguistics. *Colorado Research in Linguistics, 19*, 1-22. doi:10.25810/hnq4-jv62

Ortega, Mireia. (2008). Cross-linguistic influence in multilingual language acquisition: The role of L1 and non-native languages in English and Catalan oral production. *Íkala, Revista de Lenguaje y Cultura,* 13(19), 121-142.

Pahta, P., & Nurmi, A. (2011). Multilingual discourse in the domain of religion in medieval and early modern England: A corpus approach to research on historical code-

switching. In H. Schendl, and L. Wright (Eds.), *Code-switching in Early English* (pp. 219-251). Berlin, Germany: De Gruyter Mouton. doi:10.1515/9783110253368

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Language Resources Evaluation Conference, Malta,* 1320-1326. Retrieved from http://www.lrec-conf.org/

Piperski, A., Belikov, V., Kopylov, N., Selegey, V., & Sharoff, S. (2013). Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In S. Evert, E. Stemle, P. Rayson (Eds.), *Proceedings of the 8th Web as Corpus Workshop* (pp. 24-29).

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: Toward a typology of code-switching. *Linguistics*, *18*, 581-618. doi:10.1515/ling.1980.18.7-8.581

Rabinovich, E., Sultani, M., & Stevenson, S. (2019). CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. *arXiv*. Advance online publication. *arXiv:1908.11841*.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43-55. doi:10.1080/01638538409544580

Redouane, R. (2005). Linguistic constraints on codeswitching and codemixing of bilingual Moroccan Arabic-French speakers in Canada. In J. Cohen, K. McAlister, K. Rolstad, J. MacSwan (Eds.), *Proceedings of the 4th International Symposium on Bilingualism* (pp. 1921-1933). Somerville, MA: Cascadilla Press.

Saleem H., Dillon K., Benesch S., & Ruths D. (2016). *A web of hate: Tackling hateful speech in online social space*s. Paper presented at the Language Resource and Evaluation Conference, Portorož. arXiv:1709.10159

Sebba, M. (2012). Researching and theorising multilingual texts. In Sebba. M., Mahootian, S., & Jonsson C. (Eds). *Language mixing and code-switching in writing: Approaches to mixed-language written discourse* (pp. 1-26). London, England: Routledge. doi:10.4324/9780203136133

Sebba. M., Mahootian, S., & Jonsson C. (2012). *Language mixing and code-switching in writing: Approaches to mixed-language written discourse.* London, England: Routledge. doi:10.4324/9780203136133

Stroud, C. (1998). Perspectives on cultural variability of discourse and some implications for code-switching. *Code-switching in conversation: language, interaction and identity, 19*, 321-348. doi:10.2307/3736717

Wei, L. (1998). The "why" and "how" questions in the analysis of conversational code-switching. In P. Auer (Ed.), *Code-switching in conversation: Language, interaction and identity,* (pp. 156–179). New York, NY: Routledge.

Weinreich, U. (1968). *Languages in contact: Findings and problems*. The Hague, the Netherlands: De Gruyter Mouton.

Winford, D. (2003). *An introduction to contact linguistics*. Oxford: Blackwell.