Can Internet Data Help to Uncover Developing Preferred Multilingual Usage Patterns?
An Exploration of Data from Turkish-Dutch Bilingual Internet Fora

Margreet Dorleijn
University of Amsterdam
M.Dorleijn@uva.nl

Abstract

This paper discusses the extent to which two characteristics of digital data make such data suitable for detecting preference patterns in code switching: an absence of paralinguistic disambiguation- cues and its extra-linguistic ‘context-freeness'. This paper reports on the exploration of a 219,536 word Dutch-Turkish digital data corpus compiled from bilingual internet fora. It describes both macro-sociolinguistic patterns of language choice as well as micro-linguistic contact features in bilingual data, comparing both macro and micro results with what is known from the sociolinguistic literature in general, and Turkish-Dutch code switching and contact linguistic literature in particular. The data are analysed qualitatively and quantitatively. Focus is on the analysis of densely bilingual data of the type that has been called in the literature ‘mixed language' (Auer, 1999), ‘intimate switching' (Poplack, 1980), or ‘unmarked switching' (Myers-Scotton, 1983; 1993b). It is argued that data of this type of intensive language mixing should display a certain degree of predictability since it is generally perceived of as the most effortless way of speaking by its users. It is demonstrated that recurring patterns can be found in the data, both on the macro-level of language choice and the micro-level of lexical choice, as well as in code switching patterns, and lexico-semantic choices, and it is argued that in these patterns principles of transparency and frequency of exposure may be an explanatory factor.

Keywords

language mixing – language choice – structural contact features – internet data
1 Introduction

It is generally assumed that in relatively recent\(^1\) immigrant communities such as the Turkish Dutch community, code switching\(^2\) is to a large extent instable and unpredictable. This can be attributed to the instability of the community’s sociolinguistic circumstances in general: the process of language shift that, in migrant communities, is generally completed within three generations or so (Winford, 2003; Fase et al. 1992; Boeschoten, 1992).

The instable linguistic situation is thought to prevent code switching from stabilising into a conventional variety, and also from developing into a stabilised mixed language. Instability implies that code switching should be rather unpredictable, random and therefore quite complex for both speaker and listener; complex for the speaker because s/he needs to decide constantly which linguistic element should be selected in which (conversational or linguistic) situation (see, e.g. Myers-Scotton, 1995: 236 for a discussion of the intricate decision process of code switchers), and complex especially for the listener, who would not know which form/s his/her interlocutor is going to employ next.

The complexity of code switching can not only be attributed to its instability and unpredictability, but also to the fact that a speaker has two distinct linguistic systems (in the ‘traditional’ essentialist take on the concept of a language) to choose from, that may indeed be activated simultaneously in a ‘bilingual mode’ (Soares and Grosjean, 1984; Grosjean, 2008). Such a ‘bilingual mode’ may thus facilitate code switching, but still leaves speaker and listener with a double linguistic inventory to select, or expect, items from. The processing cost (especially for the listener) of multilingual language has been shown in experimental environments in numerous psycholinguistic studies (see e.g. Chauncy et al. 2012; Costa and Santesteban, 2004; Meuter and Allport, 1999; Moreno et al. 2002), and for a critical review of models of bilingual processing: Finkbeiner, Gollan, and Caramazza, 2006). The following quote is illustrative:

 Speakers comfortable in several languages often use more than one language within a discourse or even within a sentence, particularly when other bilinguals are present. Although such language switching or code switching may have social or communicative benefits, both comprehension and production are slower with mixed-language stimuli, and in

---

\(^1\) cf. section 2 for a discussion of the Turkish community.

\(^2\) The term code switching is for the moment used as a cover term for all types of language mixing. The definition relevant for this paper will be discussed in 2.1.
comprehension, this effect is even more marked when the input language changes unpredictably.
CHAUNCEY et al., 2012: 291

This may seem to be a rather obvious fact intuitively. However, psycholinguistic studies are carried out in an experimental, replicable setting and not concerned with naturally occurring speech. Thus, this psycholinguistic evidence contrasts with the comments often made by the actual code switchers themselves, that they frequently view code switching as the most effortless and facile way of speaking, as is illustrated by quotations such as the following:

This whole process of code-switching is done mostly out of laziness, for if I searched long enough for the correct word, I would eventually find it...'.
GROSJEAN, 1982: 184

When I speak to another Russian-English bilingual, I don’t speak carefully and often the languages blend. This also happens when I am tired or exited or angry.
GROSJEAN, 1982: 150

There is a vast body of literature in which it is described how code switching is constrained by structural, cognitive and functional factors, from which it can be concluded that code switching is not entirely random.

However, of one particular type of code switching, namely the type of frequent switching where each individual switch does not seem to be functionally motivated (see discussion in section 2), it is not yet established whether there are specific constraints on the selection of individual items. This paper seeks to investigate the extent to which code switching in the Dutch-Turkish bilingual code is really unpredictable and seeks to address the following two questions: is language mixed randomly, although constrained by either referential, structural, socio psychological and discursive factors, or do certain patterns become manifest which would make the mixed code less random?

In section 2 the central question of this paper, namely the question whether preference patterns can be observed in the Turkish Dutch bilingual mode will be elaborated on. In this section, it is argued that code mixing in recent immigrant communities may in fact be more predictable than is often assumed. In 3, the advantages and disadvantages of internet data for answering this
question will be discussed in a general fashion, and the position is maintained that this type of data may be particularly suitable to detect such preference patterns. In 4 the Turkish-Dutch data set and the research methods are presented. Section 5 reports on and interprets the results.

2 Is Dutch-Turkish Mixed Language on its Way to Becoming a Stable Variety?

Migration from Turkey to the Netherlands started in the 60ties of the former century. Those who arrived were mainly (male) migrant workers. About a decade later their families started to join them. Though it is a common occurrence for immigrant groups to shift to the societally dominant language within three generations (Winford, 2003: 237), in the Turkish community in the Netherlands Turkish is still vital and widely used, even among the third generation speakers. Language maintenance rates are very high, the Turkish language is highly valued and considered to be an essential part of one’s Turkish identity (Broeder and Extra, 1999; Extra et al., 2001; Boumans et al., 2001; Dorleijn and Nortier, 2006; Yağmur, 2009).

This can for a great part be attributed to the ardent Turkish-only language policy that has been pursued by the Turkish state ever since the advent of the Turkish Republic in 1923. As a consequence of this policy the overwhelming majority of Turkey’s inhabitants, as well as members of Turkish communities abroad, are familiar with standard Turkish. A standardized near-phonemic spelling in Latin script has been in existence since 1928. Standard Turkish is the only official language in Turkey and functions as the language of education and the media. Members of communities abroad have ample access to standard Turkish through all kinds of media, and it is common practice for Turks living abroad to send their children to Turkey in order to learn ‘proper Turkish’ and for them to complete their education there. Another factor that contributes to the resilience of Turkish in the Netherlands is, that up to 2012 the tendency among the Turkish community was to marry a partner from Turkey.3 As a consequence of this the grandchildren of the initial immigrants now grow up bilingually. At present, members of the Turkish community indicate that code switching is the most effortless way of communicating, although this mixed code is held in low esteem.

3 This situation may change however in the near future, because of the stricter immigration policy of the Dutch government since 2012.
The description of new varieties that are the result of language contact is a central issue in the field of language contact studies. There is debate as to what extent, if at all, stable varieties such as mixed languages may arise out of code switching (e.g. Bakker, 2003: 129), however, an instance of exactly such a process has been documented (e.g. McConvell, 2008; Meakins, 2011). McConvell attributes the emergence of the Australian mixed language Gurindji Kriol to the fact that code switching was pervasive among adult Gurindji speakers, such that this provided the main input in childrens’ first language acquisition process (McConvell, 2008:187; McConvell and Meakins, 2005.) O’Shannessy (2012, 2013) reports a similar process for Light Warlpiri. Meakins (2013) proposes the following criteria for the definition of a certain variety as a mixed language: the extent to which the variety can be considered autonomous and stable. Stable in the sense that diverse speakers of the mixed variety are consistent and uniform in their syntactic and lexical choices.

Among the Turkish community in the Netherlands the input for first language acquisition is probably often Dutch-Turkish mixed speech as well (Boeschoten, 1992; Backus and van der Heijden, 2002). But the input may vary much more than is the case with Gurindji Kriol or Light Warlpiri, because the sociolinguistic situation of the Turkish community is more complex, dynamic and varied and thus less stable, with varying linguistic input in the course of language acquisition as a consequence.

It is therefore unlikely that a mixed Turkish-Dutch language is in the making. However, given the fact that code switching is perceived of as the most effortless speech variety in many in-group encounters, despite its putative randomness, it is realistic to assume that patterns that make code switching predictable can be found.

To date language contact effects between Turkish and Dutch have been described extensively in studies such as: Boeschoten, H. and Verhoeven, L., 1987; Backus, 1996, 2001, 2009; Backus and Dorleijn, 2009; Backus et al. 2013; Boeschoten, 1997; Doğruöz and Backus 2007, 2009; Dorleijn, 2002; Dorleijn, Boumans and El-Aissati, 2005; Eversteijn, 2011).

However, with the exception of Backus (1996) and Eversteijn (2011), focus in these studies is mainly on contact effects in monolingual Turkish, rather than on patterning in mixed language. (Backus et al., 2013 will be discussed in more detail in 5.)

2.1 Code Switching, Language Mixing, Fused Lects
Several types of code switching are described in the literature. As this typology is well known and extensively discussed in numerous scientific contributions,
there is no need to go into very much detail here. But at this point, the definition of code switching relevant for this paper should be refined.

Relevant for this paper is the type of code switching that Auer (1999) has labelled 'Language Mixing', which is similar to Poplack's 'Intimate switching' (Poplack, 1980), Myers-Scotton's 'switching as an unmarked choice' (Myers-Scotton, 1993b) and Gafaranga's 'bilingual medium' (Gafaranga, 2007). It is the type of frequent code switching where, contrary to other types of code switching (such as 'language alternation' (Auer, 1988), emblematic switching (Poplack, 1980), and so on) each individual switch does not seem to be referentially, socio-psychologically or conversationally motivated; rather it is the whole act of switching which is meaningful and should be considered as a separate (bilingual) code. A (Turkish-Dutch) example of this type of switching is (1):5

(1)  *Bizim bir arkadas vardi,* was een jongen en hij was kei ijdel *sonra kendine acayip iyi bakiyordu,* elke ochtend ontbijt etc... *neyse* hij werd griep en ik hem aan de telefoon... *Seytanlık işte*... ik zei hey *duydunmu gece senin gibi biri hastalandı,* kreeg eerst rode vlekken op zijn gezicht (hij had dat ook) *sonra sesi gitti aldirmadi* en nou *rahmetli oldu.* hij was echt een toffe gozer felan.. *Neyse 2 saat sonra had ik hem weer gebelt cunku acayip korkmustu* aan de telefoon. *aradigimda* was hij bij de dokter.. *acayip korkmustu* hahahah

4 See for example for an overview of structurally oriented models of Code Switching: Myers-Scotton (1983) for a model of insertional code switching; Poplack (1980) for a proposal of constraints for alternational switching; Johanson (1999) for an integrated model of code switching in relation to a diachronic account of contact-induced change). Or see for an overview of these three models: Backus (1996). For a seminal and brilliant account of communicative motivations for code switching see Gumperz (1982), for a conversationalist approach Auer (1984); see for seminal papers on various aspects of code switching the volumes edited by Milroy and Muysken (1995) and Li Wei (2000); see for an accessible overview of the diverse disciplines that study code switching: Gardner-Chloros (2009). For a critical overview of the most prominent theories and models of code switching see Gafaranga (2007). See also the other contributions in this volume.


6 Examples throughout this paper are rendered as found on the web, including typos and missing diacritics, and according to the orthographical conventions used by the original source writers.

*Italics* represents Turkish, plain type represents Dutch (and occasionally English). This suggests a dichotomy between only standard Turkish and standard Dutch, but the complex reality is abstracted away from in this example (as well as in the others in this paper). In fact it constitutes an amalgam of colloquial Turkish, colloquial Dutch, regional Dutch, standard
There was this guy, this guy, he was a terribly vain creature and he well was incredibly fussy about his health every morning proper breakfast etc.... anyways he caught a flu and he phoned me... Something devilish [came over me]... I said hey did you hear someone got sick just like you, first he got those red spots in his face (he got them, too) then he got this hoarse voice but didn't pay attention to it and now he is dead. He was a nice guy and so on.. Well two hours later I phoned him again because he seemed terrified [when we talked] at the telephone. When I called he was at the gp... he was completely terrified ha, ha, ha.

From a structural point of view, the code switching in (1) is predominantly of the alternational type (the type which, according to Poplack and later researchers, can only be performed by highly skilled balanced bilinguals), but in this type of ‘intimate switching’ insertional code switching may occur as well (Auer, 1999). This type of switching will be referred to as the ‘bilingual code’ in the remainder of this paper.

2.2 Predictability

It is this bilingual code that is often referred to as the most effortless one by speakers, as illustrated by the quotations in the introduction. No systematic studies on the evaluation of – and the attitudes to – the bilingual code have been carried out in the Dutch-Turkish community. But impressionistic evidence suggests that for members of this community, too, the bilingual code is the most effortless way of speaking. The majority of Turkish-Dutch bilinguals, whenever asked which is their most relaxed, their most ‘effortless’ modus of speech, usually answered ‘mixed’. Pupils with a Turkish background were questioned about this by the author during a one year period of fieldwork on a secondary school in Utrecht. (See also Grosjean, 1982; Moreno et al. 2002; Gafaranga, 2007; Nortier and Dorleijn, 2008; Eversteijn, 2011). Also metalinguistic comments on the Internet are suggestive (see Dorleijn and Nortier, 2009).

This would suggest that, just like the other types of code switching, also the bilingual code is not random, but rather, to quote Maschler (1998):

[a bilingual code is] using two languages such that a third, new code emerges, in which elements from the two languages are incorporated into a structurally definable pattern.

MASCHLER, 1998:225

Turkish and standard Dutch, etc., with clear traces of synergy between the involved language varieties (and that is still an oversimplification).
In this paper it is assumed that the ‘third code’ referred to in this quote has two characteristics. The first characteristic would be a task division between the contributing languages (comparable to what is the case in mixed languages). It is unlikely that two complete linguistic systems would remain intact in a mixed code, since this would render the bilingual code highly complex. Maschler (1998; 2000) establishes that at an early stage of language contact such a task division can indeed be observed in the use of discourse markers (see below for more discussion). The second characteristic would be that innovative structures occur, since the intimate switching of two linguistic systems would imply a certain synergy between them. The question is, then, to what extent such innovative structure are mere incidental and ephemeral or have already been stabilised and conventionalised.

This paper seeks to investigate whether preferred usage patterns have developed, despite the relatively brief period (from the point of view of historical contact linguistics, that is) in which Turkish and Dutch have been in contact. Such patterns may appear on any (socio)linguistic level: the macro level (code choice); the pragmatic, lexical, structural, phonological level, etc. In what follows the focus will be on properties of, and the patterns within the bilingual code, but at the same time code choice on a macro level (were the choice of a particular linguistic variety reflects the linguistic constellation in society) in general will be examined as well, for two reasons. Firstly, to establish the function and domain of the bilingual code. Secondly, because code choice is also relevant on the level of the selection of individual lexical items. The lexicon is an obvious level for labour division to occur, because switched lexical items tend to be domain-specific, and, therefore, language specific (see e.g. Backus, 1996). As such, language choice has direct repercussions on the lexicon.

Thus, the central question on which this paper seeks to shed light is:

> do preference patterns and innovative structures occur in Turkish-Dutch bilingual internet data?

The hypothesis is that a. preference patterns will occur in internet data: on the macro level code choice will be consistent and predictable; on the micro level preferred items will occur significantly more frequent than its (Dutch or Turkish) alternatives, and b. innovative linguistic features will occur in internet data.

In section 3 the central theme of this special issue is addressed: to which extent is internet data suitable for answering the question above.
3 Internet Data

Internet data have obvious advantages and disadvantages when compared to naturalistic data, a number of which have been discussed in the general introduction to this volume. The similarities and dissimilarities of internet data vs. naturalistic data have been addressed there as well. As observed in the introduction to this volume, many studies of internet data have highlighted the commonalities between digital and naturalistic data. Here I want to emphasize that the fact that internet data differs from spoken data may have its benefits for structural linguistic, as well as sociolinguistic research.

3.1 Advantages

Although the corpus used for this study has features in common with speech, such as the occurrence of colloquial, low prestige language (see section 4), it is still a different mode that may be assumed to require a higher level of awareness than speaking. It is arguably an advantage that internet data differ from oral data in this respect, since writing in a context that is created on the spot, with no extra-linguistic information available, may induce communicators to use forms that they consider comprehensible to the unknown addressees, that is to say, forms that are not too idiosyncratic, and may even have been conventionalised.

Moreover, the choice of code will be less fuzzy in computer mediated communication. In actual face-to-face interaction speech is dialogic and emergent and is enacted in a dynamic, continuously shifting situation which will induce the interlocutors to reconsider and negotiate their linguistic choices continuously. Naturalistic data are therefore variable and fuzzy, whereas on the internet, interaction is more controlled and less likely to be influenced by subtle shifts in the situational context.

Thus, informal written speech, stripped of all extra-linguistic information, may in crucial ways not be similar to speech, but may because of its more controlled nature be an ideal locus to look for patterns of language choice on the macro level and language contact features on the micro level.

Another advantage of internet data (which was also discussed in the general introduction) is the fact that the analyst can select specific ‘speech’-events, such as: political discussions, fiction, poetry, informal chit-chat etcetera. Also the fact that synchronous fora are organised in topics is helpful to get a clear picture of topic-relatedness of certain linguistic choices.

3.2 Disadvantages

In the general introduction, two disadvantages concerning internet data are discussed which will briefly be touched upon here. Firstly, the communication
takes place without face to face contact, without extra-linguistic cues such as the characteristics of the interlocutors, the physical context, and the language usage practices (gestures, paralinguistic features, pragmatic usage patterns) connected with face-to-face contact. The role the latter features play in language production, and the extent to which they are in fact part of the linguistic system, have not been completely clarified yet. The absence of these features, then, may distort the picture of what naturalistic language use is like.

This disadvantage can partly be compensated for, however: avatars, nicknames and clues in the in situ-context can tell us many things about the users (see also Androutsopoulos, 2006, for a similar view). For example, if someone’s nickname is KaramanlıKız92 (’Girl from Karaman 92’), one can get important information about the person in question: a girl from the Turkish province or city of Karaman, born in 1992. We can never be sure that these details correspond to those of the ‘actual’ person, but, as Benwell and Stokoe (2006: 245, emphases in the original) argue:

Due to anonymity, freedoms of time and space, and absence of audiovisual context in cyberspace, identity is deemed to be more unstable, more performed, more fluid (and thus prone to inauthenticity and deception). It is a view of cyber identity that is remarkably consonant with the dominant view of identity as postmodern, constructed and discursive, a view that is not confined to cyberspace. In other words: in constructionist accounts, all identities are ‘virtual’: an ongoing production of an imagined, but ultimately intangible ‘real’ identity.

The absence of face-to-face contact can, if one looks for conventional formal linguistic items, even be considered an advantage: often you can get an idea about which social characteristics the actors want to convey to their audience that they have, using linguistic items that index such an identity (and may be conventionalized later on).

The second obvious disadvantage (which is related to the first one) is the loss of subtle phonetic and phonological detail in writing. Sociophonetic studies such as those reported on in Hay and Drager (2007) inform us that fine phonetic detail may give important social information, be part and parcel of specific codes and thus must in many instances be considered conventionalised. This kind of information can hardly be conveyed in writing. The same applies to possibly rather salient prosodic features such as intonation contours that are part of a particular code (see e.g., Queen, 2001). These subtle

---

7 I did not use data of which the pertaining authors have explicitly stated that they did not want me to use them. Nevertheless, KaramanlıKız92 is a pseudonym.
signals can only partly be replaced by using medium-specific means such as emoticons. Information of this kind, just like other cues given in face-to-face contact, might be crucial in the propagation and conventionalisation of certain items, and if that is the case, internet data will not reveal that.

On the other hand, some relevant information can be reflected in writing (without, of course, complete phonetic detail). For example, phonetic reduction, a feature commonly observed in conventionalised items (see e.g. Bybee, 2010: 44) might very well be reflected in writing and is, when found in internet data, in fact a diagnostic that suggests that we are dealing with a colloquial, rather than a ‘standard’ form. And, at the opposite end of reduction one can find rather salient features that may have become part of a particular code, such as the geminated /z/ in certain Dutch-based multi-ethnolects (see Hinskens, 2011, for a description of this feature). Salient features of this kind are often reflected in informal writing. Such salient features, too, can be part of a distinct variety that has undergone (ephemeral as it may be) a conventionalisation process (Nortier and Dorleijn, 2008). Also choice of script can be highly informative, as can be seen in Verschik (this volume).

To summarize this section: internet data are viewed as language use where the physical context plays no role. The context is created in situ, mainly by written verbal means, and therefore controlled and not susceptible to subtle shifts in the situational context that would possibly blur the picture (for the analyst, not for the speakers!). Representation of paralinguistic or subtle phonetic information is limited. The data lack to a large extent the dialogic nature of face-to-face communication as well. What written internet data offers us is the bare linguistic material: lexicon, word order, morphology, phonology.

The lack of extra- and paralinguistic information, it can be assumed, will stimulate authors to use linguistic means that are unambiguous and can be understood by an audience that is wider than the direct network. The control over the context and the absence of face-to-face dialogue will lead to ‘typical’ choices that are free from the fuzziness of linguistic choices in the off-line world with its continuously shifting contexts. Thus it can be assumed that internet data reflect, rather than create, conventionalized items. It may in this respect be more ‘conventional’ than speech.

4 The Data and the Methods

This section and section 5 report on an explorative study on internet data of Dutch-Turkish language internet fora. It was carried out with the aim to find out if preferred usage patterns could be observed at all. Furthermore, the aim was to provide a description of some aspects of these preferred usage patterns.
The data-set used consisted of 219,536 words. The data set was kept small-scale intentionally: if preferred usage patterns can be detected in a small-scale data set in a noteworthy way, it can be taken as an indication that such patterns surface in digital data earlier than in spoken data.

As will be discussed in section 5, even on this relatively small scale, patterns emerged on several linguistic and sociolinguistic levels.

4.1 The Data-Set

Data were collected from bilingual Turkish-Dutch internet fora. There are diverse fora catering for the Dutch-Turkish community. A detailed comparative analysis of these websites will have to be carried out as yet, but the overall impression is that all websites cater for younger users (aged between 13 and 25) and cover a comparable range of topics.\textsuperscript{8}

Typical topics on these fora include ‘news of the day’, ‘relationships and sex’, topics about what is allowed or not in Islam, and topics called ‘chit-chat’, ‘humor’, ‘jokes’ etc. Usually one member or a moderator starts a topic and others respond or react to it. Although in principle a-synchronous, sometimes (especially in the chit-chat-like topics) the contributions resemble those of face-to-face interaction, because the posts consist of brief consecutive exchanges. There are several reasonably active bilingual Dutch-Turkish internet fora available. Of three of them material was included in the test corpus, namely: www.hababam.nl, www.turksestudent.nl and www.Turkishplace.nl. These sites were accessed in May 2011 and February 2013.

Hababam was launched in 2005 and has on average 34,026 visits each month (ca. 1,300 every day). Turkishplace was launched in 2008 and has around the 193,989 registered members as of February 2013 and an average of 300 unique visitors each day. Turksestudent was launched in 2003 and is smaller, with an average of 11,500 visits a month. Hababam as well as Turkishplace seem to be frequented mainly by adolescents, Turksestudent is, as its name suggests, catering for university students and students of higher vocational training.

The data set consists of two subsets:

1. Posts from four discussion topics on Hababam, which were either in monolingual Dutch, monolingual Turkish or in the bilingual code. These were each analysed individually, both from the macro perspective of code choice, as well as for micro linguistic description. For topics and the number of contributors, see Table 1. The total body of bare text, that is,

\textsuperscript{8} As of the time of revising this article (2015) bilingual fora seem to be on the decline, their function being taken over by social media.
text stripped of meta-text, nicknames etc. consists of 17,000 words, which means that the average length of the individual posts is 13.6 words. In fact many of the posts consisted of one-word-answers. The topics selected were:

1. Güzel Sözler maybe te gebruiken op MSN (‘Beautiful words that can be used on MSN’)
2. Domst dingen die je ooit gehoord hebt (‘The most stupid things you’ve ever heard’)
3. Wat zit je te eten? (‘What are you eating this moment?’)
4. Eindexamenkandidaten 2011 (‘Final Exam Candidates 2011’).

The topics were selected based on two criteria:

a. The large number of views (>20,000) and posts (>200).

b. The expectation that each of the selected topics would trigger specific language choices. This expectation was based on field observations among the Turkish community, and on the vast literature on language choice, most specifically Eversteijn (2011), who focuses on the code choice of young Turkish people in the Netherlands.

2. The second subset consists of bilingual Turkish-Dutch storytelling. Most of the stories were found on Turkishplace.nl. These data were provided by in total 40 unique contributors. The stories were selected for two reasons:

---

9 Total number of contributors is between ≥239 and ≤ 282 unique contributors, due to overlap: i.e. contributors who contribute to more than one topic. A contributor is defined as ‘unique’ if s/he is using a unique nickname and avatar (with the obvious drawback that a contributor using two different nicknames will be counted as two unique contributors).
they featured quite dense bilingual discourse.

b. they were longer than the average contributions on fora, providing the opportunity to consider code switching in larger contexts. The total number of words of this subset is 202,536. This data set consists of one genre only: stories. Narratives are, of course, monologues rather than dyadic, but they can contain dialogue that is similar in style to conversation (cf. Paul, 2007). This was also the case in the samples of this data set, which made them suitable for the analysis of bilingual speech.

4.2  Methods
The data were approached in an exploratory way.

The first data set was analysed in detail for language choice as well as for structural and discursive aspects of code switching. The second data set was in first instance searched quantitatively by extracting word frequency lists and concordances using the concordance software MonoConc. Afterwards, the extracted data were examined in detail: each individual instance of a found pattern was examined. Because the data set was limited in size, it was possible to view each individual instance in its context.

5  Results
In the following sections results will be discussed. Patterns of code choice on the macro level will be discussed in 5.1., code choice on the lexical level including discourse particles, in 5.2., structural and discursive patterns of code switching in 5.3., and innovative structural patterns in general in 5.4.

5.1  Relation Topic-Language Choice (Macro-Level of Code Choice)
The topic Güzel sözler ‘Beautiful words’ was an ideal locus to study the relationship between language attitude and language choice. The topic was opened with the aim that contributors suggest nice slogans, catch words, song lines etc. that others could use along with their nicknames. Interestingly, the topic opener gives the topic the title Güzel Sözler maybe te gebruiken op MSN ‘Beautiful words that can be used on MSN’, which is trilingual Turkish, English and Dutch. She gets reprimanded by other contributors immediately (as often happens when the bilingual code is used): how can words be ‘beautiful’ in the bilingual code? The pattern that emerges from the posts in this topic can be summarized

as follows: when words need to be beautiful: they have to be in standard Turkish, or, to a far lesser extent in standard Dutch, very rarely in English but not in mixed code or colloquial forms of Dutch or Turkish. (2) is a typical example of a post of ‘beautiful words’, in Standard Turkish:

(2) güzel sozler kalbinden gelen sozlerdir
‘Beautiful words are words that come from the heart’

By contrast, off-topic\textsuperscript{11} side-comments were mainly in the bilingual code (as in (10) below) or in Colloquial Dutch (sometimes in standard Turkish or standard Dutch, but to a far lesser extent). See table 2 below.

(3) ...soz degilde ne güzel sozler yaziyo en tussen haakjes maybe ook te gebruiken op msn ben uzerinede yazmistim is niet voor msn bedoeld diyı amı paylasmak istedim paylasanda hata zaten!
‘well, these are words too, aren’t they, and [the topic] is [about] beautiful words and, besides, maybe to be used on msn and I had written that [my contribution] was not meant for msn I just wanted to share [these beautiful words] what is wrong with that!

Other observations concerning code choice that were made in the corpus equally suggest that the bilingual code is used in informal situations where the content of the story is central and where the contributor wants to get it across as quickly and efficient as possible, which is in accordance with what is observed in the ‘off-line world’. (1) and (3) above are examples.

5.2 Language Choice on the Lexical Level (Results from the First Data Set)
The preferred usage patterns this paper seeks to establish, may be seen as precursors of what Auer (1999) calls a Fused Lect. Auer defines a Fused Lect as follows:

(in a Fused Lect) neither type of language contact is locally meaningful. On the surface, a F(used) L(ect)may look similar to L(anguage) M(ixing). Often, the difference becomes visible at a deeper grammatical level only.

\textsuperscript{11} In the context of mailing lists, discussion groups, discussion forums, bulletin boards, newsgroups on internet a contribution is off-topic if it is not within the bounds of the current discussion, and on-topic if it is. Often contributors indicate whether their contribution is to be considered off-topic (and again on-topic) by using the abbreviation OT.
but they need not be), the use of one “language” or the other for certain constituents is obligatory in FLS; it is part of their grammar, and speakers have no choice. Thus, structural sedimentation (...) of ML into a FL presupposes a loss of variation and the stabilization of function-form relationships. *Comparing the FL grammar with that of the two languages involved, this means simplification, since alternatives are lost.* 

_Auer, 1999: 321- my emphasis_

Could this sedimentation, this simplification entail a decrease in lexical options to choose from as well, resulting in a vocabulary to which the two languages contribute in a complementary way?

The topic *Wat zit je te eten* ‘What are you eating this moment?’ was a question that mostly provoked minor responses, in which contributors simply stated with a single word what they were eating. This provided an excellent opportunity to look for lexical preference patterns. Inspection of 400 minor responses revealed patterns that were as expected: Turkish food items were in Turkish (102 items), such as *cevizli sucuk* ‘walnut sausage’, a kind of candy, or *cig kofte* ‘raw meatballs’, and Dutch food items were in Dutch (195 items), such as *beschuit* ‘Dutch rusk’ and *drop* ‘licorice’.

---

12 This neat partition in different codes should of course be viewed as an oversimplification of the gradual reality. Discussion of features typical for each code is beyond the scope of this paper.
It is conceivable that the bilingual code contains items from domain-specific vocabulary, in which either synonyms in one of the languages are lost or each synonym acquires a specialized meaning. The following quote is illustrative:13

\[(4) \quad \text{mn kanki laatst} \quad \text{illaayasee een etek zit my best friend, recently: I swear yasee (=proper noun?) a skirt is serieux fijner dan een rok really more comfortable then a skirt} \]

‘My best friend, recently: I swear, yasee, a skirt (Turkish) is really more comfortable then a skirt (Dutch).’

Apart from these rather obvious preferences, another preference emerged: the answer ‘nothing’ to the question ‘what are you eating this moment?’ It was given 13 times in Dutch, and never in Turkish. This is not a significant number, but nevertheless, it is conceivable that considerations of economy play a role here (such as discussed in, for example, Croft, 2000) After all, \textit{niks} and its variants \textit{niets} or \textit{nix} ‘nothing’, are much briefer than their Turkish counterpart: \textit{hiç bir şey} ‘not a thing’; ‘nothing’. This could also be an effect of the maxim that obtains specifically for computer mediated communication: that is to ‘be as fast as possible’(see DeDecker and VandeKerckhove, this volume), rather than with language use in the off-line world. Nevertheless, additional consultation of the larger second data set reveals 346 instances of Dutch \textit{niets} ‘nothing’ and its variants, and 146 instances of its Turkish equivalent \textit{hiç bir şey} ‘nothing’. It is also noteworthy that, again in the second data set, the context in which \textit{niets} was used were monosyllabic answers, like the ones discussed above, in mixed Turkish-Dutch stretches, whereas \textit{hiç bir şey} was mainly used in monolingual14 Turkish stretches.

Furthermore, a few instances of patterning that are less straightforward should be mentioned. For example, it is difficult to find a rationale in the fact that the term ‘chocolate’ occurred 18 times in Dutch (\textit{chocola} or \textit{chocolade}), and one time in Turkish (\textit{çikolata}).

Finally, mixed food items were mentioned (28 posts in total), all of which had Dutch as their base language or Matrix Language (as defined by

---

14 The term ‘monolingual’ is an oversimplification, of course: it should be understood as: ‘mono-phonological’ – e.i., the surface forms are Turkish, but this says nothing about the degree of ‘global’ or ‘selective’ copying. (See Verschik, this volume, for the terminology.)
Myers-Scotton, 1993a). Thus, in mixed expressions, all functional elements were in Dutch (and all lexical words in Turkish or Turkish related) – and not the other way round: no instances where Dutch food items were embedded in clauses with a Turkish matrix language were found. An example of a mixed item is (5):

(5)  
findikli cafe crown\textsuperscript{15} 3-u 1 arada met karisik cerez
with.hazelnuts cafe crown 3-of.it 1 time.in with mixed nibbles
‘cafe crown with hazelnuts, three at a time with mixed nibbles’

5.2.1 Language Choice at the Level of Utterance Modifiers. (Results from the Second Data Set)

The behaviour of utterance modifiers (Matras, 2009), such as discourse particles, interjections, conjunctions, tags, fillers, etc. have been the object of a number of code-switching studies (See Poplack, 1980; Maschler, 2000; Matras, 2000; Hlavac, 2006; Matras and Sakel, 2007 and many more). The behaviour of utterance modifiers is remarkable in that, in otherwise monolingual stretches, it is often the only item that is switched. They are (therefore) less easy to fit into the prevailing structural models of code switching.

In the literature, the positions on the production of code switched utterance modifiers are diverse. At the one end of the continuum is the concept of ‘emblematic’ switching (Poplack, 1980) where it is assumed that the production of utterance modifiers implies a high level of intentionality and agentivity. On the other end, one can find Matras’ (2009) position, who argues that the process of managing the discourse and monitoring and directing the interlocutor is demanding for the speaker. Matras argues that utterance modifiers are typically designed for this process and are typically produced at moments where the speaker is least concerned with and aware of the form of his speech. According to Matras, the bilingual speaker reduces the extra stress of having to select between competing word forms from two languages, and opts for fixed sets of utterance modifiers where for each utterance modifier only one form is used. In Maschler (1994) it is proposed that bilingual discourse markers have a metalinguistic, contrasting function.

What is important for this paper, is the fact that Maschler (1998) claims that the task division between discourse markers indicates the first stage of a ‘Fused Lect’ (but see Maschler (2000) for a modification of this claim). Hlavac (2006) offers an insight into the processes and mechanisms that are at work in the course of task division between discourse markers of two languages.

\textsuperscript{15}Cafe Crown is a Turkish cookie-brand.
Given this brief outline of utterance modifiers above, and more specifically Maschler’s claim that the task division between discourse markers is the precursor of an emerging Fused Lect, a closer examination of the use of discourse markers in the data set seems to be justified. Moreover, the sheer frequency of discourse markers, in contrast to content words, allows quantitative analysis within a limited data corpus. A feature of the digital data sample is that, contrary to most genres of written text, discourse markers do occur frequently in internet data due to its attribute of being ‘text-transposed speech’. The second subset of data consists of bilingual fiction with frequent dialogues in reported direct speech, which contain frequent instances of the use of discourse markers.

Hlavac (2006) observes that in the speech of second-generation Croatian-English bilinguals in Australia, English discourse markers co-occur with their Croatian equivalents. He also reports that high frequency English discourse markers that are polyfunctional, where their Croatian counterparts are not, tend to ‘take over’. One example that Hlavac discusses is the interjection ‘yeah’, more or less equivalent with the Dutch particle ja. Backus et al. (2013) observe the frequent insertion of ja in the monolingual Turkish of third-generation members of the Dutch-Turkish community that were forced into a monolingual Turkish mood by having to communicate with a monolingual Turkish research assistant.

The Dutch particle ja (‘yeah’, ‘yes’) has a number of discursive functions that are more or less identical with those of its English equivalent. Description of these functions is beyond the scope of this paper. Here we just want to see how often it occurs as opposed to its Turkish counterpart evet (‘yes’; ‘indeed’). In the second data set ja occurs 738 times, both in monolingual Dutch and monolingual Turkish as well as in bilingual fragments. By contrast, the Turkish word for yes, evet occurs 105 times, but exclusively in monolingual Turkish fragments. This, combined with the observations in Backus et al. 2013, is suggestive of the fact that ja is ‘taking over’. However, we must take into account that evet cannot be considered a complete equivalent of ja; for example, contrary to ja in Dutch, evet in standard Turkish is seldom employed for monosyllabic affirmative answers. In Standard Turkish, a yes/no question is generally answered (affirmatively or negatively) by repeating the main verb of the question with the proper person/number marking, as illustrated in:

(6) A: ‘Ahmet’le görüş-tü-n mü?’
Ahmet-with speak-PST-2SG Q
‘Have you spoken to Ahmet?’

---
Whether the Turkish answering strategy of repetition of the full form of the verb is entirely abandoned in the bilingual code will have to be left for future research.

5.3 Structural and Discursive Patterns of Code Switching
For the usage patterns in code switching in the bilingual code, the topic Domste dingen die je ooit gehoord hebt ‘The most stupid things you’ve ever heard’ was examined in detail. In this topic, which is part of the first data subset, contributors posted stories on (verbal) blunders of others or of themselves. This topic appeared to be a very rich data source. Laughing about the stupid things people say and do apparently calls for the bilingual code.

The 300 posts of this topic were examined in detail. Instances of code switching as well as monolingual fragments were counted and categorized. It appeared to be relevant to distinguish between brief posts that can be considered as quick reactions or exchanges, and longer posts, the actual ‘stupid things’ that were recounted as narratives. Code switching in single sentence posts appeared to show structural patterning, to be discussed in 5.3.1. In narratives another mechanism appeared to be at work as well. As already noted above, narratives are monologues rather than dyads, but they can contain dialogue that is similar in style to conversation (Paul, 2007). In these dialogues, the direct speech was often in a contrasting language mode (as was already described by Gumperz, 1982), but this mode appeared to be determined by the language that the quoted person would typically speak.

5.3.1 Single-Sentence Posts
For single-sentence posts the following categories appeared to be relevant in the sense that they showed an obvious pattern:

1. Alternational switch from Turkish to Dutch.
2. Alternational switch from Dutch to Turkish.
3. Monolingual Turkish.
4. Monolingual Dutch.

The number of instances of each category is given in Table 3 below.

The significance of the frequency of Turkish/Dutch vs. Dutch/Turkish cannot be established on the basis of these figures (but note that Turkish/Dutch switching occurs almost twice as much as Dutch/Turkish, and consider
the fact that the switches were produced by numerous contributors), but the difference in structure between the two is striking: all instances of Turkish/Dutch are either complex or compound sentences of which the second half (the dependent clause or the second independent clause) is Dutch. Typical examples are (1) and (3) above, as well as the following:

(7) küçük neefje- m-i animsat- ti bu, die ipv geit-je
little neph ew-poss.1sg-acc reminds3-sg this, who instead.of goat-dim
geit-je bleef zeggen
shit-dim continued.3sg say-inf
'It reminds me of my little nephew who kept saying 'little shit' instead of 'little goat'.

(8) bakalim of je nog grote mond zult hebben
we'll see whether you then big mouth will have
'We'll see whether you will have such a big mouth then'.

On the other hand Dutch/Turkish switching occurred exclusively between two independent clauses, such as in the following example:

(9) Yaw even serieuw wat is het?? Ik beheers de straat taal niet vloeiend of zo dus ik zou bij ALLAH niet weten wat het betekent. Bil memek de gił œgrenmemek ayıiptir.
'Come on, be serious now, what does it mean?? I don't speak street slang fluently so by GOD I wouldn't know what it means. Not to know is not a crime, not wanting to know is.'

The non-occurrence of Dutch/Turkish complex sentence switches can with fair confidence be attributed to the fact that the Turkish and Dutch systems of clause embedding are completely incongruent (in Myers-Scotton’s (1993a) sense of the term): in Turkish, embedded clauses always precede the main clause in default word order, whereas in Dutch it is the other way round for
most types (and variable for adverbial clauses). That Turkish/Dutch complex sentence-switches by contrast do occur frequently may possibly be attributed to the same fact: when starting a complex sentence in Turkish, at the clause boundary one finds out that finishing this sentence in Turkish is not possible. The psycholinguistic implications of this are fascinating but must be left for future research. Backus et al. (2013) and Onar Valk (2014) describe a convergence toward Dutch word order in Turkish embedded sentences in the monolingual Turkish of third-generation members of the Dutch-Turkish community, where the latter were forced into a monolingual mood by having to communicate with a monolingual Turkish research assistant. Onar Valk observes a syntactic change in Dutch Turkish subordination from non-finite to finite subordinate clauses in Turkish, which is well in accordance with the results described above.

5.3.2 In Narratives
In the posts with a narrative character, the structural aspects of code switching are less salient than switches that are related to the management of the narrative structure. The relevant categories were as follows:

5. Turkish story frame, direct speech in Dutch
6. Dutch story frame, direct speech in Turkish
7. Monolingual Turkish story frame and direct speech
8. Monolingual Dutch story frame and direct speech

An example of 6 is the following:

(10) Mijn broertje ging vandaag voor het eerst naar de basisschool. Komt ie enthousiast thuis; abla benim jufumun\textsuperscript{17} adi ne bilyormusun, meester Lars.
‘My little brother went to school for the first time today. He came home full of enthousiasm; sister do you know how my female.teacher is called? Master Lars.’

The number of instances of each category is given in Table 4.

As can been seen in table 4, in bilingual narratives the preference is for a Dutch frame and Turkish direct speech. By contrast, a Turkish story frame had Dutch

\textsuperscript{17} Note that ‘neefje-m-i’ in (7) and ‘juf-um-u’ in (10) are insertional code switches.
reported direct speech in only three instances. A code switch to signal direct speech is very common, as has been established frequently ever since Gumperz’s seminal work (1982). But it should be remarked here that only persons that can be considered capable of speaking Turkish (such as family members) are quoted in Turkish. Monolingual Turkish narratives are strikingly less frequent than either of the other categories. Monolingual Dutch was encountered most frequently. This points at the process of language shift going on.

5.4 Innovative Features

In studies on language contact the occurrence of innovative structures and specialised functions have been attested frequently (e.g. Winford, 2003; Thomason, 2001; Matras, 2009 for overviews and references). Auer sees the occurrence of innovative structures as part and parcel of emerging fused lects:

(... structures from language A and B which are more or less equivalent in monolingual use may develop specialized uses in the fused lect AB. Also, fused lects may have to adapt structurally to the massive combination of elements from A and B by developing new structures identical neither to those of A nor B.

_Auer, 1999: 231_

In this section, some features that can neither be attributed to influence of Turkish on Dutch, nor of Dutch on Turkish, but are innovative, will be discussed. ‘Unconventional’ features in the Turkish of Dutch-Turkish bilinguals are well described. (Backus and Doğruöz, 2007; 2009; Backus and Dorleijn, 2009; Backus _et al._ 2013; Dorleijn, 2001; Onar Valk, 2014). Most of these unconventional features can directly be attributed to contact with Dutch, though some are to be attributed to other mechanisms that are at play in language contact, such as the notion of transparency (Dogruöz and Backus, 2009). Backus and associates emphasize the fact that

<table>
<thead>
<tr>
<th>category</th>
<th>f: Turkish</th>
<th>f: Dutch :</th>
<th>Monoling. Dutch</th>
<th>Monoling. Turkish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ds: Dutch</td>
<td>ds: Turkish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instances</td>
<td>3</td>
<td>42</td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

Legend: f = story frame. ds = direct speech
(...)it must be kept in mind that the spoken data in this paper do not reflect NL-Turkish faithfully, because code switching to Dutch was not an option (because the respondents had to speak with a monolingual Turkish research assistant – MD) without doubt a frequent phenomenon in NL-Turkish. Some of the unconventional constructions that we detected are probably not established conventions in NL-Turkish, but were created on the spot because outright codeswitching was not possible.

DOĞRÜÖZ and BACKUS, 2009: 61

In the bilingual data set in the present study, innovative aspects as described in the studies of Backus and associates mentioned above were established as well, which is suggestive of the fact that the phenomena discussed in the monolingual Turkish data in these studies occur in code switched data as well. For reasons of space, innovative features in the Turkish part of the bilingual data will not be addressed here. Here, the focus is on the Dutch elements in code switched stretches. The findings presented below are based on an analysis of the second data set, which consists of bilingual story writing (containing a lot of direct speech reporting).

5.4.1 Rearrangement of Semantic Patterning of the (Auxiliary) Verb: Gaan (‘go’) The verb *gaan* ‘go’ was remarkably frequent in the second data set: when all separable verbs (e.g. *uitgaan* ‘to go out’, combinations with fixed adpositions (e.g. *gaan naar* ‘to go to’, and frequent expressions (e.g. *gaan zitten* ‘to sit down’) are subtracted, *gaan* and its inflected forms occur 2011 times, constituting almost 10% of the total number of tokens.

In standard Dutch, the verb *gaan* (‘go’) has a variety of meanings and functions. Apart from its basic function as an allative verb of movement (in combination with the adposition *naar* – ‘to’) it is also used as an auxiliary with infinitives. In this function it expresses telic or inchoative aspect or future tense (cf. Haseseryn et al., 1997).

In this corpus however, *gaan* + infinitive is used differently: it is used consistently as a narrative tense, for the depiction of events. It is frequently used in combination with verbs with an inherently telic meaning, as in (11), as well as with an iterative meaning, as in (12). This usage is divergent from standard Dutch.

(11) Esma snap-te niet waarom haar zus haar steeds
Esma understand-pst:3sg not why her sister her:obj all.the.time
An unconventional use of *gaan*+infinitive was also observed by Cornips (2000) in the speech of immigrant children (with a Turkish as well as a Moroccan background) telling jokes in Dutch. She argues convincingly that examples from her corpus similar to (11) and (12) cannot be accounted for if *gaan* + infinitive expresses inchoative aspect or a future modal reading.

A possible explanation for this pattern is, that this is an avoidance strategy. Dutch past tenses are often formed irregularly. For example, the past of *aanvallen*, ‘attack’ is *aanviel*, ‘attacked.3sg’. By using *gaan* + infinitive, this non-transparent irregular past of Dutch can be circumvented. Moreover, the construction allows for an easy insertion of Turkish infinitives, as exemplified in (12).

### 5.4.2 Comparison of Dutch Adpositions in (‘*in*’) and uit (‘*out*’)

Standard Dutch has a number of adpositions that may occur either as pre- or as postpositions, with, in many instances, a difference in meaning: as a preposition, they indicate the space where the action of the verb takes place. As postpositions they occur in combination with a verb of movement, indicating the direction of the movement. The adposition *in* ‘in’ as a pre- or postposition has an entirely transparent distribution in standard Dutch:

(13) a Hij wandelde *in* het bos.
‘He took a walk *in* the wood.’

b Hij wandelde het bos *in*.
‘He walked *into* the wood’

For the adposition *in* no remarkable patterns were attested in the corpus: *in* is used in the same way as in standard Dutch, although occasionally expressions diverging from standard Dutch occur:

(14) Ik liep *in* the auto
I walked *in* the car

meaning in standard Dutch: I walked *inside* the car.

intended meaning: ‘I went *into* the car’
As for the adposition *uit*, difference in meaning between pre- and postpositional *uit* is less transparent and largely idiomatic in standard Dutch, where postpositional *uit* is, with intransitive verbs of movement, in many instances obligatory (in any case more frequent), as in (15b):

(15) a. *Hij ging *uit de deur.
    He went *out* the door
b. Hij ging de deur *uit*.
    He went the door *out*

    ‘He walked out of the door’/ ‘He left.’

In the corpus, however, the pattern is different and seems to be dependent on the noun to which *uit* refers. In combination with *huis* (‘house’), *uit* is used prepositionally in 50, postpositionally in 4 instances. Here the pattern diverges from standard Dutch:

(16) Mijn broer was net *uit* huis gegaan toen mijn moeder thuis kwam.
    My brother was just *out* house gone when my mother home came

    ‘My brother had just left the house when my mother came in.’

In standard Dutch, with the noun *huis* postpositional *uit* would be expected, and *huis* would have a definite article *het*:

(17) Mijn broer was net *het* huis *uit* gegaan toen mijn moeder thuis kwam.
    ‘My brother had just left the house when my mother came in.’

By contrast, in combination with *deur* ‘door’ or *kamer* ‘room’, *uit* was used consistently as a postposition, which is in *accordance* with standard Dutch (cf. example 15b above). With other nouns (*straat* ‘street’; *wc* ‘toilet’; *auto* ‘car’; *douche* ‘shower’; *paskamer* ‘fitting room’ etc.) the pattern was not consistent.

A possible explanation for the use of *in* as a pre- or postposition identical with standard Dutch on the one hand, and the consistently divergent prepositional use of *uit* in combination with *huis* ‘house’ on the other, may be based on either of the following factors. First, the difference in meaning between postpositional and prepositional *in* is transparent in standard Dutch. This may explain the fact that no obvious divergence of this pattern can be found in the corpus. Secondly, in 46 of the 50 instances, ‘*uit huis*’ was used without article or other determiner. This is consistent with one rather frequent idiomatic expression in general Dutch, ‘*uit huis gaan*’ which means: ‘(for a grown up child) to leave the parental home’. Possibly the frequent occurrence of *uit huis*
in the corpus has to be attributed to modelling based on this idiomatic expression.

On the other hand, de deur uit gaan ‘leave the house’ and de kamer uitgaan ‘leave the room’ (both with article) are equally frequent units in general Dutch. In the corpus, these expressions occurred with articles as well (respectively 18 instances vs. 0 instances uit de deur gaan and 32 instances vs. 4 instances uit de kamer gaan), whereas in the corpus in general the use of articles is rather inconsistent. A possible explanation for this pattern is the supposedly frequent exposure to these idioms.

The inconsistency of pre- and postpositional use of uit in combination with other nouns may possibly be explained by a combination of lack of transparency and lack of frequency.

A detailed discussion of why this is happening in Turkish-Dutch bilingual language use is beyond the scope of this paper. But the phenomena presented in this in this section suggest that in the Dutch-Turkish bilingual code transparency but also frequency plays a role in the selection of items.

6 Concluding Remarks

In section 5 it has been demonstrated that preference patterns can be attested in a relatively small data set of digital data. This answers the question posed in section 2 above.

The examination of language choice on the macro level and language choice on the lexical level have yielded results that correspond with what is known from the literature on language choice in general (see also Montes-Alcalá, this volume), and descriptions of the Turkish-Dutch situation (e.g. Eversteijn 2011) in particular.

Concerning the regular patterns observed in code switching, more precisely in intimate alternational code switching, discussed in 5.3 above, it is interesting to note that this regularity has not surfaced before in work on Dutch-Turkish code switching and must possibly be attributed to the fact that digital data is used, indicating that digital data yield different but more regular results than oral data. As proposed above, the specific character of digital data, namely, a. the absence of paralinguistic and contextual cues and b. the controlled context, may induce the use of forms that show more regularity, especially in interactions such as code switching, that are very sensitive to the slightest change in context.

The innovative patterns discussed in 5.4 are congruent with the principles of transparency as well as frequency of exposure, in accordance with what is
often observed to be at work in language contact situations. In the data of this study morphological transparency is sought after by avoiding non-transparent features such as the Dutch irregular past and using a form of *gaan* + infinitive instead. Semantic transparency of items such as postpositional vs. prepositional *in* seems to encourage the copying of them. Frequency may serve as a possible explanation for the integral copying of frequent idiomatic expressions with the adposition *uit* ‘out’ vs. the inconsistency of pre- and postpositional use of *uit* in non-idiomatic expressions.

In looking for patterns, this study has focused on some frequently occurring items such as utterance modifiers, auxiliaries and adpositions, and has therefore not much to say about patterns in infrequent items such as content words and collocations. However, the data discussed in 5.2 suggest a tendency to use domain-specific words. The findings suggest that a follow-up study carried out on a large corpus may yield interesting results. The lexicon has been an understudied aspect in code switching studies, which is primarily to be attributed to the fact that, in such studies, researchers have worked with oral data sets, which are, for purely practical reasons, limited in size. Digital data are easier to obtain and to process, although the researcher needs to be aware of the fact that also digital data of exactly the type that one is looking for are not as easy to obtain one might have expected.

References


Backus, Ad, Demiçay, Derya and Yeşim Sevinç. 2013. Converging evidence on contact effects on second and third generation immigrant Turkish. *Tilburg papers in culture studies* 51.


