The simple view of borrowing and code-switching

Title page

The Simple View of borrowing and code-switching[i]

Jeanine Treffers-Daller (University of Reading)

Accepted for publication in the International Journal of Bilingualism (November 2022)

Abstract

Aims and Objectives: In this paper a novel approach to the distinction between borrowing and code-switching is proposed, called the Simple View of borrowing and code-switching. Under this view, listedness is seen as the key condition for classifying words or Multiword Units (MWUs) as borrowings. For MWUs, listedness is operationalised with Mutual Information scores: the higher the MI score of a given set of words, the higher the likelihood it is listed in the lexicon. Under the Simple View, the distinction between borrowing and code-switching is seen as a specific instantiation of the distinction between what belongs in the lexicon (fixed, arbitrary patterns) and what is computed online (productive rules), and should therefore be considered as part of the grammar.

Methodology: Assumptions from the Simple View were tested on a corpus of switches of single words and MWUs from a Turkish-German code-switching corpus (87,000 words), which was transcribed in CHAT format.

Data and Analysis: The frequency of switches in either direction, and their morphosyntactic integration patterns were analysed with CLAN. The formulaicity of the MWUs was analysed with Mutual Information scores through Sketchengine.

Findings: The Mutual Information scores of the donor language MWUs were found to be above 3, which is the cut off point for formulaicity in Corpus Linguistics. Thus, the MWUs were found to be borrowings. In addition, MWUs were found to be more likely to be borrowed than single words.

Originality: Insights about formulaic language from Corpus Linguistics and Second Language Acquisition were used to inform analyses of language contact phenomena, and new ways to test the model are proposed.

Significance: The Simple View offers a unified approach to borrowing of lexical items and function words, and opens a new avenue for research using neuroscientific methods to test whether items are listed in speakers' mental lexicons.

1. Introduction

In this paper, I will argue that the two-way distinction between borrowing and code-switching does not offer a good account for the variety of other-language material that can travel between languages. While single word items from a donor language are often considered to be borrowings and longer other-language stretches are seen as code-switches (Deuchar, 2020; Poplack, 2018), the length of a unit is unlikely to be the defining criterion which allows us to separate borrowing from code-switching. That length is not helpful to decide this matter is clear from the corpus linguistic literature on FORMULAIC LANGUAGE, which Wood (2020, p. 30) defines as follows:

> Formulaic language (FL) is generally defined as multiword language phenomena which holistically represent a single meaning or function, and are likely mentally stored and used as unanalyzed wholes, as are single words.

Formulaic language can take many different forms, and may include MULTIWORD UNITS (MWUs) of varying degrees of complexity, such as collocations, idioms, phrasal verbs, n-grams, and compounds (Wood, 2020). The available evidence shows that formulaic language of any length can enter a language as a borrowing, and that some types of MWUs are more likely to occur in mixed language than single words (Backus, 2003, p. 101).

The occurrences of MWUs in bilingual speech compels us to revisit the definitions of borrowing and code-switching. The alternative account offered here is called the SIMPLE VIEW OF BORROWING AND CODE-SWITCHING. Under this view the key difference between these two is that code-switching draws upon the grammars of the donor language as well as the recipient language, while borrowing does not necessarily imply activation of any grammar at all. I take the view that what defines borrowing is LISTEDNESS and not size of the donor language item[ii], nor integration (Poplack, 2018) nor frequency (Myers Scotton, 1993) in the recipient language vocabulary. Furthermore, it is claimed that the two-way distinction between INSERTION and ALTERNATION (Muysken, 2014) better captures the variability in language contact phenomena than the traditional distinction between borrowing and code-switching.

I also argue that the discussion about the variability in language contact phenomena should be informed by insights from corpus-linguistic approaches to formulaic language, and propose the FORMULAICITY CRITERION, according to which all donor language Multiword Units (MWUs) are borrowings. In the final part of the paper I compare borrowings of LONE OTHER LANGUAGE ITEMS (LOLIs) against borrowings of MWUs (compounds) from a

Turkish-German bilingual corpus, and show that compounds from a donor language are more likely to appear in a recipient language than LOLIs. This, again, illustrates that length is not a defining criterion of borrowing.

2. Distinguishing borrowing and code-switching

One of the most fiercely debated issues in the field of language contact is how to distinguish borrowed material that has become part of the lexical stock of a contact language, and code-switching between two languages. Borrowing is illustrated in (1), where we find the English verb *job*, which has been borrowed into German, and is listed in the Duden Online[iii], with the meaning "to carry out temporary work with the aim to earn a living". The German suffix *-en* has been added to indicate third person plural. In (2), by contrast, longer stretches of two languages (in this case Alsatian German and French) alternate, which is commonly seen as typical for code-switching. In all examples, Dutch, German and Welsh are given in italics, French and Spanish are underlined, English is given in capital letters, Arabic in capital letters and italics, and Turkish in regular type font.

(1) *Das Modell des Studienkontos bietet allerdings den zahlreichen Teilzeitstudenten, die viel nebenher* JOBBEN, *bessere Chancen*.
'The model of the study account offers better opportunities to the many part-time students who job alongside their studies.' (Der Spiegel, 21/2000, p. 67) in Seidel (2010, p. 55)

(2) *Ah ja, noh het er getankt*, <u>mais il devrait donc faire son plein le soir</u>.
Ah yes, now he has filled up, but he should so make his full the evening
'Oh yes, so he filled up, but he should really fill up in the evening.' (Gardner-Chloros, 1991)

In a recent volume on lexical borrowing[iv], Poplack (2018) defines borrowing as 'the process of transferring or incorporating lexical items originating from one language into discourse of another' (Poplack, 2018, p. 6). In the same volume, Poplack proposes that LOLIs are borrowings, whereas 'multiword stretches' from another language are code-switches. Furthermore, borrowings tend to be frequent and wide-spread in a community (Poplack suggest that they should be used by at least ten speakers) and morpho-syntactically integrated

into the recipient language. Thus, *jobben* in (1) would qualify as a borrowing not only because it is a LOLI but also because it has been integrated into German grammar in that a German suffix marking the third person plural has been attached to the root.

The distinction between borrowing and code-switching is, first of all, important for researchers attempting to formulate constraints on code-switching (i.e. rules for where in a sentence code-switching is (im)possible or (un)likely), because theories need to be tested against an unambiguous corpus of code-switches. Second, for psycholinguistic and neuroscientific studies of code-switching, the distinction is important too, because reaction times or ERP signals will differ for other language items that have been integrated into the lexicon of a receiving language, such as those in (1), and for code-switches, such as those in (2). However, distinguishing between both phenomena is difficult, both conceptually and empirically. An overview of the criteria that have been used to distinguish borrowing and code-switching can be found in Table 1.

Table 1 approximately here

The various criteria do not have equal importance, and there are problems with many of these. First of all, bilingual corpora are often very small by comparison with monolingual corpora, which makes it difficult to assess how frequent or wide-spread a word is. In addition, as one reviewer observes, the frequency of a particular item may differ across different bilingual communities, which makes generalizations very difficult. Second, the contact languages may have converged (at the syntactic/morphological and/or phonological levels), which makes it very difficult to assess whether a word has been integrated or not. Third, there are exceptions to all criteria listed here: in Brussels Dutch, for example, many French adverbs are not syntactically integrated (in that they do not trigger Verb Second) despite being listed in dictionaries as a borrowing from French (see below and Treffers-Daller, 1994, for discussion). Conversely, it is possible for a pre-posed Turkish adverbial clause to trigger Verb Second, as in (3), where the adverbial clause *buraya gelirken* 'when I came here' is clearly a code-switch, but it occupies the position in front of the inflected verb *hab*' 'have' (in bold). Thus, the Turkish adverbial clause triggers Verb Second[v].

(3) Ben bura-ya gel-ir-ken **hab'** ich mich gefreut.
I here-DAT come-AOR-ger
(TuGeBic03)

However, there is variability at this point too, because for some speakers a preposed adverbial clause does not trigger Verb Second, as can be seen in (4), where the inflected verb *wußte* 'knew' follows the subject pronoun *ich* 'I'.

(4) *Weißt du was ey*, ben Almanya'dan gel-ir-ken *ich **wußte** nicht wie ....*
Know you what, ey, I Germany-ABL come-AOR-ger I knew not how…
'You know what, ey, when I came to Germany I did not know how…'(TuGeBic10)

Thus, switched pre-posed adverbial clauses can be integrated into German grammar (see also Demske & Wiese, 2016 for further details on variability in the application of Verb Second in varieties of German). For further in-depth discussion of various criteria for borrowing and code-switching, and issues related to these, the reader is referred to Deuchar (2020) and Deuchar and Stammers (2016).

A completely different model emerges from Muysken's (2000; 2013; 2014) work. According to Muysken (2014), we need to distinguish between surface manifestations and underlying processes in language contact. The question then is whether borrowing and code-switching represent truly different underlying processes. In his view, this is not the case because both borrowing and code-switching make use of the same two different basic strategies[vi], namely INSERTION of lexical material into a recipient language which imposes its constraints, and ALTERNATION, where several languages each impose their constraints. Insertion and alternation are different from a grammatical point of view, because insertion involves SELECTED ELEMENTS (e.g. complements of a verb or a preposition) and alternation involves ADJUNCTS (e.g. adverbials or other phrases or words that are rather loosely attached to a clause, or other phrasal categories). Under this view, insertion can lead to borrowing, but should not be equated with it, because borrowing makes use of both insertion and alternation: while many nouns are borrowed through a process of insertion, discourse markers will generally be borrowed through a process of alternation, as they appear at the periphery of a clause and are not embedded into the grammar of the recipient language. This can be seen, for example, in borrowings of French discourse markers such as *pertang* 'however' in Brussels French, which appear at the left periphery of a Brussels Dutch clause, but cannot appear in the canonical position for Dutch adverbs and do not trigger Verb Second (Treffers-Daller, 1994). As for code-switching, on the one hand, this can involve insertion of

constituents, as is the case for insertions of French DET+N sequences in Arab-French code-switching (Bentahila & Davies, 1983), as in and Naït M'Barek and Sankoff (1988), as in (5)

(5) les gens *MABQAW JXALSU:*
'the people stopped paying.' (Bentahila & Davies, 1983)

On the other hand, it can involve alternation, as in (2), where a large chunk in German and a large chunk in French appear in succession. So, under this view, the underlying processes behind the borrowing/code-switching dichotomy, are insertion and alternation.

How insertion and alternation map onto borrowing of LOLIs versus code-switching of full determiner phrases (DPs), which are unambiguous code-switches in Poplack's framework, can be seen in Table 2.

Table 2 approximately here

While *job* in (6) is clearly an insertion, as it is part of an NP that is selected by the verb in the clause, it is possible, or even likely that some LOLIs enter the language via a process of alternation rather than insertion, if they are adjuncts rather than selected elements. This could be the case for the tags and fillers in (7) or for adverbs such as *first* in (10), which are not selected by the verb. It seems unlikely that the same process underlies both the appearance of *job* and the tags.

(10) Les anglais, ils usent leur langue FIRST, pourquoi' c'est qu'on le ferait pas nous autres?
'The English use their language first, why shouldn't we?' (Poplack, Sankoff & Miller, 1988)

Note that *un risqué de condensation* 'a condensation risk' in (8) is considered as an insertion because the DP occupies the first position in the clause and triggers Verb Second. Thus, its

position is clearly different from that of French adverbs, such as *pertang* from French *pourtant* 'however' in Brussels Dutch, which do not trigger Verb Second.

In summary, borrowing groups together phenomena that are clearly distinct (nouns and tags), and code-switching is a cover term for adjoined NPs and selected NPs, which are also syntactically very different from each other. The different processes underlying these phenomena are captured by Muysken's typology but not by the distinction between borrowing and code-switching.

3. Reconsidering the classical division

There are several additional reasons to reconsider the classical distinction between borrowing and code-switching. First of all, Muysken's (2014) division between insertion and alternation is firmly rooted in syntactic theory, and reflects the distinction between selected elements (insertion) and adjuncts (alternation), the relevance of which is attested independently in monolingual grammars. Thus, an extra bilingualism-only mechanism is not created to account for bilingual speech. Second, length (one word versus multiword stretches) is not such an attractive criterion as length is not a fundamental principle of syntactic organization. Admittedly, length plays some role in the success of borrowings, as shown by Calude, Miller and Pagel (2020), who show that shorter donor language items were more successful in being adopted by bilingual speakers than longer 'native' translation equivalents. However, this is not *absolute* length, but *relative* length and relates to the difference in length (measured in syllables) of two possible alternative ways of talking about the same thing, which are in competition in a particular speech community. This is different from assuming that length in words is a fundamental organizational principle behind the distinction between borrowing and code-switching.

Third, the principle of recursion applies to words (in that simple words can become complex through compounding) as well as phrases (in that complex phrases can be built inside other phrases). Thus, words can become very long if the process of compounding is applied recursively, and the same is true for phrases. It would indeed be rather arbitrary to say that only compounds consisting of two words still count as borrowings (e.g. *car industry*), whereas compounds consisting of three words (e.g. *car industry manager*) would be a code-switch. It seems to me that length does not play a key role here, and borrowed compounds could equally well consist of three or more words that have been combined. Thus *wide angle lenses* when used in a Welsh utterance, as in (11), which is classified as a code-switch in

Deuchar (2020), would not be a code-switch under the Simple View, because it is a compound noun that functions as a unit and is inserted as a whole into Welsh[vii].

(11) *pan dach chi 'n defnyddio* WIDE-ANGLE LENSES
when be.2PL.PRES PRON.2PLLL PRT use.NONFIN wide-angle lenses
'When you use wide-angle lenses…'

A fourth reason for reconsidering the typology is that alongside the many prototypical cases of borrowing, such as those in (1), there are also many cases that are more peripheral, in that some but not all of the criteria listed in Table 1 apply. Borrowed compounds are examples of such less prototypical cases, because they include French loan constructions such as *attorney general* or *notary public*, which retain French word order (N+A) (Bauer & Renouf, 2001). The fact that there are exceptions to typologies is, as well as exceptions to rules, is well known is inherent in any typology, because typologies are models, and models are necessarily a simplification of the complexities found in real life (see Simon & Wiese, 2011, for an in-depth discussion).

Most of the discussion about the distinction between both language contact phenomena has focused on the status of LOLIs and on whether or not these can be 'bona fide' code-switches. The problem with LOLIs is that they do not appear frequently enough in data sets to be classified as unambiguous borrowings. For Poplack LOLIs are NONCE BORROWINGS (Sankoff, Poplack, and Vanniarajan, 1990) rather than code-switches, and single word switches are 'exceedingly rare' (Poplack, 2018, p. 7). Other researchers (e.g. Jones, 2005) consider it possible for single words to be code-switches or assume there is a continuum between borrowing and code-switching (Myers Scotton, 1993b, p. 73; Treffers-Daller, 2005). That single word switches do exist, is clear from work on the phonological characteristics of English items in Welsh, as discussed in Deuchar (2020), Stammers and Deuchar (2012) and Deuchar and Stammers (2016), who show that a word's frequency of usage affects the degree of its phonological integration: the more frequently it is used, the more often a Welsh phonological process called SOFT MUTATION is applied to the English word. Thus, the authors conclude that a categorical distinction between words that are borrowed (integrated) and words that are code-switched (not integrated) cannot be made. In fact, there is a continuum between forms that are more integrated (and more frequent), and which would likely be borrowings, and those that are less integrated and less frequent, and which would be code-switches (Deuchar, 2020). In addition, in the Welsh-English data analyses, a distinction is

made, between central and peripheral morphological integration, based on Myers Scotton (1993). While both borrowed and switched single words tend to receive central morphological integration (such as the Welsh suffix *-io* on English verbs, as in *parcio* 'to park', only borrowings also receive more peripheral integration (such as the Welsh soft mutation). Less attention has been paid to semantic integration into the recipient language (but see MacAlister, 2007, for further details). However, semantic integration is not a necessary characteristic of borrowings.

4. The Simple View of borrowing and code-switching

A simpler way of looking at the distinction between borrowing and code-switching might be to say that the defining characteristic of borrowing is that it involves the addition of vocabulary to the recipient language lexical stock, or the substitution of items already in the stock (Albó, 1970; cited in Van Hout & Muysken, 1994), while code-switching does not entail such additions or substitutions. In borrowing, the donor language grammar is not actively involved, but in code-switching it is. In addition, the recipient language grammar *can be* actively involved in borrowing, in that borrowings can be morpho-syntactically, phonologically or semantically integrated, but this *not* a necessary condition for words to become borrowed into a language, as can be seen from the many bare forms that occur in mixed speech (see Budzhak-Jones & Poplack, 1997; Muysken, 2000; Owens, 2005). For approaches which see morpho-syntactic integration as a defining feature of borrowing, such bare forms are problematic. While there are solutions to this, in that patterns and rates of integration (or lack thereof) can be compared to monolingual benchmarks (Poplack, 2018), bare forms are not a problem if listedness rather than integration is considered the key defining feature of borrowing. Lack of integration can also be observed at the level of phonology, as some words retain the phonological characteristics of the donor language, such as the discourse marker *donc* 'so' in Brussels French, which is pronounced with a nasal vowel [dɔ̃k] which is not part of the Brussels Dutch vowel inventory (Treffers-Daller, 1994; see also Holden, 1976, for detailed analyses of the phonological adaptation of loanwords).

Put differently, the dilemma of distinguishing between lexical borrowing and code-switching can be seen as a specific instantiation of the problems involved in delimiting what belongs in the lexicon (fixed, arbitrary patterns) and what is computed online (productive rules), and should therefore be considered as part of the grammar. If this approach is taken, LISTEDNESS becomes the key criterion for distinguishing borrowing and code-switching and not integration (Poplack, 2018) or frequency (Myers Scotton, 1993). In Muysken (2012, p.

71) listedness is defined as "the degree to which a particular element or structure is part of a memorised list which has gained acceptance within a particular speech community." Thus, listedness is linked to the norms in a particular speech community, and such norms may vary depending on the speech community (see Hanks, 2013). On the basis of this definition, the listedness criterion is formulated under (I).

(I) THE LISTEDNESS CRITERION

The key criterion for a word to be considered as a borrowing is listedness in the mental lexicon of the speakers of the recipient language.

Here the assumption is made that speakers know whether or not a particular word or MWU is listed in their lexicon, and can identify these items as such. This issue is taken up again in section 9.

Frequency is obviously related to listedness, as LOLIs that appear more frequently are more likely to become listed, but there are many LOLIs that are infrequent, such as *clafoutis,* which is 'a type of dessert consisting of fruit, typically cherries, baked in a sweet, custard-like batter', and which is listed in the online Oxford English Dictionary (OED). It is a French borrowing, but it is not a frequent word in English according to the OED[viii]. It is no doubt also possible for LOLIs to become *less* frequent in the course of development, when they are replaced by other terms, although the trajectory over time of LOLIs has not been investigated in great detail (but see Chesley & Baayen, 2010; Poplack & Dion, 2012). Thus, frequency is not a necessary condition for LOLIs to be classified as borrowings.

An advantage of the choice of listedness over integration or frequency as the key defining feature for borrowing is that this criterion also applies to FUNCTION WORDS, many of which cannot be integrated morphologically, because they are not inflected, nor can derivational affixes be attached to them in most languages. Thus, morphological routines for integrating foreign verbs into a recipient language, such as *-ieren/-eren* for Romance verbs (*transportieren/transporteren* "to transport") in German and Dutch (Treffers-Daller, 1994) do not exist for function words. In addition, integrating function words syntactically can be difficult because of the lack of equivalence between grammars (Muysken, 2000), and as one reviewer observes, because the meanings of these do not necessarily correspond to each other in two languages. Applying the frequency criterion to function words is also problematic, because borrowed function words may be low frequency, even in large corpora (Poplack, Sankoff & Miller, 1988). A clear disadvantage of the listedness criterion is that dictionaries

of standard or regional languages are often not a good reflection of borrowings that are used in a particular bilingual speech community, although Poplack et al. (1988) and Deuchar (2020) and Deuchar and Stammers (2016) report having been able to benefit from such dictionaries in analysing borrowing in their data. However, as good dictionaries do not exist for all speech communities, what is listed in the speakers' mental lexicon is ultimately what matters (see section 9 for further discussion on how this can be measured).

The existence of Multiword Units (MWUs), such as compounds, collocations, lexical phrases (see Wood, 2020 for an overview) can throw further light on the distinction between grammar and lexis, and on the role of listedness in distinguishing between borrowing and code-switching. Research into lexical processing and corpus linguistics has shown that the lexicon is likely to contain a wide range of MWUs, which are retrieved as such, as unanalyzed units, during language processing. Delimiting what is productive and rule-bound and what is stored as an unanalyzed unit is a key issue therefore not only for the discussion about the distinction between borrowing and code-switching, but also in, for example, the literature on regular and irregular plural formation (Pinker, 2015; Simon & Wiese, 2011). Therefore we turn our attention to MWUs now.

5. Multiword Units in bilingual data: THE FORMULAICITY CRITERION

According to Poplack (2018) compounds may also qualify as borrowings if they function as a single word, as is the case for English *barbershop* in Canadian French. While canonical cases such as *barbershop* may be easily identifiable as compounds, because they are written together, that is not always the case, because different parts are sometimes separated, as the difference between *doorstep* and *front door* illustrates, and some are written with hyphens (*open-handed*). Compounds can also consist of a combination of more than two words, as in *door number plates*. Indeed, in their discussion of around 3,000 new compound formations extracted from a 360 million word corpus from articles in *the Independent*, Bauer and Renouf (2001) demonstrate there is a wide range of phenomena that can broadly be described as compounds, but which includes items such as *lady-in-waiting*, *mother-in-law* (Bauer & Renouf, 2001, p. 103).

That there are many MWUs in language has been known at least since the seminal publication of Pawley and Syder (1983). Importantly, they note 'there is a cline between fully lexicalized formations on the one hand and nonce forms on the other' (Pawley and Syder, 1983, p. 192). Their use of the term 'nonce forms' is particularly revealing because it

underlines that the issue of determining whether something is 'established' or 'nonce' is not specific to bilingual data, but a wider issue in lexicology. Corpus linguistic analyses have since shown that MWUs (also called FORMULAIC SEQUENCES, PHRASEOLOGICAL EXPRESSIONS, PHRASEOLOGY, or LEXICAL BUNDLES), are pervasive in language: according to some estimates, approximately one third to half of the words in discourse is formulaic (Schmitt & Conklin, 2012).

In the field of language contact, Backus (2003) was the first to observe that switches often consist of CHUNKS that do not necessarily correspond to syntactic phrases and appear to be retrieved as unanalyzed wholes, without attention to their internal composition, as in (12), where the Dutch expressions *politiek gesprek*[ix] 'political discussion' and mixed expressions such as *ophouden yap* 'lit. stop do, stop' are examples of such chunks.[x]

(12) *Politiek gesprek*-ler-i *ophoud-en* yap-ın la
political conversation-PL-ACC stop-INF do-IMP man
'*Stop* the *politics conversations*, man' (Backus 1992, 2003, p. 98)

Particularly relevant for the current purposes is that Backus shows that Dutch compounds are more likely to appear as an insertion in Turkish discourse than single nouns. The reason for this is that compounds such as Dutch *arbeidsbureau* 'job centre' (and other fixed expressions) are semantically more specific than single words such as *arbeid* 'work' or *bureau* 'office'. Put differently, *arbeidsbureau* is the conventional expression for the institution unemployed people in the Netherlands can visit to find work, and it is likely more convenient to retrieve this compound from the lexicon during speech production than to try and find a Turkish equivalent for it.

Thus, in the field of language contact it is by now well known that donor language items often consist of more than one word. The question then arises how MWUs are treated in code-switching/borrowing. Deciding whether or not these are borrowings or code-switches is difficult, just as it is for LOLIs, because of the wide range of phenomena that might enter the recipient language as MWUs. Moreover, the wide range of criteria do not always point in the same direction (Deuchar & Stammers, 2016). However, it seems that if a donor language item which appears in a stretch of speech in the recipient language display a degree of FORMULAICITY in the donor language, it is more likely to be a borrowing than a code-switch in the recipient language. An example would be *coûte que coûte* 'cost what it may', which is listed in the Merriam Webster's online dictionary, and can be used even by

speakers who know no or little French. Put differently, when a French-English bilingual uses an expression such as *coûte que coûte* in English, what happens is the transfer of a fixed expression from the lexicon of the donor language to the list of fixed expressions in the recipient language. An analysis of the internal structure of the expression does not take place in this process, because formulaic sequences are processed as a whole. In other words, it is not just the formulaicity in the *recipient* language that is relevant, but first and foremost the formulaicity of the expression in the *donor* language. The latter can be operationalised and measured with Mutual Information scores (MI scores) (Wood, 2020). These scores, which indicate the statistical strength of co-occurrence of two words, can be computed on large donor language databases. MI scores do not have a specific cut off point, but most researchers consider an MI score of 3.0 or higher to show that an expression is formulaic. The assumption of the Simple View is then that if a donor language item is formulaic in the donor language, it is also formulaic in the recipient language: its formulaicity does not change in the process of being transferred to another lexicon[xi]

An attractive option might therefore be to add formulaicity to the borrowing criteria for in Table 1, because for MWUs listedness can be operationalized as formulaicity. A formulation of this criterion is given in (II):

(II) THE FORMULAICITY CRITERION
A MWU from a donor language that occurs in a stretch of speech from a recipient language is likely to be a borrowing if its MI score is high. If its MI score is low, it is a code-switch.

To illustrate how this works, I have computed the MI scores for collocations of the word *Lehrer* 'teacher' in the deTenTen2012 corpus under Sketchengine, having set the left context to -1 and the right context to zero, because I only want to see collocations with words that immediately precede *Lehrer*. The results are given in Table 3.

Table 3 approximately here

Table 3 shows that function words such as *die* 'the' and *dieser* 'this' that immediately precede *Lehrer* obtain very low MI scores, while content words (adjectives) that collocate with *Lehrer* obtain much higher scores[xii]. This happens because articles and demonstratives can occur with any noun. They are not really collocates of the noun, contrary to the content words listed in Table 3. On the basis of these results, we could conclude that if we find

verbeambetete(r) Lehrer "teacher with civil servant status" in a Turkish utterance, it is a borrowing because it has a high MI score, but if we find *dieser Lehrer* 'this teacher' in a Turkish context it is a code-switch. Of course, determining the exact cut off point is difficult. If the criterion of MI scores larger than 3 is used (which is a widely used cut off point in vocabulary studies), this would mean that *unser Lehrer* 'our teacher' is a borrowing but *der Lehrer* 'the teacher' a code-switch, which is not likely to be correct. What constitutes an appropriate cut off point for identifying borrowing in a particular language pair is an empirical question that cannot be answered here. The answer to this question will depend on the speech community under investigation, as there are likely to be differences between speech communities at this point.

The formulaicity criterion is, in fact, a slight reformulation of Backus' (2003) 'unit hypothesis', given in (III).

(III) THE ''UNIT'' HYPOTHESIS:
Every multimorphemic EL insertion is a unit, inserted into a ML clausal frame.

However the formulaicity criterion differs from the unit hypothesis in two ways. First, the formulaicity criterion makes explicit reference to the concept of formulaicity, and second, it does not assume that all units are necessarily formulaic. As is well-known, free phrases are also units, although phrases are not formulaic but built from scratch on the basis of productive grammar rules. In other words, MWUs and free phrases are both units, in the sense intended in the unit hypothesis, but they are different kinds of units, even if the differences are scalar rather than absolute.

As Poplack (2018) also suggests that compounds might qualify as borrowings, this is an obvious group of items on which to test the formulaicity criterion. If formulaicity is indeed an important characteristic of borrowing, one wonders whether MWUs might, in fact, be more successful than single words in entering a recipient language. There is some evidence for this: Compounds were indeed more likely to appear as insertions than single nouns in Backus (2003).

To test the formulaicity criterion, we would need to look beyond nominal compounds, however, because not all languages make productive use of this type of compounds (Sadock, 1998). In French, for example, N+N compounding is not productive: Instead of N+N compounds, NOMINAL GROUPS, such as *sens unique* 'one way street', which resemble syntactic phrases and which may or may not be fixed, are used to express the functions

fulfilled by compounds in English. Thus, if compounds qualify as borrowings, alternative constructions such as the French nominal groups might qualify as borrowings too. But this means opening pandora's box, because in Brussels Dutch, not only N+A phrases such as *sens unique*, but also N+PP phrases such as *femme d'ouvrage* 'cleaning lady' can be borrowed from French (Treffers-Daller, 2005). As a consequence, if structures which resemble syntactic phrases can be borrowings, this challenges the assumption that borrowing mainly involves single words.

It is not currently known which of the other constructions which fall under the broad label of MWUs would also be likely to qualify as borrowings in bilingual data, but the default assumption should be that this is the case for all types of MWUs listed in Wood (2020). However, compounding is a highly productive process in some languages (see also section 7), and therefore language users may well create novel compounds which are not listed in the mental dictionaries of either language. In order to do so, the speaker would need to use the grammar rules for creating compounds in the respective languages. A clear indication that bilinguals can be very creative with compounds is the existence of MIXED COMPOUNDS, as in (13), which has a German head (*Häuser* 'hauses') and an English non-head (*beach*).

(13) BEACH+*häus-er*
'beach hous-es' (Clyne, 2003)

To create novel mixed compounds, the speaker or writer would need to combine grammar rules of both languages, which would be typical of code-switching but not borrowing. Over time, however, such mixed compounds may become part of the lexical stock of a recipient language, and thus be listed as is the case for mixed compounds in Brussels Dutch (Treffers-Daller, 2005). The formulaicity of such mixed compounds can then be measured in the same way as for monolingual compounds.

6. Compounding in Turkish and German

The focus is here on nominal compounds, as this type is productive in both Turkish (Kornfilt, 1997) and German (Donalies, 2007) and the properties of nominal compounding have been described in great detail for both languages. I will begin by defining compounds and summarizing compounding rules for Turkish and German.

6.1 Defining compounding

For the purposes of the current paper, the definition of compounds provided by Granger and Paquot (2008, p. 40) will be used[xiii].

> 'Compounds are morphologically made up of two elements which have independent status outside these word combinations. They can be written separately, with a hyphen or as one orthographic word. They resemble single words in that they carry meaning as a whole and are characterized by high degrees of inflexibility, viz. set order and non-interruptibility of their parts.
> Examples: *black hole*, *goldfish*, *blow-dry.*'

Although many criteria are used to distinguish between free forms and compounds (see Trips & Kornfilt, 2015 for a review), finding criteria that cover all cases and are universally applicable remains very difficult. According to some observers (e.g. Gross, 1996; ten Hacken, 2021), the differences between compounds and phrases are relative rather than absolute. Ten Hacken (2021, p. 19) puts this very nicely:

> (…) for individual speakers (linguists), there are prototypical instances and a gradual transition to non-instances without a clear, natural borderline.

There is a great variety of structures which can be described as compounds. As it is not possible to cover all of these in this paper and verbal compounds have been treated extensively in the literature on bilingualism (e.g. Muysken, 2000), in this paper the focus is on nominal compounds, and in particular on N+N compounds, such as *door knob*.

7.1. Nominal compounds in Turkish

In Turkish, the head of the compound is the right hand side element. That can be seen by comparing (14), where *kitap* 'book' is the head, and the compound represents a kind of book, not a kind of school, while in (14), where *kitap* is the non-head, and the meaning of the compound is a kind of fair rather than a kind of book. In N+N compounds, such as those in (14) and (15), there often is a compound marker on the head of the construction, which

resembles the possessive marker. The specific form of the suffix is determined by the rules for vowel harmony (see Kornfilt, 1997, p. 498-500).

(14) okul kitab-ı
School book-CmpM
'textbook' (Kornfilt, 1997, p. 474)

(15) kitap fuar-ı
book fair-CmpM
'book fair'

There are also other options for N+N compounds, because in some cases the compound marker can be left out, as in (16).

(16) çoban salata
shepherd salad
'shepherd's salad' (Göksel & Kerslake, 2011, p 34)

7.2. Nominal compounds in German

Like in Turkish, nominal compounds[xiv] are right-headed, as can be seen in (17), where *Haus* 'house' is the head, and *Holz* 'wood' the non-head, and the construction refers to a type of house, rather than a type of wood. In (18), by contrast, *Holz* constitutes the right-hand element, and the expression refers to a type of wood. In German compounds are normally written together, which is not common in Turkish.

(17) *Holz+haus* (Donalies, 2007)
wooden house
'wooden house'

(18) *Brenn+holz*
fire wood
'fire wood'

In many cases there is an INTERFIX (German: *Fugenelement*) between the non-head and the head in N+N compounds, as in (19) where *-es* and in (20) where *-er* forms this linking element. Further examples of the range of interfixes in German compounds, their historical development and constraints on their use can be found in Wegener (2003).

(19) *Tag+es+licht*
day+ITF+light
'daylight' (Wegener, 2003, p. 426)

(20) *Kind+er+krankheit*
child+ITF+illness
'childhood illness' (Wegener, 2003, p. 426)

7. The current project

In this paper I set out to test a number of key assumptions of the Simple View of Borrowing and Code-switching against a corpus of Turkish-German code-switching collected in the 1990s, and recently made available to the research community (Treffers-Daller & Çetinoğlu, 2022).

The research questions for the current project were as follows:

1) How frequent are LOLIs by comparison with donor language compounds in both directions?
2) To what extent can donor language compounds be shown to be formulaic using MI scores?
3) To what extent are LOLIs and compounds integrated into the recipient language?
4) Are there any mixed compounds?
5) Which underlying processes (insertion or alternation) are used most frequently for LOLIs and compounds in the data?

A Turkish-German bilingual corpus (TuGeBic) was used to test the assumptions of the Simple View. The corpus contains transcripts of conversations of twelve males and 24 females between the ages of 18 and 50, who were recorded in the 1990s in either Turkey or Germany. Participants were Turkish-German bilinguals, the majority of whom were students, and others were personal friends and family members of the Turkish-German research assistant who collected and transcribed the data (see Treffers-Daller & Cetinoğlu, 2022, for

further details)[xv]. The corpus consists of 87,681 tokens, roughly equally divided between Turkish (43,785 tokens) and German (43,210) and 686 tokens which consist of morphemes from both languages. There are 10,141 monolingual utterances in the dataset and 4,510 bilingual utterances[xvi]. The occurrence of LOLIs and donor language compounds in utterances where there was *only one* switch/borrowing (namely the LOLI or the donor language compound) were coded in the first 10 transcripts of this corpus (N = 20,566 tokens). There were also 88 utterances where more than one unit was mixed, but these have not been included in the current analyses, as determining the matrix language becomes very problematic in such cases. These deserve a more detailed treatment at a later stage. Nouns and compounds in monolingual stretches were also coded, to facilitate a comparison of the frequency of LOLIs and donor language compounds against each other and against monolingual items in the corpus[xvii]. In total 625 single nouns and 64 compounds were coded manually in Turkish utterances, and 483 single nouns and 83 compounds were coded in German utterances (see Tables 4a and 4b for details).

8. Results

In this section the results of the analyses for each research question will be presented.

8.1 The relative frequency of LOLIs and donor language compounds

Table 4a gives an overview of the absolute and the relative frequency of nouns and compounds in monolingual utterances, as well as of LOLIs and donor language compounds in recipient language utterances. It shows that single German single nouns are far less likely to be selected for inclusion in a Turkish sentence (3.52%) than German compounds (20.3%), and this difference is statistically significant (χ^2 = 1434.09, df =3, p <.001). Interestingly, Backus (2003) also reports that Dutch compounds have a 20% chance of being selected for insertion in Turkish discourse, so our findings confirm those observations for German-Turkish code-switching. Table 4b provides the same information for German utterances. The likelihood of Turkish compounds appearing in German is lower (8.9%) but still higher than

that for nouns (5.59%). Again this difference is statistically significant (χ^2 =928.56, df =3, p <.001).

Table 4a and Table 4b approximately here

8.2 The formulaicity of donor language compounds

As there are no dictionaries of German borrowings in Turkish or Turkish borrowings in German, this criterion cannot be used to evaluate the listedness of donor language compounds. Instead, MI scores were computed for donor language compounds. For those Turkish donor language compounds in the TuGeBiC corpus that were attested in the TrTenTen2012 Turkish corpus, the MI scores computed with Sketchengine were higher than 3, which is a cut-off point widely used in vocabulary studies (see Table 5a). For German, MI scores cannot be computed in the standard way with Sketchengine because German compounds are written together, and MI scores can only be computed under this software for words that are written apart. The MI scores for the German compounds were therefore computed by hand[xviii] (see Table 5b). All MI scores found were above 3. We can therefore assume that the compounds found in each language are indeed formulaic.

8.3 The integration of the donor language compounds into Turkish and German

First of all, we note that the compounds in monolingual and bilingual utterances are all right-headed, as might be expected, given the fact that this is the canonical word order for compounds in both languages. Determining to what extent word order in mixed compounds is German or Turkish is therefore not really possible. Second, Turkish N+N compounds that are found in German utterances receive a compound marker as in (21), as would be expected.

(21) *Die ganzen* lağım+su-lar-ı!
the whole sewage water-PL-CmpM
'All the sewage water!' (TuGeBic10)

While *lağım suları* 'sewage water' is clearly integrated into the German DP, in that it is accompanied by a German determiner and an inflected adjective, in most cases, Turkish compounds are not integrated into German DPs, because there is no article in constructions

where this would be expected. In other words, the compounds in (22) and (23) are examples of bare nouns. In (21), the preposition would need to be followed by the determiner *dem*, which is marked for dative case or instead the form *im* (which is a merger of the preposition *in* and the determiner *dem*) would need to be used.

(22) *Wir wurden jetzt geprüft in* Türkçe+ders+i
We were now tested in Turkish lesson+CmpM
'We were then tested in (a) Turkish lesson.' (TuGeBic10)

(23) *Das ist* irade mesele+si
That is will matter+CmpM
'That is (a) matter of will.' (TuGeBic10)

It is possible that determiners are left out because choosing determiners automatically implies allocating a gender to the noun too. This could be problematic for some speakers, as there is no nominal gender in Turkish, and it may not be clear to the speaker which gender would be appropriate. It seems that using a bare noun strategy is the preferred option for inserting Turkish compounds into German. A second strategy in the data is to use the plural form of the compound as for *lağım+su-lar-ı* instead of *lağım+su-yu* in (20), which could be seen as an avoidance strategy on the part of the speaker, as the dative plural inflection is the same for masculine, feminine an neuter nouns. A third option is to use the dislocation strategy, as in (24), because an article is not required there either. The use of *Dings* 'thingie' in (24) could signal the speaker has word finding difficulties on this occasion too.

(24) *Ich hab' Dings mal gelesen*, Çin+horoskop-u
I have thingie once read, Chinese horoscope-CmpM
'I read (a) Chinese horoscope once.' (TuGeBic10)

The absence of German articles before some LOLIs can be a sign of lack of mastery with some speakers of the first generation, who learned German as adults as in (25), which stems from a Turkish-German returnee who lived in Germany for about fourteen years and learned German as an adult. However, most of the speakers in the TuGeBic corpus were born in Germany and had attended German schools.

(25) *Gibt es hier* kapıcı?
Is there here concierge
'Is there (a) concierge here?' (TuGeBic08)

Conversely, German compounds are clearly integrated into Turkish, in that they are combined, for example, with regular Turkish possessive suffixes, as in (26), where the Turkish first person singular possessive marker *-m* 'my' is attached to *Fremdsprache* 'foreign language'. It automatically erases the compound marker, as would be expected according to the Turkish grammar rules (see Kornfilt, 1997 for details).

(26) Çünkü benim asıl *Fremd+sprache*-m Türkisch-dir.
Because my main foreign language-1sg Turkish+COP
'Because my main foreign language is Turkish.' (TuGeBic07)

Integration of LOLIs into Turkish can also be seen in the addition of Turkish plurals, and case marking to German nouns, as in (27), where German *Kassette* is marked with the German plural suffix *-en* as well as the Turkish plural in *-ler*, and an accusative case marker.

(27) *Kassett-en*-ler-i sen al-acak-sın.
Cassette-PL-PL-ACC you buy-FUT-2sg
'You buy the cassettes.' (TuGeBic07)

Turkish LOLIs can be integrated into German, but adding an *-s* plural to a Turkish noun, as in (28) is much rarer than adding a Turkish plural form (*-ler*/*-lar*) to a German noun. Note that *dönem* 'semester' is well integrated into a German DP in that it has the appropriate case and gender markers on the preposition and the adjective.

(28) *Und dann habe ich auch im zweiten* dönem *keine* zayıf-*s*.
And then have I also in.the-DAT second semester no low marks-PL
'And then I do not have low marks in the second semester either.' (TuGeBic03)

That donor language compounds can consist of more than two parts, can be seen in (31), where we find a German compound with a complex non-head in a Turkish utterance, as in (29).

(29) Yani *Sport+lehrer+Ausbildung* mu?
so sports teacher education
'So, sports teacher education?' (TuGeBic07)

While the compound in (29) contains an interfix, there are no German compounds in the Turkish data that contain an interfix. This is probably just accidental, as interfixes are not obligatory for all German compounds, and there are no cases where an interfix is missing in compounds that require one. Among the German compounds in monolingual German utterances from the TuGeBic corpus there are several (30 and 31) with such interfixes.

(30) *Stand+es+amt*
registry+ITF+office
'registry office'

(31) *Mark+en+hose*
brand+ITF+trousers
'brand trousers'

There are hardly any errors with compounding. Two erroneous structures were found, one of which involved the selection of an incorrect derivational suffix (*-ier* instead of *-iv* for deriving an adjective from a verb), as in (32), and the other was an error with a mixed compound (see section 9.4). However, to what extent this indicates the speakers' ability to build compounds was compromised cannot be determined on the basis of these examples.

(32) *Ja , das ist mein Adoptier+sohn*
Yes that is my adopted son
'Yes, that is my adopted son.' (TuGeBic 10)

8.4 Mixed compounds

There are eight mixed compounds in the data (see Table 6), and for most of these the head is Turkish, which may reflect the fact that Turkish is the matrix language of many clauses. One of the exceptions is (33), where the compound consists of a Turkish non-head *kayın* 'in-law'

and a German head *Sohn* ‘son’, which are combined in a mixed compound. Although many kinship terms for in-laws are formed with *kayın*, ‘son-in-law’ is not one of them. The standard Turkish for *son-in-law* is *damat*, which the speaker does not use. Instead, the speaker produces the German translation equivalent *Schwiegersohn* as the next token.

(33) Almanya’da ben-im kayın+*sohn*, *Schwieger*+*sohn*
Germany-LOC I-1sg.POSS in-law+son, in-law+son.
‘My son-in-law is in Germany.’ (TuGeBic01)

For all but one of these mixed compounds, the non-head consists of one lexical item only. The exception is (34), where the non-head is also a compound.

(34) *Nee*, işte *so* *Einkauf*+*s*+*bummel*+yer-ler-e
No, here, such shopping+ITF+spree+place-PL-DAT
‘No, here, to such shopping spree places.’ (TuGeBic10)

Table 6 approximately here

8.5 Underlying processes

As for the underlying processes through which donor language compounds are introduced into recipient language clauses, the overwhelming majority are INSERTIONS, as can be seen in (21) and (26), for example. There are only a few cases of donor language compounds that are ADJOINED, see (24). This is not so surprising, as we noted in section 3, alternations tend to consist of longer phrases. However, the compounds in the current study generally consist of combinations of only two nouns. It is interesting that some of the longer compounds, as in (34) appear on the periphery of the utterance, as is typical for alternation, or are in verbless clauses, as in (29). As the verb is responsible for setting the syntactic frame of an utterance, in the absence of a verb, it is not clear what the syntactic frame consists of. In such cases, it is probably better to assume alternation of chunks in different languages is the underlying strategy.

9. Discussion and conclusion

In this paper I have argued that the key criterion which makes a fundamental distinction between borrowing and code-switching is listedness, and that the advantage of the Simple View over other views of borrowing is that it is applicable not only to lexical borrowing, but also to borrowing of function words. To measure listedness one could investigate whether borrowings are listed in a dictionary of the recipient language (if available), but more important is whether an item is listed in the mental lexicon of speakers of the recipient language. For MWUs, listedness can be operationalized as Mutual Information scores (MI scores), which "indicate how likely a given set of words are to occur together in a set sequence by comparison to chance" (Wood, 2020, p. 38). Here MI scores are chosen to operationalize listedness because an expression with a high MI score is likely to be listed as such in the lexicon. The other borrowing criteria that are often mentioned in the literature apply variably: Borrowings *can* but do not need to consist of just one word, they *can* but do not need to be morpho-syntactically integrated, and finally, they *can* be frequent in a dataset but often they are not.

This view of borrowing and code-switching was subsequently tested against a Turkish-German corpus, the TuGeBic corpus (Treffers-Daller & Çetinoğlu, 2022). The findings largely confirmed the assumptions of the model: It was found, first of all, that compounds had a much higher chance of being selected as a borrowing than single words, in both directions, which confirmed findings of Backus (2003), but would be unexpected if the defining characteristic of borrowing was that borrowings consist of a single word. Second, evidence for the formulaicity of the Turkish compounds was found through the computation of MI scores. Third, morpho-syntactic criteria turned out not to be very useful to establish whether words were borrowings, because a) compounds are right-headed in both languages, and b) Turkish inflection was regularly applied to German nouns, but this process was much less productive in the opposite direction. Indeed, Turkish compounds were often inserted into German as bare forms, without articles that would indicate gender and case. In this context it may be relevant to note that omission of articles is quite common in some varieties of German, in particular Kiezdeutsch, a new urban dialect that was developed by young people with or without a migration background (Wiese, 2011). This means that it may not be appropriate to assume that the standard norms for the use of articles apply to the current data set, which makes studying how borrowings/code-switches are integrated syntactically

becomes very difficult. Third, the frequency criteria could hardly be used because most of the donor language items were very infrequent.

A limitation of the study is that the corpus was small by comparison with other bilingual corpora. It is not impossible that some forms are more frequent if more data are being analysed, but it is unlikely that this would change the overall picture dramatically, because as Poplack, Sankoff and Miller (1988, p. 57) report, 'borrowed words tend not to be recurrent'. This is, in fact, common in all corpora: As Kornai (2007) notes, about 40% to 60% of all word types in large corpora appear only once. Thus, borrowings are not different from indigenous words with respect to their frequency distribution.

In future research, it will need to be established what constitutes an appropriate cut off point for MI scores in studies of borrowing. As some combinations of function words and content words (e.g. *the teacher*) obtain MI scores above 3, it is possible that the cut off point of 3 for MI scores is not the appropriate level for identifying borrowing. It is also possible that different cut off points apply to different speech communities, reflecting community-specific norms.

Further information about community norms could be obtained from experimental approaches. As Deuchar (2020) points out, we need more information about community norms for code-switching, but the same is true for borrowing. Experiments could take the form of a frequency judgement task (Hofweber, Marinis & Treffers-Daller, 2019) for which bilinguals are asked how frequently they encounter a particular mixed utterance with donor language single words or MWUs in their environment. One would expect utterances with LOLIs or donor language MWUs that have been added to a speaker's recipient language mental lexicon to be encountered as frequently as monolingual utterances. Crucially, in such a task, participants are NOT asked to give a grammaticality judgement or say whether they use such sentences themselves. Instead, they are asked to provide a judgement which reflects personal perceptions of community norms. The effects of size, morpho-syntactic integration and frequency on these judgements could then be measured precisely.

Mixed compounds would be of particular interest here because such compounds are novel creations. Thus, these will not be listed in the donor language dictionary. Such mixed compounds should therefore trigger a response that is different from those given to unmixed donor language compounds. Further evidence would come from neuroscientific approaches, as one would expect borrowings of listed items and switches of non-listed items to trigger different ERP signals (see Moreno, Federmeier & Kutas, 2002; Zeller, Hentschel & Ruigendijk, 2015, for neuroscientific approaches to code-switching).

Because the distinction between borrowing and code-switching is essentially a specific instantiation in bilingualism of the basic distinction between words and rules, studying how single words and MWUs from one language are used in another language can also contribute to theory building on the distribution of labour between vocabulary and grammar. This can, however, only be done if the focus shifts from a language contact-internal discussion about the distinction between borrowing and code-switching, towards a discussion of how language contact patterns can contribute to a better understanding of the wider issue of what is rule-bound productive language behaviour and what is stored and retrieved as an unanalyzed whole. This would have the added benefit of research from the field of language contact having a greater chance of being perceived by researchers in neighbouring disciplines (e.g. Corpus Linguistics and Second Language Acquisition).

References

Albó, X. (1970). *Social constraints on Cochabamba Quechua*. Cornell University.

Backus, Ad (1992). *Patterns of Language Mixing. A Study in Turkish-Dutch Bilingualism.* Wiesbaden: Otto Harrassowitz.

Backus, A. (2003). Units in code switching: Evidence for multimorphemic elements in the lexicon. *Linguistics* 41(1), 83–132.

Bauer, L., & Renouf, A. (2001). A corpus-based study of compounding in English. *Journal of English Linguistics*, *29*(2), 101-123.

Bentahila, A., & Davies, E. E. (1983). The syntax of Arabic-French code-switching. Lingua, *59*(4), 301-330.

Budzhak-Jones, S., & Poplack, S. (1997). Two Generations, Two Strategies: The Fate of Bare English–origin Nouns in Ukrainian. *Journal of Sociolinguistics, 1*(2), 225-258.

Calude, A. S., Miller, S., & Pagel, M. (2020). Modelling loanword success–a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, *16*(1), 29-66.

Chesley, P., & Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics, 48(6),* 1343–1374.

Clyne, M. (2003). *Dynamics of language contact: English and immigrant languages*. Cambridge University Press.

Demske, U., & Wiese, H. (2016). Vorfeld, das. In: C. Bluhm, J. Hopperdietzel & L.E. Zeige (Eds.), *Glossarium amicorum. Festschrift für Karin Donhauser*. Berlin: Institut für deutsche Sprache und Linguistik.

Deuchar, M. (2020). Code-switching in linguistics: a position paper. *Languages*, *5*(2), 22.

Deuchar, M., & Stammers, J. R. (2016). English-origin verbs in Welsh: Adjudicating between two theoretical approaches. *Languages, 1*(1), 7.

Donalies, E. (2007). Basiswissen. *Deutsche Wortbildung. Tübingen, Basel: A. Francke Verlag*.

Gardner-Chloros, P. (1991). *Language selection and switching*. Oxford: Oxford University Press.

Göksel, A., & Kerslake, C. (2011). *Turkish: An essential grammar*. Routledge.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. *Phraseology: An interdisciplinary perspective*. In Granger, S. & Meunier, F. (eds.). Phraseology: An Interdisciplinary Perspectives (pp. 27-49). Amsterdam & Philadelphia: Benjamins,

Gross, G. (1996). *Les expressions figees en francais: Noms composes et autres locutions*. Collection l'Essentiel français. Paris: Ophrys.

Hofweber, J., Marinis, T. & Treffers-Daller, J. (2019). Predicting executive functions in bilinguals using ecologically valid measures of code-switching behaviour. In: D. Miller, F.

Bayram, J. Rothman and L. Serratrice (Eds.). *Bilingual Cognition and Language. The State of the Science across its subfields* (pp. 161-180). Studies in Bilingualism, Benjamins Publishing.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.

Holden, K. (1976). Assimilation rates of borrowings and phonological productivity. *Language*, 52(1), 131-147.

Jones, M. C. (2005). Some structural and social correlates of single word intrasentential code-switching in Jersey Norman French. *Journal of French Language Studies*, *15*(1), 1-23.

Kornai, A. (2007). *Mathematical linguistics*. Springer Science & Business Media.

Kornfilt, Jaklin (1997). *Turkish*. London & New York: Routledge.

Macalister, J. (2007). Weka or woodhen? Nativization through lexical choice in New Zealand English. *World Englishes* 26, 492–506.

Matras, Y., & Sakel, J. (Eds.). (2007). *Grammatical borrowing in cross-linguistic perspective* (Vol. 38). Berlin: Mouton de Gruyter.

M'barek, M. N., & Sankoff, D. (1988). Le discours mixte arabe/français: emprunts ou alternances de langue?. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *33*(2), 143-154.

Moreno, E. M., Federmeier, K. D., & Kutas, M. (2002). Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and Language*, *80*(2), 188-207.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Muysken, P. (2013). Language contact outcomes as the result of bilingual optimization strategies. *Bilingualism: Language and Cognition*, *16*(4), 709-730.

Muysken, P. (2014). Deja voodoo or new trails ahead. In R. Torres Cacoullos, N. Dion and André Lapierre (eds.) Linguistic Variation. Confronting Fact and Theory (pp.242-262). London: Routledge.

Owens, J. (2005). Bare forms and lexical insertions in code-switching: A processing-based account. *Bilingualism: Language and Cognition*, *8*(1), 23-38.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and communication*, *191*, 225.

Poplack, S. (2018). *Borrowing. Loanwords in the Speech Community and in the Grammar*. New York: Oxford University Press.

Poplack, S., & Meechan, M. (1995). Patterns of language mixture: Nominal structure in Wolof-French and Fongbe-French bilingual discourse. In L.Milroy and P. Muysken (Eds.) *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching* (pp.199-231). Cambridge: Cambridge University Press.

Myers-Scotton, C. (1993). Common and uncommon ground: Social and structural factors in codeswitching. *Language in Society*, *22*(4), 475-503.

Pinker, S. (2015). *Words and rules: The ingredients of language*. Basic Books.

Poplack, S. (1980). Sometimes I'll start a sentence in spanish y termino en espanol: toward a typology of code-switching. *Linguistics*,18(7/8): 581-618

Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1), 47-104.

Sadock, J. M. (1998). On the autonomy of compounding morphology. In S. G. Lapointe, D. K. Brentari & P. M. Farrell (Eds.), *Morphology and its relation to phonology and syntax* (pp.161 – 187). Stanford: CSLI Publications.

Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language variation and change*, *2*(1), 71-101.

Seidel, U. (2010). *The usage and integration of English loanwords in German a corpus-based study of anglicisms in Der Spiegel magazine from 1990–2010* (Doctoral dissertation, The University of Alabama).

Simon, H. J., & Wiese, H. (Eds.). (2011). *Expecting the unexpected: Exceptions in grammar*. Berlin and New York: De Gruyter Mouton.

Stammers, J. R., & Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, *15*(3), 630-643.

ten Hacken, P. (2021). The nature of compounding. *Cadernos* de *Linguística*, *2*(1), 1-21.

Treffers-Daller, J. (1994). *Mixing two languages: French-Dutch contact in a comparative perspective* (Vol. 9). Walter de Gruyter.

Treffers-Daller, J. (1997). Treffers-Daller, J. (1997). Variability in code-switching styles: Turkish-German code-switching patterns. In R. Jacobson (ed.).*Code-switching worldwide* (pp. 177-200). Berlin: Mouton de Gruyter.

Treffers-Daller, J. (2005). Evidence for insertional codemixing: Mixed compounds and French nominal groups in Brussels Dutch. *International Journal of Bilingualism*, *9*(3-4), 477-506.

Treffers-Daller, J & Çetinoğlu, O. (2022). TuGeBiC – A Turkish German Bilingual Code-Switching Corpus. Language Resources and Evaluation Conference, Marseille, 20-25th June 2022

Trips, C., & Kornfilt, J. (2015). Typological aspects of phrasal compounds in English, German, Turkish and Turkic. *STUF-Language Typology and Universals*, *68*(3), 281-321.

Van Hout, R., & Muysken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, *6*(1), 39-62.

Wegener, H. (2003). Entstehung und Funktion der Fugenelemente im Deutschen, oder: warum wir keine ∗Autosbahn haben. *Linguistische Berichte*, 196, 425–458.

Wiese, H. (2011). Ein neuer urbaner Dialekt im multiethnischen Raum: Kiezdeutsch. *Stadt und Urbanität. Transdisziplinäre Perspektiven*, 146.

Wood, D. (2020). Classifying and Identifying Formulaic Language. In S.Webb (ed.) *The Routledge Handbook of Vocabulary Studies* (pp. 30-45). Abingdon: Routledge

Zeller, J. P., Hentschel, G., & Ruigendijk, E. (2016). Psycholinguistic aspects of Belarusian-Russian language contact. An ERP study on code-switching between closely related languages. *Slavic Languages in Psycholinguistics. Chances and Challenges for Empirical and Experimental Research*, 257-278.

---

[i] I am most grateful to Mareike Keller, Rosemarie Tracy, and participants in the workshop *Constructing Languages: Usage-based approaches to multilingual first language acquisition* (Munich, 6-7th October 2022) for their comments on a previous version of this paper. I am very much indebted to Philipp Wasserscheidt for showing me how to compute mutual information scores for German compounds that are written together. All remaining errors are mine. Finally, I would like to thank the editors of the current special issue for including my paper in it.

[ii] The term DONOR LANGUAGE ITEM is a theory neutral term to cover items that could be either borrowing or code-switching (Deuchar & Stammers, 2016; Poplack & Meechan, 1995).

[iii] Duden Online, https://www.duden.de/woerterbuch [accessed, 15th September 2022]

[iv] Borrowing of functional items or grammatical borrowing is not discussed in this volume. For details see Matras and Sakel (2007).

[v] There are two subjects in this sentence. A Turkish one at the start (ben 'ich') and a German one which follows the inflected verb. This kind of doubling occurs regularly in bilingual data.

[vi] Congruent lexicalization (activation of the grammars and the lexica of both languages in one clause) is seen in Muysken (2014) as a form of codemixing that takes place when two languages are similar to each other, but not as a fundamentally different strategy. Backflagging (the use of L1 discourse markers in L2 discourse) is seen as a subtype of alternation.

[vii] *Wide-angle lense* is also listed in the Merriam Webster's dictionary. According to the information provided by Sketchengine it has a mutual information score between 12 and 14 (depending on whether *lense* is used in singular or plural).

[viii] In the entry in the OED it receives a score of 2 on a scale of 8 (https://www.oed.com/view/Entry/264374?redirectedFrom=clafoutis#eid). A highly frequent borrowing such as *aid* receives a score of 6 on this scale.

[ix] In standard Dutch, the adjective should be inflected with an *-e*, as in *politieke gesprekken* 'political discussions'.

[x] The following abbreviations were used in the glosses: 1sg = 1st person singular; 2sg = 2nd person singular, ACC=accusative, CmpM = compound marker, COP = copula, DAT = dative, FUT = future, ger = gerundive, IMP = imperative, INF = infinitive, ITF = interfix, LOC = locative, PL = plural, POSS = possessive.

[xi] One reviewer observes that large bilingual corpora could also be used for the purpose of computing MI scores. This would certainly be useful if such large reference corpora exist, but that is not the case for most bilingual communities.

[xii] There are different MI scores for *pensionierter Lehrer* (nominative, MI score of 10.33) and *pensionierten Lehrer* (accusative, MI score of 9.69). For reasons of consistency, MI scores are reported for the nominative forms of the collocations only. One reviewer points out that *Lehrer* can be plural as well as singular. Under Sketchengine, it is not possible to select collocations in the singular only, but it is possible to select only singular collocates from the entire list of collocations with *Lehrer*.

[xiii] A defining feature of compounds not mentioned in the above definition is that in some languages, for example English, there are specific stress patterns for compounds. Because of lack of space, this cannot be discussed here any further.

[xiv] As in German compounds are normally written together, a '+' is used to indicate the separation point between the two parts of the compounds.

[xv] One reviewer notes that there could be differences in code-switching patterns between bilinguals living in Germany and in Turkey. There are indications that this is indeed the case (see Treffers-Daller, 1997), but further analysing the similarities and differences between participants belonging to these two groups is beyond the scope of the current paper.

[xvi] It is important to bear in mind that this is a *code-switching corpus*, which means that sections with code-switching were prioritised for transcription. While the relevant (monolingual) context is available for code-switches in this corpus, the number of monolingual utterances is relatively low by comparison with sociolinguistic corpora for which all speech that was recorded was also transcribed. However, this paper does not aim at explaining how frequently code-switching/borrowing occurs in absolute terms, but only at comparing the relative frequency of LOLIs and donor language compounds with reference to a set of monolingual data. It is thus the proportion of LOLIs and donor language compounds in the data that is the focus of attention, and not the proportion of donor language items by comparison with all the monolingual items.

[xvii] One reviewer asked whether the data contained MWUs other than compounds. This is indeed the case. There are many examples of light verb constructions with *yap-* "to do/make", such as *bunlar bestehen yapmadı* "they pass did not" (they didn't pass), where *yap-* is combined with the German verb *bestehen* "pass". It is not really possible to analyse these constructions in the framework of this paper.

[xviii] Personal communication from the Sketchengine team (4th April, 2022). However, as Philipp Wasserscheidt pointed out to me, it is possible to compute the frequency with which each part of a compound co-occurs with the other part, and with other nouns in German compounds. Thus, for example, *Kinderfilm* 'children's film' occurs 16,246 times in the German TenTen corpus. *Kinder-* is found as the first part of a compound 8,282 times and *-film* as the second part of a compound 1,238,978 times. The entire corpus consists of over 17 billion words. We computed MI scores for *Kinderfilm* on the basis of this information.

Table 1. Overview of criteria that have been used to distinguish borrowing and code-switching

feature	Borrowing	Code-switching
Single lexical item	+	-
Multiword unit	- (except for compounds)	+
Syntactic integration	+	-
Central morphological integration	+	+
Peripheral morphological integration	+	-
Phonological integration	+	-
Semantic integration	+	-
Widespread in the bilingual community	+ (except for nonce borrowings)	-
Listed in the mental lexicon of bilinguals or in a dictionary of the recipient language	+	-
Frequent in the recipient language (as shown in a bilingual corpus)	+	-
Replaces (or is in competition with) a recipient language item	+	-
Monolingual users of the recipient language use it	+	-

Table 2. Mapping of borrowing and code-switching onto insertion and alternation			
Element switched/ borrowed	Examples from the literature	Muysken, (2014)	Poplack (2018)
(6) single nouns (LOLIs), complements of the verb	Lorsqu'il trouve un JOB ça va être difficile 'When he finds a job that will become difficult.' (Poplack, 2018)	insertion	borrowing
(7) discourse markers, fillers	Este 'umm', ¿entiendes? 'understand?', I MEAN,	alternation	borrowing

etc, attached to clause in the 'other language'	YOU KNOW, etc. (Poplack, 1980)		
(8) selected DP	Un risque de condensation *heb je* A risque of condensation have you. 'You have a risk of condensation.' (Treffers-Daller, 1994)	insertion	code-switching (subtype: constituent insertion)
(9) left-dislocated DP	Les étrangers, *ze hebben geen geld, he*? 'The foreigners, they have no money, right?' (Treffers-Daller, 1994)	alternation	code-switching

Table 3. Collocations of Lehrer 'teacher' with words immediately preceding it and their MI scores as computed under Sketchengine

Left context	MI score
dieser 'this'	0.71
der 'the'	1.62
kein 'no' (lowercase)	2.17
unser 'our'	3.53
erfahrener 'experienced'	5.84
ausgebildeter 'trained'	7.73
muttersprachlicher	8.98
pensionierter 'retired'	10.33
verbeambteter 'teacher with civil servant status'	12.73

Table 4a. Nouns and nominal compounds in Turkish - tokens (types)

	monolingual Turkish	German items in Turkish	total	percentage
single nouns	602 (455)	23 (22)	625 (477)	3.52 (4.6)
compounds	51 (49)	13 (12)	64 (61)	20.31 (19.6)
total	653 (504)	36 (34)	689 (538)	5.23 (6.31)

Table 4b. Nouns and nominal compounds in German - tokens (types)

	monolingual German	Turkish items in German	total	percentage
single nouns	456 (261)	27 (25)	483 (286)	5.59 (8.74)
compounds	83 (57)	8 (8)	91 (65)	8.79 (12.31)
total	539 (318)	35 (33)	574 (351)	6.10 (9.40)

Table 5a. Turkish compounds in German: translation equivalents, standardized frequency in the Turkish web 2012 corpus (trTenTen12) and MI-scores

Turkish compounds	German translation equivalent	English translation equivalent	Frequency per million	Mutual Information score
Çin+horoskopu	Chinesisches Horoskop	Chinese horoscope	Less than 0.01	n.a.
öğrenci+bileti	Studententicket/ Studierendenticket	student ticket	0.08	5.17
Türkçe+dersi	Türkischunterricht	Turkish lesson	0.39	5.71
geçme+şansı	/Durchfallquote	Chance to pass (pass rate)	0.14	8.61
irade+meselesi	Willenssache	matter of will	0.04	6.02
lağım+suları	Abwasser	sewage water	0.37*	12.61**

yaz+kursu	Sommerkurs	summer course	0.12	6.63
yaz+semineri	Sommerseminar	summer seminar	Not found	-

*Includes singular and plural ** based on the plural form only

Table 5b. German compounds in Turkish: translation equivalents, standardized frequency in the German web 2018 corpus (deTenTen18) and MI scores

German compounds	Turkish translation equivalent	English translation equivalent	Frequency per million	MI scores
Fremd+sprache	yabancı dil	Foreign language	6.58	11.31
Hoch+türkisch	yüksek Türkçe/ İstanbul Türkçesi	High Turkish	Less than 0.01	15.16
Kinder+film	çocuk filmi	kids' movie	0.77	14.76
Real+schule	orta okul	secondary school	6.03	15.95
Sau+bohnen	bakla	broad beans	0.05	15.19

Sport+lehrer+ausbildung	beden eğitimi öğretmeni yetiştirmesi / beden eğitimi öğretmeni meslek öğrenimi	sport teacher education	0.02	10.04
Stern+zeichen	burç	star sign (zodiac)	2.87	16.65
Text+sorte	metin türü	text type	0.75	16.30
Umwelt+minister	Çevre Bakanı	Environment Minister	1.53	16.60
Weiter+ausbildung*	meslekiçi eğitimi	Further Education	27.27	18.54

*In standard German, the expression *Weiterbildung* is preferred. MI scores are based on this compound.

Table 6. Mixed compounds

Mixed compounds	Turkish translation equivalent	English translation equivalent
Dependenz+grameri	bağımlılık grameri	dependency grammar
Lehrer+Diploması*	öğretmenlik diploması	teacher diploma
Einkaufs+bummel+yerlere	alışveriş gezme mekanları	Shopping spree places
Mathe+sınavı	matematik sınavı	maths exam
kayın+birader	Damat/ kayın birader imho	son-in-law
günlük+sprache	günlük dil	everyday language
kazık+frage	tuzak soru	trick question
orta+eins	ortaokul bir (birinci sınıf)	Middle school one (first year)

*The speaker cannot find the word for teacher diploma, but is helped by the interlocutor