Explaining Russian-German code-mixing

A usage-based approach

Nikolay Hakimov

Contact and Multilingualism

language science press

Contact and Multilingualism

In this series:

1. Lucas, Christopher & Stefano Manfredi (eds.). Arabic and contact-induced change.
2. Pinto, Jorge & Nélia Alexandre (eds.). Multilingualism and third language acquisition: Learning and teaching trends.

ISSN (print): 2700-8541
ISSN (electronic): 2700-855X

Explaining Russian-German code-mixing

A usage-based approach

Nikolay Hakimov

language
science
press

Nikolay Hakimov. 2021. *Explaining Russian-German code-mixing: A usage-based approach* (Contact and Multilingualism). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/289

ISBN: no digital ISBN
no print ISBNs!
ISSN (print): 2700-8541
ISSN (electronic): 2700-855X
no DOI
Source code available from www.github.com/langsci/289
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=289

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

To my father, who gave me books, and my mother, who read them aloud to me

Contents

Acknowledgements

This is a slightly revised version of my doctoral dissertation that I defended in 2016 at the University of Freiburg. A lot of people have contributed to the success of this work and I am happy to have the opportunity to thank them. First and foremost, I owe my sincere gratitude to my supervisors Peter Auer and Juliane Besters-Dilger, who, from our first meeting on, were very enthusiastic about my project and supported me from my first days in Freiburg. Your acumen, diligence and insight have been a source of inspiration to me. I thank them and John Nerbonne, the reviewer of my dissertation, for their valuable comments on the previous version of this work.

I am also greatly indebted to the *Deutsche Forschungsgemeinschaft, DFG* and the research training group *Graduiertenkolleg* 1624 "*Frequenzeffekte in der Sprache*" (Frequency effects in language) for funding this work and to Stefan Pfänder, the former speaker of the graduate school, for invaluable personal support. I would also like to thank my colleagues from the graduate school as well as from the German and the Slavic Department of the University of Freiburg.

I remain deeply indebted to the participants of my research for their invaluable contributions and patience as well as to those who so generously assisted me in establishing contacts with eligible participants, including Olga Held of The *Landsmannschaft von Deutschen aus Russland e.V.* (Homeland Association of Germans from Russia), Group Lahr, and Tabea Maire of The *Jugendmigrationsdienst des Caritasverbandes Freiburg-Stadt e.V.* (Youth Migration Service of the Caritas Association in Freiburg) and particularly to the Russian teachers Bettina Lipinski and Friederike Posega from the *Kaufmännische Schule Integriertes Berufliches Gymnasium* (Occupational and Business High School) of Lahr. Without their help this project would never have been realised.

Many thanks go to Ad Backus, Raffaela Baechler, Pia Bergmann, Javier Caro Reina, Eugenio Goria and Marjoleine Sloos, who were willing to discuss my research, and to James Walker for pointing out an important article to me. I am also grateful to the audiences at the 9-th and 10-th International Symposia on Bilingualism and the SKY conference on language contact in Helsinki for commenting on my work, especially in its early stages. For the assistance and advice

on statistical and computational matters I thank Uli Held, Christoph Wolk and Benedikt Szmrecsanyi.

Thanks are also due to Isabelle Léglise and Stefano Manfredi, the editors of the Contact and Multilingualism series of the Language Science Press, for having accepted the manuscript to this series. I am likewise grateful to one anonymous reviewer of the original manuscript for insightful comments and helpful suggestions.

I would finally like to thank all my friends and family for their support and patience, and particularly Tobi, who not only shared his time with me during all these years but contributed to the success of this project by his keen advice, encouragement and technical help.

Transliteration of Russian

Cyrillic	Transliteration	Cyrillic	Transliteration
а	a	р	r
б	b	с	s
в	v	т	t
г	g	у	u
д	d	ф	f
е	e	х	x
ё	ë	ц	c
ж	ž	ч	č
з	z	ш	š
и	i	щ	šč
й	j	ъ	’’
к	k	ы	y
л	l	ь	’
м	m	э	è
н	n	ю	ju
о	o	я	ja
п	p		

Introduction

Bilingual speakers often describe language mixing, which is notorious for its variability, as "messy", but we know at the latest since Shana Poplack's seminal (1980b) work that variation in this inherently manifold phenomenon is largely orderly. Further, as pointed out first by Bokamba (1989) and summed up later by Muysken et al. (1996: 487), variation in code-mixing results from an intricate interplay of structural, psycholinguistic and sociolinguistic factors. Looking back on more than three decades of rigorous research into code-mixing, we can state with certainty that despite the intense research activity in the field, many questions about the structure of this variation and the driving forces behind it have remained unanswered. The main question still open concerns the nature of the motivations underlying code-mixing patterns in bilingual speech. This overarching question can be broken down to more specific questions, such as "What are the explanations for the code-mixing patterns we observe?", "How psychologically plausible are they?" and "How can various motivations be adequately examined in a testable multilevel fashion?". In search of answers to these questions, I will investigate code-mixing between two inflecting-fusional languages, Russian and German.

My research draws on a corpus of Russian-German bilingual speech that I recorded in some of Germany's communities of repatriates from the former Soviet Union and its successor states. The speakers sampled in the corpus include, for the most part, immigrants of the intermediate generation. Code-mixing in their speech is mainly of the insertional type. This means that stems or longer constituents from German are regularly inserted in otherwise Russian sentences. Crucially, German multiword and multimorphemic constituents systematically alternate with mixed constituents, consisting of German stems and Russian inflectional suffixes. The bulk of the insertions appearing in the corpus is constituted by German nouns and their combinations with German and Russian adjectives and prepositions. This tendency accords with the observation reported for other bilingual communities that nouns and nominal constituents are among the most frequent insertions of in the discourse framed by the bilinguals' other language. The high rate of code-mixing observed in contexts involving nouns determined the choice of the specific linguistic phenomena for the distributional

and structural analysis. The grammatical contexts in which patterns of mixing were scrutinised thus include the adjective-modified noun phrase, the prepositional phrase and the marking of plural on noun insertions.

The purpose of this research is to describe variable code-mixing patterns and account for them in terms of competition among several factors. These include usage frequency, linguistic and discursive context as well as distinctive and overlapping properties of Russian and German. Although most of the foregoing motivations have been discussed, or at least adumbrated, in the literature (e.g., Myers-Scotton 1993, 2002, Backus 1996, 2003, Boumans 1998, Muysken 2000), they have neither been subjected to systematic analysis, nor have they been studied in interaction with each other. In exploring the relationship between these factors and the competing patterns, my approach builds on usage-based approaches to language. These theories integrate the gradience and variability of linguistic structures and a psychologically plausible theory of mental representations and thus provide an adequate framework to examine the structure of code-mixing and the motivations behind it.

This book is organised into six chapters. Chapter 1 will set the scene for the empirical chapters that follow. It will define the scope of the term "code-mixing", introduce Muysken's (2000) code-mixing typology, outline social factors influencing the patterning of bilingual speech and survey, albeit briefly, the key approaches to insertional code-mixing, including Myers-Scotton's Matrix Language Frame model (2002, 1993) and Backus's unit hypothesis (2003). By doing so, it will emphasise empirical shortcomings in both approaches and suggest possible solution paths. The remainder of the chapter will discuss the code-mixing versus borrowing controversy. Although from a synchronic view, code-mixing and borrowing can theoretically be viewed as a continuum, I will argue that they are different phenomena by virtue of their different distributions in bilingual speech.

Chapter 2 will provide the theoretical backdrop for the conducted analyses. I will begin by summarising usage-based exemplar models of language, a central tenet of which is that linguistic structure is represented in the mind as memories of specific language experiences as well as in form of generalisations over these memories. Special emphasis will be given to the role of recurrent multiword sequences and multimorphemic words in language acquisition and language processing because they stand at the centre of my analysis of code-mixing. The chapter will close with a presentation of a usage-based perspective on language variation.

Chapter 3 will introduce the participants of my study, Russian German youths and young adults of an intermediate-immigrant generation. I will demonstrate

that in regard to their bilingual abilities and linguistic backgrounds, they constituted a sufficiently homogeneous group so that their speech was well-suited to study insertional code-mixing. The chapter will open with a description of German repatriates from the Soviet Union and its successor states as part of Germany's sizeable Russian-speaking community, and will proceed with an overview of Russian Germans' sociolinguistic history prior to emigration. After outlining the selection criteria for participation and giving details of the participant recruitment, the chapter will introduce the participants first as a group and then individually. Finally, the chapter describes the methods underlying the construction of the corpus and presents the speech situations in which the conversations were recorded.

Chapters 4, 5 and 6 will constitute the core of this book, they will present three case studies tapping into variation in code-mixing patterns in specific morpho-syntactic contexts. Chapters 4 and 5 will investigate code-mixing at the level of syntax, and Chapter 6 will investigate a phenomenon pertaining to the morphological structure of bilingual speech. Specifically, Chapters 4 and 5 will analyse code-mixing in the prepositional phrase and the adjective-modified noun phrase, respectively. In these syntactic contexts, German constituents inserted in otherwise Russian sentences, sometimes referred to as embedded-language islands, alternate with mixed constituents. The alternating patterns studied in Chapter 6 will concern German noun insertions which either retain their German plural marking and thus form the so-called internal embedded-language islands, or receive Russian inflectional suffixes and form mixed plurals.

Each of the three chapters capitalizes on the distributional properties of a specific structural template. A noun occurring in a given structural template combines with the other part of the template – a plural-marker, a preposition, or an adjective – with varying probabilities, depending on the number of forms participating in the distribution. While the use of nouns in plural contexts is linked to the competition between two forms: the German noun stem, coinciding with the base form, and the German plural, the distribution of a noun's collocates in the prepositional phrase usually involves some ten prepositions, and in the adjective-modified noun phrase, a noun may appear with a virtually unlimited number of adjectives. Hence, the three structural contexts complement each other since the different properties of the explored structural patterns have varying repercussions for the effect of co-occurrence frequency on mixing patterns and are thus worth comparing.

Taken together, the three case studies embrace a range of understudied phenomena of bilingual speech, covering typical embedded-language islands (Chapters 4 and 5) and internal embedded-language islands (Chapter 6). The chapters

have the following structure: The existing explanations for the scrutinised phenomena will be reviewed, including structural non-equivalence and frequency of co-occurrence, and complemented by further possible motivations, such as word repetition in discourse and morphophonological regularities. Systematic analysis and its results will be presented for each of the examined factors, their interplays will be evaluated statistically. The remainders of the chapters will summarise and discuss the results.

I will conclude this work by recapitulating its main findings, comparing predictors of the variation examined in each of the case studies and sketching some promising avenues for future research.

1 Previous research on the grammar of code-mixing

This chapter presents an overview of current approaches to the grammar of code-switching/mixing, which is defined as the juxtaposition of two or more languages, or varieties, in discourse. After more than thirty-five years of thriving research, research in code-switching/mixing has developed itself into a well-established branch of linguistics, which has found its way into introductory linguistic textbooks. The study of code-switching/mixing has strong interdisciplinary links to diverse disciplines, such as sociolinguistics, conversation analysis, descriptive linguistics, language contact, language acquisition, linguistic anthropology, psycholinguistics and neurolinguistics. Given the substantial increase in the amount of literature devoted to code-switching/mixing, it is impossible to provide a complete overview of the field. Most comprehensive surveys of recent work are *The Cambridge handbook of linguistic code-switching* edited by Bullock and Toribio (2009) and the volume *Code-switching* authored by Gardner-Chloros (2009). An apposite review of the earlier literature on grammatical aspects of code-mixing/switching is provided by Boumans (1998: Chapter 1). For this reason, I will only outline key modern advances in the field, which are relevant for the research presented in the following chapters.

Before delving into theoretical aspects of code-mixing, the chapter will address an important terminological issue requiring disambiguation, the notorious dichotomy code-mixing versus code-switching. The chapter will thus open with a brief overview of the history of the terms and explain the division adopted in the present work. On tackling this issue, the chapter will proceed with the typology of code-mixing proposed by Muysken (2000). Recognising the variability inherent in code-mixing patterns, Muysken attempts an effective general classification and relates the identified types of code-mixing to structural, social, and psychological factors. A separate section will be devoted to social factors, which have been considered crucial to the emergence of code-mixing patterns and their variability. As code-mixing in my Russian-German data is expected to be of the insertional type, I will present and discuss one of the most elaborate approaches to insertional code-mixing the Matrix Language Frame (MLF) model, authored

by Myers-Scotton (1993, 2002). multimorphemic and multiword insertions, a distinct type of insertional code-mixing, will be covered in a separate section. I will specifically look at these insertions from the perspective of the unit hypothesis and the "conceptual unit" hypothesis, articulated by Backus (1999a, 2003). The closing section of the chapter will focus on the ongoing controversy over the status of code-mixing and borrowing as distinct processes, and the proposed diagnostic criteria to distinguish between them.

1.1 A preliminary remark on terminology

A juxtaposition of two or more languages in discourse has been studied from various perspectives and been referred to by different terms: code-alternation (Johanson 1993, Boeschoten & Broeder 1999, Thomason 2001, Migge & Léglise 2013), code-switching (Blom & Gumperz 1972, Poplack 1980b, Myers-Scotton 1993, Backus 1996), code-mixing (Muysken 2000, Muhamedowa 2006), language alternation (Auer 1984, Maschler 1998) and language-mixing (Pfaff 1979b, Backus 1992, Lanza 2004). Researchers often use more than one of these terms to refer to different phenomena of bilingual speech. For example, Thomason (2001) employs the term *code-switching* for "the use of material from two (or more) languages by a single speaker in the same conversation" (p. 132), whereas the term *code-alternation* is reserved for the diglossic use of languages by the same speaker (p. 136). Among the aforementioned terms, *language-mixing* and *code-switching* have the longest history.

The use of the word *language-mixing* in academic discourse goes back to Haugen (1953a). He adopts this layman's term (cf. p. 58) to refer to the "confusion of patterns" observed in bilingual speech (p. 53). Alongside *language-mixing*, Haugen speaks of *switching* from one language to another. Although he does not provide a definition of the term, he relates it to the notion of *switch*, which he defines as "a clear break between the use of one language and the other" (Haugen 1953a: 65). The term *language-mixing*, as employed in Haugen (1953a), can be interpreted as the umbrella term for both switching languages and borrowing. However, in the second volume of *The Norwegian Language in America: A Study in Bilingual Behavior* (1953b) he abandons this term in favour of *borrowing*. He explains his choice by the inadequacy of the term *mixing* when applied to bilingual speech.[1]

[1]Haugen writes, "Mixture implies the creation of an entirely new entity and the disappearance of both constituents, or a jumbling of a more or less haphazard nature. But speakers have not been observed to draw freely from two languages at once, aside from abnormal cases. They

The first mention of the term *code-switching* in relation to the juxtaposition of two or more languages is attributed to Vogt (1954). In his review of Weinreich's *Languages in Contact*, Vogt refers by this word to the alternate use of "languages as well integrated systems, as codes" (p. 81). However, Alvarez-Cáccamo (1998) reports that the idea of "switching code" was previously articulated by Jakobson et al. (1976 [1952]), who adapted this notion from information theory and related it to "coexistent phonemic systems" in a speaker's mind (p. 11). In the terrain of morphosyntax, it was Haugen (1974 [1956]), again, who explicitly defined *code-switching* as "the alternate use of two languages" (p. 40) and contrasted it with *interference* and *integration* (p. 50). According to Alvarez-Cáccamo (1998), the work of the 1950s and early 1960s is characterised by a lack of consistency in the use of the terms *code-switching* and *language-mixing*. This situation is obviously the source of the terminological discrepancy observed in the field to date (for overviews, see Poplack 2004, Treffers-Daller 2005a).

In the present work, I will approach phenomena of bilingual speech using the typology proposed by Auer (1999). Auer distinguishes between the cases of *code-switching* and *code-mixing*, depending on the meaning participants ascribe to these events. Participants may perceive and interpret the juxtaposition of two codes, or languages, as locally meaningful, i.e., they may use this juxtaposition as a conversational device contributing to the local management of conversation. In this case we deal with code-switching. In the case of code-mixing, the alternate use of two languages is meaningful only in a global sense, that is, "as a recurrent pattern, but not in each individual case" (Auer 2011: 467). The traditional division between these terms is based on whether switching occurs on the sentential level or below, it is thus distinguished between inter-sentential code-switching and intra-sentential code-mixing (cf. Auer 2011: 467, Clyne 2003: 70–73). The extent of the term *code-switching*, as defined by Auer, coincides with that of the traditional definition in so far as code-switching in Auer's sense indeed occurs more often among full syntactic units. However, the meaning-based perspective taken by Auer is more pervasive and therefore preferable than the traditional, form-based dichotomy because the former can account for transitional cases of code-switching, such as intra-sentential switching with local meaning attributed to it. Below, I will build on the opposition *code-switching* versus *code-mixing*[2] to

may switch rapidly from one to the other, but at any given moment they are speaking only one, even if they resort to the other for assistance" (1953b: 362).

[2]Margaret Deuchar (p.c.) has pointed me to the fact that the term "code-mixing" has a negative connotation. This is probably one reason why this term is indeed infrequent and perhaps even intentionally avoided in studies originating in the USA and the UK. Another reason may be the tradition traced back to Haugen's explicit rejection of the term "mixing". Nevertheless, in

describe linguistic data and to review relevant theoretical approaches. Following Muysken (2000), I will occasionally use the term "switch" to refer to a particular site where the juxtaposition of languages occurs.

1.2 Typology of code-mixing

Until recently researchers have not discriminated between different kinds of code-mixing. In studies that originated in the late 1990s, however, we can observe an interest in greater differentiation of phenomena pertaining to bilingual speech and their relation to each other (cf. Auer 1999). A most comprehensive typology of code-mixing, which draws on samples of bilingual speech encountered in various bi- and multilingual communities worldwide and involving different language constellations, has been proposed by Muysken (1997). He further elaborates this model in the volume *Bilingual speech: A typology of code-mixing* (2000). In this approach, the grammatical patterning in code-mixing yields three major structural types, which are brought into relation with typological, psycholinguistic and sociolinguistic factors. The latter circumstance makes this typology particularly appealing for analyses of bilingual speech and I will sketch it briefly below.

According to Muysken, code-mixing covers "all cases where lexical items and grammatical features from two languages appear in one sentence" (2000: 1). Although this definition ignores the issue of meaning that participants can attribute to each case of juxtaposing codes in conversation, the scopes of this definition and the definition given by Auer (1999) seem to roughly overlap (cf. §1.1). Yet, Muysken's definition is broader than Auer's in that the former includes cases of juxtaposition of codes within one sentence which may be perceived and interpreted by conversation participants as meaningful.

The typology distinguishes between three basic code-mixing processes: *insertion* of material from another language, *alternation* between structures of the involved languages and *congruent lexicalisation*. In *insertional code-mixing*, a lexical item, or an entire constituent, of language A is used in a structure of language B, where language B, called base language, usually maintains the frame for the overall clause. Insertions thus exhibit a nested B A B structure such that an insertion from language A is preceded and followed by fragments of language B that are grammatically related (cf. Muysken 2000: 61–69). This type subsumes lexical borrowing, which may be regarded as a special case of insertion. Insertions

order to maintain consistency with the established typologies of bilingual speech, I will use the term "code-mixing" without intending to evoke negative associations.

are open-class words that function as complements rather than adjuncts, and they are usually subject to morphological integration. It is useful to distinguish between two kinds of insertion: minimal and maximal insertions. A minimal insertion occurs when a stem from another language combines with grammatical formatives of the base language, a maximal insertion refers to the case when both the inserted stem and the accompanying grammatical formatives come from the same language (cf. Auer 2014), for instance:

(1) Swahili-English (Myers-Scotton 1993: 80)
Leo *si-ku*-come *na* books *z-angru* […]
today 1SG.NEG-PST.NEG-come with CL10-my
'Today I didn't come with my books.'

In (1), the English verb *come*, inserted in the Swahili matrix structure, receives Swahili verbal prefixes, but the nominal insertion *book* bring in its English plural suffix. Whilst the minimal insertion *come* undergoes morphological integration here, the plural *books*, being a maximal insertion, is subject only to syntactic integration.

The next type of code-mixing is *alternation*. As the name suggests, the two languages alternate in a clause so that "there is a true switch from one language to the other, involving both grammar and lexicon" (Muysken 2000: 5). Thus, the speaker begins a sentence in one language and switches over to the other language. Alternation can occur at any point in the clause provided that the syntactic structures of the languages involved exhibit linear word order equivalence at that point (cf. Muysken 2000: 114). For example, in (2) the alternation occurs after the Moroccan Arabic relativiser *lli* 'who'; at this point the syntax of Moroccan Arabic is equivalent to that of French.

(2) Moroccan Arabic-French (Bentahila & Davies 1983: 311)
kajn bzzaf djal nna: lli ne font rien
there are many people who NEG do.PRS3PL nothing
'There are many people who do nothing.'

In (2), French is maintained throughout the switched fragment: from the switch site to the end of the clause. In other words, there is no 'return' to the language of the initial fragment after the switch. This case is one subtype of alternational code-mixing, its structure can be represented as A … B. In this case, the switched sequence involves several immediate constituents (cf. Muysken 2000: 96). In (2), for instance, the French fragment encompasses a verb phrase and a noun

phrase. The switched sequence may occasionally coincide with a whole subordinate clause, as shown by Pfaff (1979b: 312) and especially by Treffers-Daller (1994: 196–200). In the extreme case, it may exceed the clause boundary, as is observed in the example below:

(3) English-Kashmiri (Bhatt 1997: 228)
… and they also made *gadɨ kyazyiki sanyi bil-as chi bad kʰof kəran.*
fish because our Bily-DAT is very happy does
'And they also made fish because our Billo (son's name) is very happy.'

The beginning of the switched fragment in (3), the Kashmiri *gadɨ* 'fish', may remind one of an insertion. We could assume that the Kashmiri noun may have been inserted into the matrix of the English sentence produced so far. However, an insertion would be possible here only if the noun were not followed by another syntactic unit in the same language, either a phrase or a subordinate clause. This pattern, although not "clearly alternational" (Muysken 2000: 102), still exhibits some of the characteristic features of alternation. Notably, the sentence in (3) has a distinct point of alternation from English to Kashmiri, and the switched sequence comprises a noun phrase and a subordinate clause. According to Auer (2014: 303) instances like (3) are quite common in bilingual corpora. As can be gleaned from examples (2) and (3), switched fragments pertaining to the discussed subtype of alternation are long and complex.

Another subtype of alternational code-mixing exhibits a non-nested structure A … B … A (cf. Muysken 2000: 97–103). This means that a fragment of language B is preceded and followed by fragments of language A, but these latter fragments are not related grammatically. In this regard, the two Hungarian fragments in (4) – *szóval* 'well' and *hanem kottára* 'but with notes' – are not in syntactic relation, and we can thus analyse the example as an instance of alternational code-mixing.

(4) German (dialect)-Hungarian (Szabó 2010: 260)
szóval net abgedreckt, *hanem kottá-ra.*
well not printed but note.SG-SUB
'Well, not a printed [hymnbook], but one with notes.'

Alternational code-mixing of this subtype commonly involves discourse markers, adverbs, adverbials, coordinated constituents, clefts and tags. Many of these elements usually occur at the periphery of the clause (Muysken 2000: 97–99). Both switching sites in (4) can be described as peripheral: First, the peripheral position of the Hungarian discourse marker *szóval* 'well' is unambiguous. Second, the Hungarian sequence *hanem kottára* 'but with notes' is a coordinated

constituent, which means that the speaker could have finished her turn just before this sequence, i.e., after producing the participle *abgedreckt* 'printed'. Thus, the second Hungarian sequence is also located at the periphery.

At first sight, the two subtypes of alternational code-mixing, as exemplified in (2) and (4), may appear to be at odds with each other. In (2), the switch is placed within a complementiser phrase and the switched fragment is long and complex, whilst in (4), the switches are placed at the clause periphery and the switched fragments are short. Nevertheless, the two subtypes can be conflated because in both cases none of the involved languages is superimposed and the switched fragments are autonomous of each other. Furthermore, the two subtypes are understood as abstractions, and, of course, they can combine in bilingual speech.

One aspect of the outlined approach has been identified as problematic. According to Muhamedowa (2006: 7), certain syntactic structures cannot be clearly assigned to one of the two types: alternation or insertion. She illustrates this point with the case of prepositional phrases, which are handled as instances of alternation in the typology. Yet, she contends that for mixed sentences containing prepositional phrases in the other language, the identification of the base language is non-ambiguous, and therefore, the other-language constituents may rather count as insertions (p. 8). Interestingly, in an earlier version of his model Muysken (1997) illustrates insertional code-mixing with the following example:

(5) Spanish-English (Pfaff 1979b: 296)
Yo and-uv-e *in a state of shock* pa dos día-s.
1SG walk-PST-1SG for two day-PL
'I walked in a state of shock for two days.'

Here, the English prepositional phrase *in a state of shock* is analysed as an insertion. According to the later version of the model, this example would obviously be regarded as an instance of alternation. Generally speaking, the case of prepositional phrases shows that insertional and alternational code-mixing should be considered as two non-discrete categories (cf. Backus 1996: 95).

The third type of code-mixing is *congruent lexicalisation*. Muysken (2000: 3–4) speaks of congruent lexicalisation when the grammatical structure shared by the languages in contact is filled by lexical items belonging to either language. The involved languages have to therefore exhibit a great extent of both linear and categorical equivalence such that lexicalisation is made possible. Congruent lexicalisation is also likely when the languages at play manifest a low degree of linear equivalence, but a similar vocabulary including homophonous words, which can trigger code-mixing. Muysken (2000: 123) draws on Dutch-English

bilingual speech to illustrate this pattern. Although Dutch and English exhibit some divergence in word order patterns, their vocabularies overlap to a considerable degree. In this case, homophonous words function as triggers for code-mixing. This scenario can arguably unfold in a situation of contact between any two closely related languages as they often share a large stock of words. Let me illustrate congruent lexicalisation which results from contact between two closely related languages Russian and Ukrainian:

(6) Ukrainian-Russian (Vahtin et al. 2003)
To ž oce pered *prazdnik-om* *plit-k-u*
PTCL PTCL PTCL before holiday-INSTR.SG.M cooker-DIM-ACC.SG.F
pomy-l-a.
wash-PST-SG.F
'Well then before the holiday [I] washed the cooker.'

Russian and Ukrainian not only have a similar vocabulary like English and Dutch, but also share a large part of grammatical structure. Although both languages are traditionally described as free-word order languages, the constituent order observed in (6) is only relatively free: it serves to express topic-focus relations. The topic generally precedes the focus in both Russian and Ukrainian. As the grammars of the contact languages require, the topic in (6), realised by the mixed phrase *pered prazdnikom* 'before the holiday', is fronted. The adpositional phrase in question is headed by the Ukrainian preposition *pered* 'before', which projects the instrumental case on the Russian singular noun *prazdnik* 'holiday'.[3] The remainder of the sentence *plitku pomyla* 'washed the cooker' expresses the focus. As both languages use the same syntactic pattern to code topic-focus relations, it is impossible to say whether the constituent order in (6) is Russian or Ukrainian. The word order in the adpositional phrase is again identical for both languages since both heavily rely on prepositions. Crucially, not only linear but also categorical equivalence plays a important role in the examined instance. Notably, patterns of case assignment observed in the prepositional and verb phrases are the same: the preposition *pered* 'before' assigns its complement the instrumental case in both languages, and the Russian verb *pomyt'* 'wash', just like its Ukrainian counterpart *pomyti*, assigns its complement the accusative case. Following Muysken (2000: 129), the switch within the prepositional phrase *pered prazdnikom* is conditioned by the structural equivalence in the prepositional phrase and the

[3] Juliane Besters-Dilger (p.c.) has drawn my attention to the fact that the use of the preposition *pered* in the temporal meaning is far less common in Ukrainian than in Russian and the preposition could thus be analysed as Russian rather than Ukrainian. I thank her for this observation.

noun phrase. He asserts that structural equivalence leads to multi- and non-constituent mixing, another feature of congruent lexicalisation (Muysken 2000: 129).

Let me now turn to the lexical correspondences observed in (6). Here, the Russian *plitku* 'cooker', used in the accusative case, is virtually identical to the corresponding Ukrainian form *plytku*. The two words differ in the stem vowel – Russian /i/ versus Ukrainian /ɪ/ – and the preceding consonant, i.e., Russian /lʲ/ versus Ukrainian /l/. The verb form *pomyla* is Ukrainian and deviates from its Russian equivalent only in the vowels, namely, the Ukrainian [pɔˈmɪła] contrasts the Russian [pʌˈmiłɑ], or [pʌˈmiłə]. We can thus consider the words *plytku* and *pomyla* as homophonous diamorphs. They illustrate the dependence of congruent lexicalisation on a common vocabulary stock. With the exception of the Ukrainian discourse marker *to ž oce* 'well then', the only lexical item in (6) that contributes to lexical divergence is *prazdnik* 'holiday', its Ukrainian equivalent is the word *svjato*. Since the instrumental marking on the noun *prazdnik* is also identical in both languages, we may assert that the structural equivalence maintains lexical insertion here. Crucially, structural congruence as such does not have to be total. For example, congruent lexicalisation may be observed when mixing occurs between languages such as Spanish and English. In this case congruent lexicalisation results from partial congruence (cf. Muysken 2000: 6).

A important generalisation concerning congruent lexicalisation is that categorical and linear equivalence lead to a situation in which mixing is syntactically unconstrained. Any category can be switched, and even word-internal switching is possible. For this reason Muysken views congruent lexicalisation as akin to style shifting and language variation, such as observed between a standard and a dialect (2000: 127–128). We can thus conclude that congruent lexicalisation depends on structural equivalence, lexical correspondence, or both, as in (6), and it is distinguished by non-nested *a b a* structure (Muysken 2000: 129).

The comprehensive character of the typology makes it an attractive tool for investigating the linguistic patterning of bilingual speech. Analysis of code-mixing aimed at probing into linguistic variation and change as well as their correlates draw on the mixing types (and their subtypes) to describe patterns in naturally occurring bilingual speech. For instance, Hoi Ying Chen (2015) identifies distinctive code-mixing styles in Hong Kong, with one style allowing for insertion and alternation, and the other involving only insertion. Another example is the work by Goria (2018, 2021), in which he reports differing alternational patterns in Gibraltar's Spanish-English bilingual speech and attributes them to three generational cohorts of bilingual speakers and ultimately to an ongoing language shift from Spanish to English. Although these studies demonstrate the usefulness of

this approach, some of its caveats and intricacies should be mentioned. First, all three code-mixing types can co-occur in a corpus of bilingual speech (Muysken 2000: 229). It is therefore necessary to determine the dominant code-mixing type in the corpus. Second, "intermediate cases may exist" (Muysken 2000: 229). As discussed above, prepositional phrases form a category with an intermediate status. Adverbials, which count as indices of alternational mixing, may in principle be analysed as inserted into a base language.

Another merit of the outlined typology is that it relates the identified code-mixing types to linguistic, or typological, and extralinguistic, or socio- and psycholinguistic, factors (Muysken 2000: 221–249). The evaluated correlations allow predictions about the predominant code-mixing type in a specific community given the contact languages' typological profiles, the speakers' language dominance patterns, the specifics of the sociolinguistic situation and other factors. Considering Muysken's conclusions, it is possible to assume, for instance, that the predominant code-mixing type in the bilingual speech of Russian Germans in Germany, the subject matter of this book, will be insertion. At the grammatical level, Russian and German, being fusional languages, exhibit a high degree of typological proximity, but their word order patterns and core vocabularies are not similar enough for congruent lexicalisation to emerge. The social conditions of the examined situation, namely, repatriation after a language shift to Russian (for details, see Chapter 3), also favour insertion. Finally, the speakers' bilingual proficiency in Russian and German allows for intensive code-mixing. The three groups of factors work jointly and are often included in analyses of bilingual speech as individual independent variables (cf. Muysken et al. 1996). At the same time, many researchers have emphasised that social factors take priority over other factors in language contact. In what follows I will discuss this issue and illustrate the role of social factors in code-mixing by providing examples from the literature.

1.3 Social factors in code-mixing

This section begins with an outline of the existing attempts to systematise social factors influencing the linguistic structure of bilingual speech and then showcases the factors that are particularly relevant to the linguistic community and the speakers whose bilingual speech is analysed in this book. Before I turn to these topics, I discuss the claim that among other factors, social factors play a major role in code-mixing, and in language contact in general.

1.3.1 Social factors versus structural and psychological factors

Researchers who relate differing outcomes of language contact to distinct types of sociohistorical contexts in which language contact occurs, particularly Sarah (Sally) Thomason, have strongly advocated for the view that "when social factors and linguistic factors might be expected to produce opposite results in a language contact situation, the social factors will be the primary determinants of the linguistic outcome" (2008: 42). This position is not alien to Muysken (2000). In a treatment of structural factors, he acknowledges the role of categorial equivalence between linguistic structures of the contact languages, but emphasises that historical and sociolinguistic factors are more reliable for determining the likelihood of a specific mixing pattern in a given situation because "categorial equivalence is not a purely objective notion" (p. 247). Although quite plausible, this reasoning does not elucidate the nature of the relation between social factors and congruence, even when the latter is regarded as a subjective phenomenon. An account of this relation may build on the view that identification of equivalence, or what Weinreich (1979 [1953]) labelled as "interlingual identification", is bound to a single individual (see the discussion in Gardner-Chloros 2009: 37) and is intrinsically embedded in social interaction. Since interlingual identification, being an effect of a more general process of similarity detection, depends on an individual's previous linguistic and interactional experience (for more details, see Hakimov 2017, Hakimov & Backus n.d.[a]), the outcomes of this process are diverse and indeterminate. Yet, it is more likely than not that they are constrained by the individual's interactions in social networks, or a larger speech community.[4] Innovative usage patterns attributable to individual unconventional interlingual identifications may either go unnoticed in interactions, or be sanctioned by the speaker's interaction partners; alternatively, they may be perceived and adopted by the interaction partners and gradually diffuse in the speech community. In other words, the emergence of innovative usage patterns resulting from subjective interlingual identifications, or individually established congruence, is inseparable from the interactions in the social networks and the community. Innovations may also spread to neighboring bilingual communities and give rise to local norms.

A recent example of such a locally emerged innovation is provided by Bullock et al. (n.d.). The authors report an unconventional use of the Spanish verb *agarrar*, meaning 'grab' or 'grasp', in a corpus of Spanish spoken in Texas. Utilising

[4]Pertinent mechanisms at play are accommodation (Giles 1980), and focusing (Le Page & Tabouret-Keller 1985).

variationist and corpus-linguistic methodology, they demonstrate that the combinations of this verb with abstract nouns such as *ayuda* 'help', or *experiencia* 'experience' are calqued on English *get*+NP support verb constructions such as *get help*. Crucially, the reported usage patterns of the verb *agarrar* are not registered to the same degree elsewhere. It is highly plausible that local norms affect code-mixing to a similar extent as they influence linguistic transfer.

Of the psycholinguistic factors at work in code-mixing, Muysken (2000: 224–227) pays particular attention to bilingual proficiency and cites several studies probing into the relationship between bilingual proficiency and the rate and type of code-mixing. He concludes that although most studies report positive correlations between bilingual proficiency and the extent of code-mixing, the relation is complex and often mediated by social factors. Among them, he mentions network membership, prestige and generational membership in a migrant community (for details, see below).

The interplay between psycholinguistic aspects of multilingualism and the social nature of linguistic boundaries is evaluated by Law (2014). Although he does not specifically discuss code-mixing, he emphasises the social nature of language separation in the bilingual mind, asserting that "[a] central process in 'language contact' is the merging and separation of different elements of linguistic systems in the bilingual mind, and that separation is intimately social and extremely variable and dynamic, not only from person to person, but within a single individual's own mind over time" (p. 162). When applied to bilingual speech, this view assumes that selection and use of linguistic structures of one language in the discourse framed by another language depends on the individual's ideological awareness of distinction between mixed and unmixed speech as well as the interactional context permitting language mixing. Empirical evidence in support of this position is indirect but encouraging. In two language comprehension experiments, Adamou & Shen (2019) found that processing costs in mixed utterances are reduced if code-mixing is socially acceptable and frequent. In other words, routine use of elements of one language in juxtaposition with elements of another language have ramifications for patterns of activation of linguistic representations in the mental grammar/lexicon and its overall organisation.

All in all, the view that social factors are primary determinants of linguistic patterning in bilingual speech has become widely accepted. Yet, work is still lacking that systematises and evaluates different aspects of social behaviour and power relations that have been linked to various settings in which code-mixing takes place. Existing classifications of the social factors affecting code-mixing include the proposals by Muysken (2000) and Gardner-Chloros (2009). Although both approaches are hardly exhaustive, they are genuinely useful. I will describe and

contrast them below and then complement the outline by a presentation of the individual factors relevant to the studies reported in the subsequent chapters.

1.3.2 Types of social factors

In his attempt to categorise social factors favouring the emergence of code-mixing, Muysken (2000: 222–223) puts the contexts in which code-mixing occurs center stage. Although we find no explicit reference to social factors as such in this approach, contexts in which communication takes place are socially determined and are thus interpretable in terms of social factors. Muysken allocates social contexts to one of three analytical levels: the macro, the meso, and the micro level. The macro level draws on aspects of social and political structure that are characteristic of the bilingual situation on a large scale. The bilingual settings at this level include frontier regions between languages, clusters of multilingual tribal groups with reciprocal bi- and multilingualism, dialect/standard language relations, minority language islands, bilingualism of native elites, colonial and post-colonial settings, and migrant communities. The meso level in Muysken's catalogue describes bilingual communities in terms of their sociolinguistic profiles. The list encompasses aspects such as "the degree of acceptance of code-mixing in the community, attitudes towards bilingualism, structures of linguistic domination, whether it is a transplanted or endogenous bilingual community, the distribution of patterns of language use, including bilingual speech across generations" (p. 222). The interactional setting corresponds to the micro-level. Among the investigated contexts at this level we find peer group and family interactions, institutional interaction in class rooms and in public-authority bodies, marketplace transactions, and exploratory conversations between relative strangers. Muysken emphasises that the proposed list of contexts is incomplete. An apparent consequence of this approach is that the bilingual individual's linguistic behaviour is viewed as a product of contexts located at different analytical levels. In an account of code-mixing, the analyst's task is thus to identify these contexts, to relate them to each other and eventually to the observed mixing patterns.

Muysken further links the various contexts and other social aspects of multilingualism such as, for instance, attitudes towards bilingualism and the existence of strong linguistic norms to specific structural types of code-mixing, namely, insertion, alternation and congruent lexicalisation. As a dominant code-mixing pattern, insertion is common in (post-)colonial settings and in immigrant communities in the first and intermediate generations (see below). The language providing insertions is usually associated with political power, it is the language of

the new country in a situation of immigration and the language of the (former) metropolitan country in (post-)colonial settings. According to Muysken, alternational code-mixing is common among immigrants of the second and following generations, and is also typical in communities characterised by strong norms concerning linguistic behaviour (p. 249). Finally, congruent lexicalisation is facilitated by loose linguistic norms, a balance between the involved languages, and structural parallels.

The reference points of the typology of social factors proposed by Gardner-Chloros (2009: 42–43) are the speaker and the interaction in which they are involved. Factors of the first type include those that are situated beyond the speaker and the specific interactional context; these are factors that "affect all the speakers of the relevant varieties in a particular community, e.g., economic 'market' forces such as those described by Bourdieu (1991), prestige and covert prestige (Labov 1972, Trudgill 1974), power relations, and the associations of each variety with a particular context or way of life (Gal 1979)" (p. 42). Factors of the second type refer to the bilingual individual, also as a member of social sub-groups. These factors pertain to such aspects of social structure as social networks and relationships, linguistic attitudes and language ideologies, perception of self and of others. Also included in this group is proficiency in each variety. According to Gardner-Chloros, the individual's proficiency (competence in her terminology) is essentially "a product of their (reasonably permanent) psycholinguistic make-up", but it has sociolinguistic implications because it is influenced by social factors such as age, network, identity, etc. Although this decision is justified theoretically, in practice, the task of assessing the speaker's proficiency in a variety would inevitably require collecting research data through experimental work (e.g., the application of vocabulary-based proficiency tests) in order to supplement traditional fieldwork data.[5] The final bundle of factors encompasses factors operating within the interactional context. In this case, bilingual speakers employ the juxtaposition of languages, or varieties, in conversation as a contextualisation cue (see e.g., the papers in Auer 1998). With the distinction between code-switching and code-mixing in view, we may assert that the factors at this level pertain to code-switching, i.e. a situation in which speakers perceive and interpret the juxtaposition of codes in a specific instance as a conversational device, but not to code-mixing, which the speakers consider meaningful only in the global sense but not in each individual case.

[5] Another approach to this issue may be illustrated by a study carried out by Muysken et al. (1996), in which the authors refrain from collecting naturally occurring spontaneous conversations as the basis for the analysis and draw on a range of other data instead, including the recording of bilingual parent-child reading sessions.

The classifications of social factors by Muysken (2000: 222–223) and Gardner-Chloros (2009: 42–43) exhibit a considerable overlap; each distinguishes between three types of factors, or three levels of analysis, respectively, and the first and second order categories in both approaches coincide to a great extent. The third order category is conceptualised differently in each case: In Muysken's classification it refers to the interactional setting as the global context in which the interaction takes place, whereas Gardner-Chloros' conversational factors pertain to local contexts in which two, or more, languages are meaningfully juxtaposed in discourse. Another discrepancy between the two approaches lies in the fact that alongside the social dimension of language contact, Muysken introduces its historical dimension, namely the duration of contact, as a fourth level. A more substantial difference is in the object of analysis: While Muysken's concern is with contexts allowing for code-mixing, Gardner-Chloros' focus is on factors. A specific characteristic of Gardner-Chloros' typology is the inclusion of factors pertaining to the organisation of conversation, i.e., the use of codes as a resource for managing interaction. Obviously, this level of analysis relates to switching codes as a bilingual practice, but not to language mixing. As the corpus analysed in the present work contains only few instances in which the speakers employ the juxtaposition of codes as a contextualisation cue, e.g., for quotation, the dominant pattern in the corpus is language mixing. I thus abstain from giving a detailed outline of factors pertaining to code-switching in the present overview, while I acknowledge the possibility that code-switching, and particularly its directionality, may influence patterns of mixing, should both types of code juxtaposition be accepted in a bilingual community. Below, I will showcase factors belonging to the first two types.

Of the factors independent of the bilingual individual, i.e., those operating at the macro level, patterns of sociopolitical dominance appear to be crucial because they influence the direction of insertion in code-mixing (Muysken 2000: 224). Among the factors pertaining to the bilingual individual, of particular importance to the sampled group of speakers is generational membership, which is often related to the speakers' bilingual proficiency.

Differing patterns of sociopolitical dominance in social contacts may produce distinct patterns of linguistic structure. For instance, mixing patterns in bilingual speech may be inverted if an imbalance in power/status is reversed. Such is the case in the speech of Russian Germans after the shift to Russian and following their repatriation to Germany (for details, see Chapter 3) when compared to the speech of Siberia's Russian Germans who have not shifted to Russian. Both groups employ the same strategies of verb integration in bilingual speech: the

verb from the other language is adapted by adding either phonological or morphological material between the stem and the grammatical suffix, (7a, 8a), or it is inserted as a frozen form (7b, 8b). The data from Siberia provided by Blankenhorn (2003) include the following examples:

(7) German (dialect)-Russian (Blankenhorn 2003: 103,93)

a. […] die hen sich so ge-wunder-t, dass ah mir scho
they have.3PL REFL so PTCP-wonder-PTCP COMP also we already
all da *vstreča*-i-t hen, un *provoža*-i-t […]
all here meet-SF-PTCP have.3PL and see_off-SF-PTCP
'…they were so surprised that we also met people here and saw them off…'

b. […] *nu vot tAk vot*, die hen/ die leit hen
PTCL PTCL PTCL PTCL they have.3PL DET.DEF.PL people have3PL
uns *vytjanu-l-i*.
us pull_through-PST-PL
'Well that's just how it is. The people pulled us through.'

My data from Germany show the mirror image. While the Germans sampled in Blankenhorn's 2003 study insert verbs from Russian, the politically dominant language in that case, the speakers recorded for the present research in Germany do the opposite: they insert verbs from German, the "new" majority language, for instance:

(8) Russian-German (own field data; for details, see Chapter 3)

a. prosto referat skaž-u xoč-u
simply presentation[SG.AKK] say.PFV.PRS-1SG want.PRS-1SG
halt-ova-t'.
make-SF-INF
'I'll just say I want to make an oral presentation.'

b. …čë za nedelj-u *passier-t*
what during week-ACC.SG.F happen-PTCP
'…what happened during the week.'

Comparisons of bilingual speech emerging in contexts with opposite dominance relations between the contact languages are scarce, but Muysken (2000: 223–224) provides one example of the case. He compares the patterns of verb integration

in Central American English Creole (based on Herzfeld 1980, 1983) and Mexican American Spanish (based on Pfaff 1979a) to show that the turnabout of the linguistic patterns can reflect the reversal of sociopolitical dominance. These observations lead to a more general conclusion that the structural patterns found in bilingual speech may mirror asymmetric quality of contact due to an imbalance in status between social groups.

The next factor pertinent to the group whose speech is the subject of this book is generational membership. Several studies have provided evidence that structural patterns of code-mixing may be related to generational differences (for a review, see Muysken 2000: 224–227). To illustrate this factor, I will draw on two studies, one conducted in a (post-)colonial setting, the other carried out in the context of migration.

The study by Goria (2021) analyses mixing patterns in Spanish-English bilingual speech of Gibraltar[6] across three generations of speakers: speakers with age over 60 years, speakers between the ages 30 and 60 years, and speakers younger than 30 years. They are labelled as 'elderly', 'adult' and 'young' speakers. The examined corpus of bilingual speech contains instances of insertional mixing as well as as tokens of congruent lexicalisation, but the dominant and most variable mixing pattern is alternation. Specifically, Goria looks at the patterns of clause-peripheral code-mixing and its distribution across the three generations. The results indicate that the factor "speaker generation" is an important predictor of the language of the extra-clausal constituent. For example, the use of Spanish conjunctions and complementisers linking English clauses comprises only six per cent of all cases in the speech of the elderly group, but amounts to 23 and 32 per cent in the speech of the adult and the young group, respectively. The author interprets these generational differences in mixing in terms of an ongoing shift from Spanish to English as a sociopolitically dominant language.

In his analysis of code-mixing in the Turkish-speaking community in The Netherlands, Backus (1996: 387–391) reports a correlation between generational membership in the migrant community and a specific type of code-mixing. He observes that the dominant pattern in the speech of first-generation immigrants is the insertion of Dutch content words and their morphosyntactic integration into Turkish (I have referred to this pattern as minimal insertion above). The vernacular of the intermediate generation speakers – these immigrants arrived in the Netherlands when they were between 5 and 12 years old – has as much insertion as alternation, and insertional mixing is highly varied as it includes both minimal and maximal insertions, the latter being fully-fledged Dutch constituents in

[6]The political status of Gibraltar is disputed. The description of this British dependent territory as a colony has been criticised by the UK authorities.

otherwise Turkish sentences. Finally, the speech of the second-generation immigrants is characterised by alternational mixing. An overview of these findings is given in Table 1.1. As evident from the table, the first and the intermediate generation share the same propensity to use Dutch words in otherwise Turkish sentences, whereas the second generation also uses Turkish words in Dutch sentences. Furthermore, Backus links the generation-specific mixing patterns to the generations' language choice preferences, with a gradual shift from Turkish to Dutch. In his 2006 publication, he cites several studies documenting the same pattern of intergenerational variation in other Turkish-speaking communities across Europe and concludes that this pattern may be very general.

Table 1.1: Distribution of main types of code-mixing, and base language in code-mixing across first, intermediate, and second generations in Turkish-Dutch code-mixing data (adapted from Backus 2006: 702).

Generation	Type of code-mixing			Base language in code-mixing		
	insertion	both	alternation	Turkish	both	Dutch
First	✓			✓		
Intermediate		✓		✓		
Second			✓		✓	

In view of the Russian-German community in Germany, whose mixing patterns are reported in the subsequent chapters of this book, we can expect that the pattern of intergenerational variation in their speech will largely coincide with the pattern reported by Backus. It is important to emphasise however that the described Turkish community and the Russian-German community in Germany differ in their official status. Unlike the Turkish immigrants in the Netherlands, Germans from the Soviet Union and its successor states are repatriates to Germany. Yet, as will be demonstrated below (see Chapter 3), the patterns of their language choice preference, considering their shift to Russian in the 1970s and 1980s, are comparable with those of immigrants. Against this background, and in view of the goal to explore patterns of insertional code-mixing, it is reasonable to assume that the speech of intermediate-generation Russian-German repatriates will provide diverse loci of variation in insertional code-mixing, including minimal and maximal insertions in the same language, and will hence suit the envisaged goal.

In addition to the sociolinguistic perspective, patterns of insertional code-mixing have been approached along the lines of structural analysis (e.g., Halmari 1997,

Boumans 1998, Verschik 2008), but one of the most elaborate structural approaches to insertional code-mixing is the Matrix Language Frame model (cf. Muysken 1997: 363). To this, I turn next.

1.4 The Matrix Language Frame model and its extensions

The Matrix Language Frame model was proposed and has been further developed by Carol Myers-Scotton (1993, 2002). Since its first formulation, this model has become the object of a heated controversy. Some scholars have successfully tested the model on various language pairs (see Haust 1995, for Mandinka, Wolof-English; Backus 1996, for Turkish-Dutch; Hlavac 2003, for Croatian-English; Amuzu 2010, for Ewe-English) or applied it to other language contact phenomena, such as creole formation, long-standing language contact, child bilingualism, and adult second language acquisition (cf. Myers-Scotton & Jake 2000). Other researchers, however, have criticised some of the model's assumptions (Meechan & Poplack 1995; Halmari 1997; Bentahila & Davies 1998; Boumans 1998; Auer & Muhamedova 2005; Muhamedowa 2006; Chang 2009; Zabrodskaja 2009). From the beginning, Myers-Scotton has developed and modified the original model further. I will therefore discuss the proposed models in the chronological order, i.e., the Matrix Language Frame model first and then its extensions: the Abstract Level model and the 4-M model.

1.4.1 The Matrix Language Frame model

As outlined above, in the case of insertional code-mixing an asymmetry is observed between the languages involved because only one language is responsible for providing the frame for the sentence. This language is called the *matrix language* (ML), whereas the other language is referred to as the *embedded language* (EL). According to Myers-Scotton (1993), this terminology goes back to Joshi (1985). His approach to code-switching builds on the observation that speakers and hearers are capable of identifying the language "the mixed sentence is 'coming from' " (Joshi 1985, quoted from Myers-Scotton 1993: 35). In contrast to the perceptual view taken, both Joshi and Myers-Scotton adopt a structural perspective on the matrix language-embedded language asymmetry (cf. Myers-Scotton 1993: 35–37). As this asymmetry is the crux of the outlined model, the question of determining the matrix language is essential to this approach.

According to Myers-Scotton (1993: 68), more morphemes come from the matrix language than from the embedded language, where morphemes are counted

in a discourse sample and cultural borrowings are excluded from the counts. This criterion for matrix language identification is not unproblematic. For example, Muhamedowa (2006: 19) claims that her bilingual data contain lengthy monolingual passages which intervene with instances of code-mixing and code-switching (Haust 1995: 102; Boumans 1998: 154; and Hlavac 2003: 196, argue a similar point). That is, the discourse-dominant language may not coincide with the matrix language of a clause (cf. Auer 1988). In this circumstance, adequate morpheme counts are infeasible. In her later work, Myers-Scotton (1995: 237) mentions two further criteria for matrix language definition: First, the matrix language is the unmarked choice in bilingual communication, one of the functions of which is solidarity building. Second, speakers' self-reports on which language is the matrix language are a reliable indication of the matrix language. In spite of the two suggested criteria, Myers-Scotton's analysis of bilingual sentences is virtually always based on solely structural criteria, related to another important premise of the matrix language frame model, i.e., the hierarchy between content and system morphemes.

Myers-Scotton differentiates between content and system morphemes. Prototypical system morphemes subsume function words and inflectional affixes, whereas prototypical content morphemes include verb and noun stems. A more detailed morpheme classification is based on such discreet categories as [±Quantification], [±θ-role assigner], [±θ-role receiver] and their role at the discourse level. System morphemes are characterised as [+Quantifier]. Syntactic categories with the feature [–Quantifier] are further classified with regard to their potential to assign or receive theta-roles. For example, nouns, pronouns, adjectives and adverbs derived from adjectives are theta-role receivers and thus content morphemes, but dummy pronominals *there* and *it* are not theta-role receivers and are therefore system morphemes. Because some syntactic categories assign and receive theta-roles, depending on the specific items that belong to these categories, they can function as either content or system morphemes. Prepositions are one of such categories; for instance, the English preposition *in* and its French counterpart *dans* are content morphemes because they assign both theta-roles and case, whereas the English *of* and the French *de*, marking genitive objects, are system morphemes as they assign only case (cf. Myers-Scotton 1993: 98–102). Among verbs, which are typical theta-role assigners, the copula and the English *do* in *do*-constructions are considered system morphemes. The system morpheme versus content morpheme hierarchy and the matrix language versus embedded language hierarchy are related because system morphemes participate in constituent frame formation and can be controlled by only one language

at one point in time (Myers-Scotton 1995: 235). The two hierarchies are at the core of the Matrix Language Frame model.

The model draws builds on several hypotheses. The *matrix language hypothesis* seeks to explain the structure of mixed, or ML + EL, constituents in code-mixing, it says "[...] the ML provides the morphosyntactic frame of ML + EL constituents" (Myers-Scotton 1993: 82). This hypothesis determines the Morpheme Order and the System Morpheme Principles, namely:

The Morpheme Order Principle: Morphemes in mixed constituents are ordered according to the ML.

The System Morpheme Principle: Syntactically relevant system morphemes in mixed constituents come from the ML. (cf. Myers-Scotton 1995: 239)

A system morpheme is syntactically relevant if it is involved in agreement relations external to its head constituent. In (9), for instance, the matrix-language (i.e., Croatian) system morpheme *-u* expressing the accusative case is considered syntactically relevant because it is required by the head of the prepositional phrase and not the noun as the head of the noun phrase.

(9) Croatian-English (Hlavac 2003: 115)
... sad ć-e ić u Hrvatsk-u za ov-u
now FUT-3SG go.INF to Croatia-ACC.SG.F for this-ACC.SG.F
treć-u *term*-u [...]
third-ACC.SG.F -ACC.SG.F
'...he will be going to Croatia for this third term [...]'

According to Myers-Scotton (1993: 110), the System Morpheme Principle is maintained if any of three possible strategies is employed: The first, prototypical, case is the occurrence of mixed constituents, with system morphemes coming from the matrix language. System morphemes may also come from both languages simultaneously, so that *double morphology* emerges. This is the second strategy. As such, morphological doublets are only possible when the system morphemes of the embedded language do not have relations external to their heads. As a third strategy, EL content morphemes may appear as *bare forms*. Myers-Scotton (1993: 112) asserts that bare forms are produced if an embedded-language system morpheme and the corresponding matrix-language system morpheme are incongruent. The outlined constraints are integrated into a production model, which is largely based on the work by Levelt (1989, quoted in Myers-Scotton 1993).

The language production process according to Myers-Scotton (1993: 116–119) involves four steps. As the first step, in order to meet the requirements of the communicative situation, speakers take, mainly unconsciously, intentional and socio-pragmatic decisions at the conceptual level. Steps two and three concern the functional level. Step two involves building the frame into which content morphemes are inserted. Specifically, matrix-language lemmas are selected from the speaker's mental lexicon in accordance with her conceptual specifications. At step three, the selected lemmas send information to the "formulator", or processing centre, which regulates grammatical encoding procedures. These operations are considered responsible for controlling matrix-language specifications for system morphemes. Myers-Scotton remarks that concrete morphemes may yet be actualised at a later stage, "nearer the surface" (p. 118). She further asserts that the essential procedures carried out in the formulator can be covered by the matrix language hypothesis and both the Morpheme Order and the System Morpheme principles. Once the frame is built, lexemes attached to lemmas are realised, and a unified structure is produced. This final step in the production process concerns the positional level. This implies that information about the surface structure is activated. The only structures that violate this model are *embedded-language islands*, defined as well-formed embedded-language constituents occurring in a matrix language clause. That is, embedded-language content morphemes appear in this case together with embedded-language system morphemes in a clause framed by the matrix language. Myers-Scotton (1993: 119) claims that embedded-language islands are produced "when ML procedures are entirely inhibited by EL procedures". The model distinguishes between obligatory and optional embedded-language islands. Obligatory embedded-language islands are triggered by incongruent morphosyntax. (The idea of obligatory embedded-language islands is elaborated further as the EL Island Trigger hypothesis, see below.) Moreover, not only the matrix language, but also the embedded language can be subject to inhibition. The inhibition of the embedded language is crucial to the System Morpheme Principle as well as the Blocking Hypothesis.

Whilst in the case of the System Morpheme Principle, a filter in the formulator prohibits embedded-language system morphemes, in the case of the Blocking Hypothesis, inhibition applies to embedded-language content morphemes. The *Blocking Hypothesis* postulates that "[i]n ML + EL constituents, a blocking filter blocks any EL content morpheme which is not congruent with the ML with respect to three levels of abstraction regarding subcategorization" (Myers-Scotton 1993: 120). The following kinds of incongruence are considered relevant: first, a mismatch in the morpheme status, i.e., the matrix language uses a system morpheme to code a given grammatical category, whereas the embedded

language employs a content morpheme for the same purpose; second, incongruence between embedded-language and matrix language content morphemes in respect of thematic role assignment; third, a mismatch between embedded-language and matrix language content morphemes regarding their discourse or pragmatic functions. Although the Blocking Hypothesis predicts the blocking of any incongruent content morpheme, the presented analysis covers only the cases of incongruent pronouns and prepositions. Pronouns, for example, may be realised as content morphemes in one language and as system morphemes, i.e., as clitics and dummy pronominals, in the other. Myers-Scotton (1993: 126–128) argues that such lack of congruency in the morpheme status is the reason why pronouns analysed as content morphemes do not occur in mixed constituents. However, the discussion of the Blocking Hypothesis is silent on what content morphemes with divergent patterns of thematic role assignment or different discourse or pragmatic functions could be blocked.

The final hypothesis underlying the matrix language frame model is the *EL Island Trigger Hypothesis*. It predicts when obligatory embedded-language islands must occur: "Activating any EL lemma or accessing by error any EL morpheme not licensed under the ML or Blocking Hypotheses triggers the processor to inhibit all ML accessing procedures and complete the current constituent as an EL island" (Myers-Scotton 1993: 139). The examples provided as evidence of the EL Island Trigger Hypothesis include insertions of English noun phrases modified by demonstrative or possessive pronouns, i.e., system morphemes. The following example illustrates the case of possessive pronouns:

(10) Swahili-English (Myers-Scotton 1993: 141)
Tu-na-m-let-e-a *our brother* wa Thika
1PL-PROG-him-take-APPL-INDIC of Thika
'We are taking [it] to our brother of Thika.'

Myers-Scotton regards the modifier *our* in (10) as a trigger for the corresponding island because the order of this modifier and its head in English, the embedded language, is at odds with the order of the corresponding constituents in Swahili, the matrix language. According to the production model assumed as well as the System Morpheme Principle, the lemmas which correspond to system morphemes come from the matrix language as early as step two of the production process, i.e., in the formulator, where the frame is built. Consequently, the only possible explanation for the activation of the EL lemma supporting the system morpheme *our* is by error. Crucially, it is owing to this morpheme that the whole EL island is produced. The analysis of a similar case in Myers-Scotton &

Jake (1995) assumes an alternative scenario: at first, the lemma corresponding to the head of the nominal phrase is activated in the mental lexicon and then "[t]his lemma activates morphosyntactic procedures in the formulator, such that ML procedures are inhibited for the maximal category projection (here, NP) associated with that lemma. The result is an EL island" (p. 995). In this case, the Embedded-Language Island Trigger hypothesis is dismissed. As such, the explanation by triggering would be more straightforward provided that linearity is possible at an abstract level. I assume that an approach to language production taking into consideration the linear character of speech could arrive at a more realistic account of instances like (10) than a purely top-down production model.

A specific type of embedded-language islands, according to Myers-Scotton (1993: 142, 144) include set phrases. Unfortunately, she does not specify the diagnostic features of a set phrase.

The Embedded-Language Island Trigger hypothesis appears to be problematic inasmuch as it considers the access to any EL morpheme as erroneous. The regular occurrence of EL islands in such language pairs as English and Spanish (cf. Poplack 1980b), or Dutch and French (cf. Treffers-Daller 1994), could hardly be only due to erroneous access, the incongruence of the embedded-language material with the matrix language specifications (Myers-Scotton 1995: 250), or the formulaic character of the morpheme string involved. After all, in the model version outlined in Myers-Scotton (1995), the Embedded-Language Island Trigger hypothesis is reformulated with no mention of erroneous access (p. 249), and Myers-Scotton & Jake (1995) discuss embedded-language islands without a reference to the aforementioned hypothesis. Myers-Scotton (1993: 137) acknowledges that embedded-language islands are "the potential Achilles' heel of the MLF model"; therefore, her later work (2001; 2002), including joint research with Jake (1995), focuses on mechanisms constraining embedded-language islands. The ideas that were initially presented in their 1995 paper "Matching lemmas" laid the ground for the Abstract Level model.

1.4.2 The Abstract Level model

As indicated above, the work by Myers-Scotton & Jake (1995) examines structures that result from either the Blocking Hypothesis or the Embedded-Language Trigger Hypothesis of the matrix language frame model, i.e., the so-called compromise strategies: embedded-language islands, bare forms and *do*-constructions. First and foremost, the Abstract Level model is aimed at explaining these phenomena. Second, it is claimed to shed light on the structure of entries in the mental lexicon.

Myers-Scotton & Jake (1995) proceed from the premise that abstract grammatical structure contained in lemmas underlying lexical items is distributed at three levels: (i) the level of lexical-conceptual structure (semantic/pragmatic features), (ii) the level of predicate-argument structure, and (iii) the level of morphological realisation patterns. According to the production model assumed (Myers-Scotton 2002: 23–25, 76–78), these levels are activated in the following way. At first, the speaker's intentions select a language-specific semantic-pragmatic feature bundle at the conceptual level, which in its turn elects a lemma underlying a content morpheme. If a lemma is activated that supports an embedded-language content morpheme, Myers-Scotton & Jake (1995) hypothesise that this lemma is matched for congruence against a matrix language counterpart lemma at every level of abstract grammatical structure mentioned above. As the matrix language may lack a counterpart for the given embedded-language lemma, congruence checking is assumed possible against Lexical Knowledge ("generalized but specific to the matrix language", Myers-Scotton 2002: 97). The idea that lemmas are checked for congruence in bilingual production is another premise of the Abstract Level model. In case of sufficient congruence between the embedded-language lemma and the matrix language counterpart at all three levels, the embedded-language content morpheme will surface as fully integrated into the matrix language frame, that is, a mixed constituent will be produced. If there is a lack of congruence at one of the levels of abstract grammatical structure, a bare form or an embedded-language island will emerge. It is necessary to note that Myers-Scotton & Jake (1995) no longer maintain the division of embedded-language islands into obligatory and optional, which was present in Myers-Scotton (1993). In essence, the Abstract Level model ascribes the use of compromise strategies to the lack of congruence in one of the levels of abstract grammatical structure. Below, I will discuss mismatches between lemmas at all the three levels of abstract grammatical structure.

Following the model, a compromise strategy may be employed when there are differences in lemmas' lexical-conceptual structure, or their semantic and pragmatic features. We can thus assume, for instance, that the embedded-language island in (11) is produced because there is not sufficient congruence between the lemmas in the lexical-conceptual structure.

(11) Kazakh-Russian (Muhamedowa 2006: 41–42)
Mustafa ata-m on bes žas-ï-nda *protiv*
Mustafa grandpa-POSS.1SG ten five year-POSS3SG-LOC against

basmačestv-a ket-ken.
basmachi.movement-GEN.SG.N go-PRF3SG
'My grandpa Mustafa fought against the basmachi movement at fifteen.'

According to Muhamedowa (2006), Kazakh has no counterpart for the Russian preposition *protiv* 'against', and the speaker's use of this preposition triggers the production of the embedded-language island. Following the scenario envisaged by Myers-Scotton & Jake (1995: 995), the choice to use Russian should be made at the conceptual level, although with consequences for the morphological realisation pattern. This is because Kazakh and Russian take different patterns to encode spatiotemporal relations: Kazakh postpositions contrast with Russian prepositions. The observed incongruence at the conceptual level is argued to account for the occurrence of the embedded-language island.

Similarly, the lack of congruence at the functional level, i.e., in either predicate-argument structure, or morphological realisation patterns, is regarded as a motivation for the appearance of embedded-language islands and bare forms. The example below illustrates the occurrence of a bare form, which is hypothesised to be the result of a mismatch in the predicate-argument structure.

(12) Croatian-English (Hlavac 2003: 213)
... to je možda najbolj-e što sam ja,
that be.PRS.3SG perhaps best-NOM.SG.N REL.ACC be.PRS.1SG 1SG.NOM
što ja *remember-∅ yeah.*
REL.ACC 1SG.NOM -∅ PTCL
'...that is perhaps the best thing that I have, that I remember, yeah.'

Here, the English verb *remember* and its Croatian counterpart *sjećati se* exhibit divergent argument structures: the Croatian reflexive verb realises the theme as a genitive NP, whereas the English *remember* assigns the theme the non-nominative case. The theme in (12) is an accusative NP, expressed by the relative pronoun *što* 'that'; yet, it is the genitive form of this pronoun *čega* that concurs with the required configuration. The use of the bare verb *remember* seems to be a compromise strategy that resolves the apparent problem of incongruence.[7] An alternative account of instances like (12) could be one that allows for linearity at an abstract level. In this case, we could assume that the bare form is triggered

[7]Hlavac (2003: 213) analyses the pronoun *što* as a nominative form. As the latter exhibits form syncretism, it can be interpreted as expressing either the nominative or the accusative case. Here, it seems more appropriate to assume the accusative case because the nominative NP is also realised in the clause by the personal pronoun *ja*.

by the preceding structure, incongruent with the argument configuration of the Croatian verb but with a higher degree of congruence with the configuration of the English verb. In other words, the speaker, without finding *le mot juste* in Croatian, has to make a compromise and switch to English.

With regard to the third level of abstract grammatical structure, the authors claim that insufficient congruence in patterns of morphological realisation employed by the languages should bring about the occurrence of an embedded-language island. This seems to be the case in the following example:

(13) Finnish-English (Lehtinen 1966: 226)
...ja sitte eh *in the afternoon* isä vai minä men-i...
and then HES father[NOM.SG] or 1SG.NOM go-PST.3SG
'...and then in the afternoon father or I went...'

We can hypothesise that the embedded-language island *in the afternoon* in (13) arises because the morphological patterns that express spatiotemporal location in Finnish and English diverge: this meaning is expressed by prepositions in English, but by postpositions in Finnish. The apparent conflict in the patterns is resolved in that the embedded language is produced.

Although the proposed model seems well elaborate to explain the emergence of mixed constituents as well as many instances of bare forms and embedded-languages islands in code-mixing with various languages, some of its aspects may be considered problematic. First, the model seems capable of predicting the use of compromise strategies, but it remains inexplicit about when a particular strategy is followed, i.e., when a mismatch in congruence will result in an embedded-language island, and not in a bare form. Second, as the use of compromise strategies is explained *ex-post*, it is unclear which level of abstract linguistic structure matters when. For instance, we can assume for examples (11) and (13) that the mismatch relevant for the occurrence of embedded-language islands may be at the level of morphological realisation patterns as well as at the level of lexical-conceptual structure. Finally, embedded-language islands and bare forms may occur even when the features of lemmas supporting the respective morphemes match, or when some features of the morphosyntactic structure are shared by both languages. For example, Treffers-Daller (1994) provides instances of embedded-language islands structured as PPs, such as (14), which can be explained by neither incongruence in lexical-conceptual and predicate-argument structure, nor by incongruent morphological realisation patterns.

(14) Dutch-French (Treffers-Daller 1994: 208)

Wat gaa-t ge do-en *chez ce-s* *pauvre-s vieux*?
what go-PRS.2SG 2SG do-INF at DEM-PL poor-PL old[PL]
'What are you going to do in the house of these poor old people?'

In (14), the French PP *chez ces pauvres vieux* 'at these poor old people's (place)' is an embedded-language island. Apparently, the island and its Dutch equivalent *bij die arme oudjes* do not demonstrate incongruence at any level of abstract grammatical structure: The concept 'old people' exists in both languages and is encoded in a similar way, i.e., by nouns formed by conversion from adjectives (i.e., Dutch *oud* and French *vieux*). With regard to morphological realisation patterns, both French and Dutch use prepositions and pre-nominal determiners, and in the present case they both rely on overt plural markers (the demonstratives are plural forms: *ces* and *die*, and Dutch also marks plural on the noun *oudje* with the suffix *-s* and on the adjective *arm* with the suffix *-e*). Therefore, it is necessary to examine other factors that may influence the emergence of embedded-language islands. As I will show in the subsequent chapters, the frequency with which words co-occur in the embedded language can explain occurrences of embedded-language islands more effectively.

1.4.3 The 4-Morpheme model

A more recent development in the work by Myers-Scotton is the 4-Morpheme model (2001; 2002: 73–82; for a recent overview, see Myers-Scotton & Jake 2016). As was the Abstract Level model, this is a submodel of the matrix language model. The 4-M model is motivated by the need to provide a more precise account of "certain types of congruence problems [that] arise in codeswitching between certain language pairs" (Myers-Scotton 2001: 42) and in specific contact phenomena, such as creole formation, language attrition and second language acquisition (which abound with counter-examples to the original matrix language frame model, proposed in Myers-Scotton 1993). The 4-M model introduces a subdivision of system morphemes based on their distribution in code-mixing. The classification relies on the premise that system morphemes are activated at two different levels. Myers-Scotton claims that the proposed morpheme classification is valid for language production in general.

By and large, the model distinguishes between four types of morphemes: content morphemes, early system morphemes, bridge late system morphemes and outside late system morphemes. The classification builds on three abstract oppositions. The first opposition concerns the conceptual activation of lemmas underlying morphemes. The hypothesis here is that some of the lemmas have

a more direct connection to speaker's intentions than others. Under this opposition, represented as [±conceptually activated], content morphemes and early system morphemes have the feature [+conceptually activated], whereas the other two morpheme types, i.e., late system morphemes, have the feature [–conceptually activated]. Myers-Scotton (2002) claims that speakers' intentions activate language-specific semantic/pragmatic feature bundles, which select lemmas in the mental lexicon supporting content morphemes. Such a [+conceptually activated] element has semantic content. Lemmas that underlie content morphemes are hypothesised to be elected directly. Directly elected lemmas may in their turn activate other lemmas at the same level. These other lemmas are thus elected indirectly and underlie early system morphemes. In other words, morphemes which are activated at the level of the mental lexicon (and have the feature [+conceptually activated]) are subdivided into content morphemes and early system morphemes depending on whether they are elected directly or not. This opposition is formalised as [±thematic role]. *Content morphemes* have the feature [+thematic role], and *early system morphemes* are [–thematic role]. The prototypical content morphemes are nouns and most verbs. Early system morphemes express the *phi*-features of person, number and gender and therefore include determiners and plural morphemes.

According to Myers-Scotton (2002), "Together, the lemmas underlying content morphemes and early system morphemes send directions to the Formulator to build larger linguistic units" (p. 77). This is how lemmas underlying late system morphemes are activated. Depending on the type of grammatical information carried by the activated lemma supporting a late system morpheme, Myers-Scotton distinguishes between bridge late system morphemes and outside late system morphemes. *Bridge late system morphemes* integrate content morphemes and early system morphemes into larger constituents, whereas *outside late system morphemes* provide for coindexical relations across maximal projections. Formally, this opposition is represented like this: [±refers to grammatical information outside of Maximal Projection of Head]. While outside late system morphemes refer to grammatical information outside of Maximal Projection of Head, bridge late system morphemes do not refer to grammatical information outside of the maximal projection of their heads. Myers-Scotton (2002: 75) illustrates 'bridges' with English possessives *of* and *'s* and the French *de*, as in *beaucoup de gens* 'a lot of people'. "Outsiders" are exemplified by case affixes and morphemes expressing subject-verb agreement.

In a nutshell, the model differentiates between the following four types of morphemes on the basis of the proposed abstract oppositions:

- content morphemes: [+conceptually activated], [+thematic role];

- early system morphemes: [+conceptually activated], [–thematic role];
- bridge late system morphemes: [–conceptually activated], [–grammati-cal information outside of Maximal Projection of Head]
- outside late system morphemes: [–conceptually activated], [+grammatical information outside of Maximal Projection of Head]

Myers-Scotton introduces an important amendment concerning the order of morpheme activation in the case of fusion of grammatical features. In some languages, grammatical features may be fused under the same morpheme. For example, German determiners express gender, number and case simultaneously. Russian and other Slavic languages abound with such portmanteau morphemes: inflections within the Russian nominal declensional system encode number, gender and case, and also the category of animateness. Myers-Scotton (2002) calls such portmanteau morphemes multimorphemic elements (p. 305), or multimorphemic lexemes (p. 81). The hypothesis regarding multimorphemic elements is this (p. 305):

> In multimorphemic elements (consisting of two or more system morphemes and including a late system morpheme), the late system morpheme takes precedence. This means that the entire element shows distribution patterns as if it were a late system morpheme. This is the 'pull down' or 'drag down' principle.

In other words, a late system morpheme such as one expressing case pulls down the portmanteau morpheme which also expresses gender and number. As such, Russian nominal inflections should be classified as outside late system morphemes because morphemes expressing case are the last to be activated of the discussed morpheme types.

However, structures examined in Muhamedowa (2006) seem to provide evidence against this hypothesis. The following is an example of Kazakh Russian discussed by the author:

(15) Kazakh Russian (Muhamedowa 2006: 92)
edinstvenn-yj ja by-l-a semejn-yj i
sole-NOM.SG.M 1SG.NOM be-PST-SG.F with.family-NOM.SG.M and
eščë s det'-mi
additionally with children-INST.PL
'I was the only one with a family and what is more with children.'

Before presenting the argument, I will mention the relevant features of the Russian verbal phrase. First, in the past tense subject-verb agreement involves the inflectional values of number and gender. Second, predicative adjectives also agree with the subject in inflectional values and are additionally assigned either the nominative, or the instrumental case. Two standard-Russian versions of (15) are *Edinstvennaja ja byla semejnaja i eščë s det'mi* or *Edinstvennoj ja byla semejnoj i eščë s det'mi*. In both versions, we can observe agreement in number and gender between the verb *byla* and the predicative adjectives *edinstvennaja*, or *edinstvennoj*, and *semejnaja*, or *semejnoj*, respectively. Following the hypothesis given above, the nominal inflections here are late outside system morphemes because they refer to grammatical information outside of the maximal projection of their heads. Yet, the verbal phrase in (15) lacks the necessary agreement: the inflections of the predicative adjectives correspond to the nominative singular masculine, and not the nominative (or instrumental) singular feminine. We can thus state that agreement between the verb and the predicate adjectives is only in number and case, and is thus partial. Therefore, (15) could be considered a violation to the aforementioned hypothesis and further to the suggested classification according to abstract oppositions.[8]

Despite the identified problems with the matrix language frame model, it is necessary to emphasise the importance of the seminal work by Myers-Scotton and associates, which opened up the field towards psycholinguistic theorising by linking code-mixing phenomena to existing language production models.

The analysis of insertional code-mixing reported in the remainder of this book will build on the matrix language-embedded-language dichotomy, without adopting the specific assumptions of the Matrix Language Frame model as well as the underlying production model. I will thus distinguish between mixed constituents, or minimal insertions, and embedded-language islands, or maximal, multimorphemic insertions. My principal focus will be on the nature of embedded-language islands, whose pervasive use in bilingual speech has been explained by their their special status in the bilingual mental lexicon/grammar. Multimorphemic insertions have been argued to correspond to units in the mental lexicon/grammar, which are comparable with entries of mono-morphemic words. According to this account, multimorphemic sequences with unit status are perceived and produced online as mono-morphemic words. To this approach I turn

[8] According to Muhamedowa (2006: 92), the speaker in (15) is Russian-dominant, she acquired Kazakh late, as a second language. The speaker produces well-formed predicative adjectives in the same conversation. However, numerous instances of gender neutral Russian adjectives are common in Kazak-Russian code-mixing, especially as modifiers in Russian noun phrases inserted in Kazakh (cf. Muhamedowa 2006: 77–93).

next.

1.5 Multimorphemic units in insertional code-mixing

An account of embedded-language islands as multimorphemic units in the lexicon is proposed by Ad Backus (1996, 1999a, 2003). His analysis of code-mixing proceeds from the matrix language-embedded-language asymmetry, crucial for the MLF model, but it differs from that model substantially. While acknowledging the semantic basis for distinguishing between content and system morphemes, the MLF model, he criticises, excludes semantic factors; he contends that "they are not presented as such" (Backus 1996: 115) in the MLF model. By integrating semantic and structural factors, Backus (1999a, 2003) manages to shed light on the nature of embedded-language islands. The approach presented in his 1996 monograph and subsequent publications (1999a, 2003) draws on premises from Langacker's Cognitive Grammar (1987, 1991) and Goldberg's Construction Grammar (1995). In this section, I will first outline the premises relevant for the current overview and then introduce the features of multimorphemic units. Finally, I will discuss the unit hypothesis and the "conceptual unit" hypothesis put forth by Backus.

One premise postulates that the grammar of a language and its lexicon are not separate modules but form a continuum. The lexicon is considered an inventory of symbolic units, i.e., form-meaning pairings. Some of these units are abstract schemas, which are broadly equivalent to morphosyntactic rules in other frameworks, while others are specific (lexical) items. Schemas, or constructions, sanction novel combinations of lexical items. These two types of units form the poles of the lexicon-grammar continuum.

Another premise concerns the nature of lexical units: "lexical units can be of any length and complexity" (Backus 2003: 84). Therefore, whether morpheme-sized or multimorphemic, all units are supposed to be accessed directly in production and not composed on-line, even if multimorphemic units can be decomposed into their constituent parts (cf. Backus 1996: 129). Support for this view comes from psycholinguistic experimental research, notably from behavioural studies on processing of multimorphemic units: noun plurals and idiomatic expressions. Whilst Baayen et al. (1997) provide evidence that Dutch speakers store high-frequency regular noun plurals (see Chapter 2, for further details), Libben & Titone (2008), and Tabossi et al. (2008) underpin the role of decomposition in idiom comprehension. The study by Libben & Titone (2008) confirms that decomposition plays only a marginal role in idiom comprehension, and Tabossi et

al. (2008) show that semantic compositionality of idiomatic expressions does not influence their syntactic flexibility.

From these premises it follows that a string of morphemes or words, sanctioned by a schema, gains the status of a lexical item once it is entrenched, i.e., when it receives an allocated mental representation. This scenario asks for a definition of multimorphemic lexical units. Backus (2003) defines *units* as "any recurrent combinations of two or more morphemes that together exhibit idiomatic meaning" (p. 90).[9] Further lexical units include high-frequency composite forms as well as forms with irregular morphosyntax. That is, in order to qualify as units, multimorphemic elements have to either (1) demonstrate irregular morphosyntax (e.g., the past *caught* and the plural *children*), (2) express non-compositional meaning (e.g., the collocation *play tennis*, the idiom *this is a piece of cake*, meaning 'X can be accomplished with ease', and the discourse marker *the thing is*), or (3) be of high frequency (e.g., the regular past forms *worked* and *helped*). Moreover, multimorphemic units can be discontinuous and have open slots. For instance, the aforementioned idiom *this is a piece of cake* allows for variability in the first two elements: the first element is an open slot filled by a nominal phrase, and the second element is any form of the copula *be*. This idiom can hence be represented formally like this: [X BE *a piece of cake*]. The open slots accompanying discourse markers, such as *the thing is*, are filled by clauses. As open slots can possibly be instantiated by an unlimited range of lexical items, only their frozen elements are used as a diagnostic feature of units.

In his work on Dutch insertions in Turkish sentences, Backus (1999a) claims that maximal insertions consisting of co-occurring embedded-language, i.e., Dutch, morphemes, frequently correspond to Dutch multimorphemic lexical units. Backus (2003) articulates this idea as the *unit hypothesis*, which stipulates that "[e]very multimorphemic EL [= embedded-language] insertion is a unit, inserted into a ML clausal frame" (p. 91). This hypothesis is plausible; bilingual corpora abound in instances of lexical-unit insertion, for instance:

(16) Hindi-English (Bhatt 1997: 228)
kəl hi *Israel government* ne kəha ki Asaad$_i$ *peace talks* ke
yesterday PTCL ERG said that
prəṯi *serious* nəhĩ: hai aur pro$_i$ *political games* khel rəha hai
toward not is and play PROG is
'Only yesterday the Israel government said that Asaad is not serious about peace talks and that he is (instead) playing political games.'

[9] The term "idiomatic meaning" is not restricted to idioms alone, it should be understood broadly, as an equivalent to 'non-compositional meaning' (cf. Backus 2003: 86).

Here, the word strings *Israel government*, *peace talks* and *political games* are lexical units inserted into Hindi clauses, the first two present compound nouns and the third is a collocation. The compound *Israel government* is one of the few instantiations of the pattern [COUNTRY *government*]. This pattern is productive only to some degree: in *The Corpus of Contemporary American English (COCA)*[10], only 16 lexemes appear as realisations of the first pattern element[11], with the exception of initialisms such as *US* and *UK* and plural forms such as *United States* and *Seychelles*. The pattern is restricted to nouns denoting country names whose corresponding adjectives end in *-i* or *-ese*. Whereas the schemas [COUNTRY-ADJ *government*] and [COUNTRY*'s government*] – illustrated by word strings *Israeli government* and *Israel's government* – are the common, productive ways to refer to 'the government of a country', the pattern used in (16) can be applied only to a limited set of lexemes, resulting in a small, presumably unproductive set of word combinations (cf. Bauer 2001: 74). It is thus possible to analyse the word sequence *Israel government* as a specific lexical item. The word string *peace talks* is the other inserted compound noun in (16), it is a *plurale tantum* containing the suffix *-s* and thus a frozen morphological form. This feature enables one to consider this compound noun as a lexical unit. Finally, the collocation *political games* is distinguished by its idiomatic meaning and also qualifies as a lexical unit.

The three instances of multiword unit insertion in (16) coincide with syntactic phrases. However, multimorphemic units in the embedded language do not have to fit the constituent structure neatly. Such units can encompass more than one syntactic phrase. Backus (2003) demonstrates this point by providing numerous examples of the predicate-complement construction, which is lexically realized as recurrent collocations of the embedded language. This type of insertional code-mixing may be found in other bilingual corpora, for instance:

(17) Moroccan Arabic-Dutch (Boumans 1998: 245)
n-dir-u *pauze houd-en*?
1-do-PL break take-INF
'Shall we take a break?'

(18) Moroccan Arabic-French (Bentahila & Davies 1983: 315)

[10]Unfortunately, I could not use *The international corpus of English: Indian corpus* because of its modest size; the word sequence *Israel government* is not attested in it.

[11]These nouns are as follows: *Hong Kong*, *Singapore*, *Pakistan*, *Taiwan*, *New Zealand*, *Iraq*, *Kuwait*, *Sudan*, *Afghanistan*, *Bangladesh*, *Israel*, *Botswana*, *Myanmar*, *Palau*, *Flanders*, *Cameroon*.

tajbqa j-*confronter* *ces* *idées*
keep.3SG IMPRF-oppose these ideas
'He keeps opposing these ideas.'

(19) Tamil-English (Sankoff et al. 1990: 80)
anta *car*-ei *drive* paNNanum
that -ACC do.must
'We must drive that car.'

The multimorphemic insertions in these examples can be analysed as lexical units. In (17), the Dutch word string *pauze houden* 'take a break' is inserted into the Arabic clausal frame and forms an embedded-language island. Syntactically, the string is an instantiation of the predicate-complement construction; from the lexico-grammatical perspective, it is an idiomatic collocation. Boumans (1998: 246) contends that the collocational status of the string explains the absence of a determiner before the noun, which is obligatory in Dutch usage. The sentence in (18) allows for two different analyses. Following the MLF model, one has to admit that *j-confronter* 'oppose' is a mixed constituent and the nominal phrase *ces idées* 'these ideas' is an embedded-language island. Yet, following the unit hypothesis an analysis is preferred according to which the whole sequence in French is a lexical unit, and only one of its elements acquires an Arabic morpheme. Support for the latter treatment of the instance comes from the corpus-driven dictionary *Wortschatz: Corpus français*, which lists the nouns *idées* 'ideas', *réalité* 'reality' and *expériences* 'experience' as the most frequent collocates of *confronter* 'oppose'. (19) is a similar case: the lexical items *car* and *drive* are inserted into the Tamil matrix structure. While the noun receives the Tamil accusative suffix, the verb does not take any infinitive suffix required in spoken Tamil (i.e., neither *-kka* nor *-a*, cf. Schiffman 1999: 73), and is thus analysed as a bare form. From the lexico-grammatical perspective, however, the two words form a collocation: the noun *car* is the most frequent collocate of the verb *drive*.[12] The insertion of *car* and *drive* can thus be regarded as unit insertion, even though the unit is separated by the Tamil suffix *-ei*.

The presented evidence, even if sporadic, is in favour of the unit hypothesis. Nevertheless, owing to the broadness of the generalisation that the hypothesis expresses, the acceptance of the hypothesis would be unreasonable if any embedded-language morpheme sequence inserted in the context of the matrix language is granted unit status (cf. Wray 2002: 42). However, Backus (2003: 91)

[12]This fact is confirmed by the observations of nominal collocates of the verb *drive* in the *BYU-BNC* corpus and *The Corpus of Contemporary American English (COCA)*.

is aware of the generality of the hypothesis and points to two types of counterexamples (also see Backus 1999a: 105–107). One type involves multiword embedded-language insertions that are not lexical units in the embedded language. The other type includes insertions of single words instead of expected multimorphemic units. To account for the second type of counterexamples, Backus (2003) puts forward the *"conceptual unit" hypothesis*. The hypothesis predicts that "[t]he use of EL conceptual structure in CS [=code-mixing] can, but does not have to, lead to EL units. The actual morphemes do not have to be from the EL. ML morphemes will have semantically basic meanings in such cases" (p. 92). To illustrate the "conceptual unit" hypothesis, I draw on an example from Irish-English code-mixing data:

(20) Irish-English (Stenson 1990: 184)
Tagann sé isteach *handy*.
come-PRS it inward
'It comes in handy.'

We can regard (20) as a partial loan translation of the corresponding English idiomatic expression (Stenson 1990: 184). In line with the "conceptual unit" hypothesis, the matrix language items – here Irish – express semantically basic meanings, whereas the item with a specific meaning is realised in English. However, Backus provides counterexamples for his second hypothesis as well. These include cases in which a collocation element bearing a specific meaning is in the matrix language, for example:

(21) Turkish-Dutch (Backus 2003: 111)
koffie dök
Dutch: *koffie inschenken*
'pour coffee'

Backus (2003: 113) asserts that possible reasons for the occurrence of such mixed collocations are the grammatical incongruence between the languages, as in (21), or processing factors such as the effect of recency. At the same time he admits that not every counterexample can be explained.

Corroborative evidence regarding the role of multimorphemic units in code-mixing is offered by Treffers-Daller (2005b), who analyses Dutch-French code-mixing involving compounds and nominal groups, and by Muhamedowa (2006: 67), who discusses the insertion of plural forms in Kazakh-Russian code-mixing. Moreover, Myers-Scotton, in her more recent work (2006), acknowledges that

two types of lemma entries exist in the lexicon, she states that "[e]ntries for some content morphemes combine with other entries cross-linguistically in 'fast and clean' on-line production. The entries for some other content morphemes are parts of holistic multimorphemic units that are readily integrated into ML frames…" (p. 211). Despite the support for Backus's hypotheses, one type of counterexamples for the unit hypothesis are not handled explicitly in his approach. As mentioned above, such examples include multiword embedded-language insertions that are not lexical units in the embedded language. In other words, determining unit boundaries, except when relying solely on code-mixing data, is problematic. Specifically, it can be difficult to decide whether an embedded-language word string in the matrix language context corresponds to one lexical unit, as the hypothesis predicts, or a combination of such units. For instance, several units can be posited for the multimorphemic insertion in (22):

(22) Ewe-English (Amuzu 2013: 22)
Míe *download journal article* ma-wo katã xle nyitsɔ
1PL that-PL all read a.day.removed
'We downloaded all those journal articles and read [them] a few days ago.'

One possibility, following the unit hypothesis, is to regard the string *download journal article* as a unit. Another option is to analyse the string as consisting of two units *download* and *journal article*. A third possibility is to presume that the string is a blend of two collocations combined online *journal article* and *download article*. Remember that in order to qualify as a unit, the string has to either convey a non-compositional meaning or be very frequent. These criteria make the first analysis implausible, i.e., the string does not have a non-compositional semantics, nor do its parts co-occur frequently. Whether the second, or the third analysis should be adopted needs further exploration.

The problem with this specific instance and both hypotheses is the need for verification testing (cf. Wray 2002: 42). Although I relied on corpus data as additional evidence when I discussed some of the multimorphemic insertions above, the evidence provided is clearly unsystematic. A systematic examination of the hypotheses will involve an analysis of a bilingual corpus. Nevertheless, my random analysis of instances of mixing as well as the evidence provided by other researchers point at the plausibility of the hypotheses articulated by Backus (2003). To put the unit hypothesis to a test and to gather evidence in its favor or against it through a systematic study of a bilingual corpus is the aim of this book. I will explore multimorphemic insertions by comparing them with mono-morphemic, or minimal, insertions in a bilingual corpus and by analysing their distributions in large monolingual corpora.

An issue that may have ramifications for the analysis of minimal insertions in a bilingual corpus is their status in the lexicon. Some studies of bilingual speech classify single-word insertions as pertaining to either code-switching/mixing, or borrowing (e.g., Poplack 2018). Other analyses abandon this distinction, either partially (e.g., Muysken 2000: 78–81), or altogether (e.g Backus 2013, 2015). In the next section, I will present and discuss the various views on this issue.

1.6 Insertional code-mixing versus lexical borrowing

There is a widespread agreement in the field that in a synchronic analysis of code-mixing, minimal insertions fall under either switched, or borrowed words (e.g., Haugen 1953a, Poplack et al. 1988, Myers-Scotton 1993, Muysken 1995, Thomason 2001). Borrowed items differ from switched items in their status: the former are not alien but nativised in the variety of the language spoken in a bilingual community. The process of word nativisation, i.e., its introduction and assimilation in the recipient language, can only be studied diachronically (cf. Poplack & Dion 2012) and is thus beyond the scope of this study. However, a synchronic differentiation between switched and borrowed forms, as I will argue, is possible and useful. (Backus 2015 is a proponent of a dynamic approach to code-mixing and borrowing, which integrates the synchronic and diachronic aspects of these phenomena.) A confusion of switched forms with native forms may lead to a skewed database (Myers-Scotton 1993: 164), resulting eventually in an inadequate analysis. As argued by Poplack (2011), the division of lone embedded-language items into borrowed and switched should not be done *a priori* and has thus to rely on the these items' specific characteristics. Before I discuss the diagnostic features of borrowed forms, I will briefly sketch the typology of language mixing phenomena proposed by Poplack and associates.

The approach developed by Poplack and presented in her volume 2018 *Borrowing: Loanwords in the speech community and in the grammar* assumes three types of phenomena: code-switching (equivalent to my use of the term "mixing"), nonce borrowing and established borrowing. The first criterion for distinguishing between code-switching and borrowing is the number of words switched: a multiword sequence is an "unambiguous" code-switch. A lone other-language item may either be a switch or a borrowing (cf. Poplack 2011: 2). If this item is integrated morphosyntactically into the recipient language, it qualifies as a borrowing (cf. Sankoff et al. 1990). Morphosyntactic integration is the second criterion used in this classification. A bilingual speaker may borrow an item for a moment, or the whole bilingual community can use it on a regular basis. While

the borrowed form in the former case is classified as a nonce borrowing, it is regarded as an established borrowing, or a loanword, in the latter case. That is, the final criterion, on which the typology is based, involves the processes of spread and nativisation. The category of borrowing as used in at the beginning of this section will thus correspond to Poplack and collaborators' category of *established borrowing*.

In order to embed their classification in a more general discussion of borrowing, it is useful to contrast their typology with the central categories adopted in the Matrix Language Frame model. Poplack's multiword switches are akin to Myers-Scotton's embedded-language islands. Nonce-borrowings correspond to mixed, or bilingual, constituents, and morphosyntactically unintegrated single-word switches are identical to bare forms. Finally, the category of established borrowing coincides with Myers-Scotton's category of lexical borrowing (cf. Myers-Scotton 1993: 163–170). These correspondences are given in Table 1.2.

Table 1.2: Contrasting comparison of the approaches to language mixing by Poplack (2018) and Myers-Scotton (2002).

Author	multiword insertion	Single-word insertion		
			integrated	
		unintegrated	sporadic	recurrent[a]
Myers-Scotton	embedded-language island	bare form	mixed constituent	lexical borrowing
Poplack	code-switch	code-switch	nonce borrowing	established loanword

[a](and nativised)

Despite the differences between these approaches in the methodological procedures applied and the theoretical issues raised, the structures that both approaches identify and investigate exhibit a high degree of overlap. However, similarities end as soon as we examine how lexical borrowings are identified on synchronic grounds. A lack of agreement on this issue is characteristic of the whole field.

Researchers who distinguish between switched and (established) borrowed forms apply the following diagnostics: morphological integration into the base language (MacSwan 2000), recurrence (Poplack & Sankoff 1984, Myers-Scotton 1993), diffusion in the community (Poplack et al. 1988, Poplack 2018), listedness

(Muysken 1995, 2000, Muhamedowa 2006, Stammers & Deuchar 2012) and speakers' acceptance (Poplack & Sankoff 1984). Whilst MacSwan (2000) treats all Spanish words integrated morphologically into Nahuatl as borrowings, *morphological integration* is not a sufficient criterion for Myers-Scotton (1993). She claims that incomplete morphological integration may apply to both borrowed and switched forms (Myers-Scotton 1993: 191). According to Myers-Scotton (1993: 183–188), languages can use several patterns for marking a particular feature: one pattern presupposes full morphological marking, whereas another involves partial marking, or a lack of marking. Consequently, when applied to the morphological integration of foreign words, these patterns result in either full or incomplete integration. Myers-Scotton refers to these two types of marking as central and peripheral. Elaborating on the idea of central and peripheral marking, Boumans (1998: 52–53) links different marking types to productivity; he states that some of the morphological processes in a language are more, or less, productive than others. That is, a borrowing can demonstrate a lack of integration when it follows a specific, morphologically unproductive pattern (cf. Dressler 2003). Muhamedowa (2006: 45) illustrates this point by considering a case of morphological non-integration in Russian: foreign nouns with a vowel in the stem-final position, such as *kakáo* 'cocoa', *kófe* 'coffee', *pal'tó* 'coat' and the like, form a class of indeclinable nouns within the Russian nominal system. Thus, these long-established loans cannot be regarded as fully integrated into the Russian morphological system. To take another example, German verbs borrowed from Latin or French which include long-established loans ending in *-ieren*, such as *regieren* 'rule', *probieren* 'taste', *studieren* 'study' and many others, are integrated into the German verbal system also only to a degree. Unlike the majority of German verbs, they do not take the prefix *ge-* as part of the circumfix in the form of past participle, following the rule that *ge-* attaches only to initially stressed items (cf. the pair *mach-en* 'make-INF' – *ge-mach-t* 'PTCP-make-PTCP' and the pair *regier-en* 'rule-INF' – *regier-t* 'rule-PTCP'). These examples show that established loanwords may follow a certain pattern in a language which does not allow for full morphological integration. Note that the material used for illustrations includes forms pertaining to monolingual grammars. Therefore, only a synchronic analysis of a monolingual system, possibly supplemented with a diachronic analysis, can enable inferences about the degree of integration. I argue that when bilinguals produce switched or borrowed forms on-line, they can draw on all patterns available to them, regardless of their status, whether central or peripheral. In other words, if a peripheral, or unproductive recipient language pattern is used to accommodate an item from another language, this item should be analysed as well-integrated, for the produced mixed form concurs with the morphosyntactic requirements

of the recipient language (cf. Sankoff et al. 1990).[13] In this respect, a synchronic analysis of code-mixing alone does not seem appropriate for determining productive patterns of a language, and the only possible outcome of such an analysis is the identification of forms that either exhibit morphosyntactic integration or lack it (cf. Poplack & Dion 2012). As such, morphosyntactic integration is not an adequate criterion to distinguish between insertional mixing and (established) loanwords, but a prerequisite for a form to qualify as a borrowing (cf. Poplack 2018).

As Myers-Scotton (1993) refutes morphological integration as a criterion for distinguishing borrowing from insertional mixing, she makes the division between the two categories by virtue of *recurrence*. She asserts that "CS [=code-switching/mixing] forms have little recurrence value, in contrast with B [=borrowed] forms" (Myers-Scotton 1993: 163) and suggests frequency of occurrence in absolute and relative values as a reliable criterion. Her book *Duelling languages* contains two case studies which operationalise recurrence as frequency, these studies investigate the realisation of numerals in Shona-English code-mixing and the use of English *because* and *but* as borrowed forms in Shona. The discussion of the results entails concern regarding relative frequency: like every researcher dealing with the factor 'frequency', Myers-Scotton is eager to know "'how much' relative frequency is 'enough'" (1993: 204), and admits that setting the threshold is an arbitrary decision. Despite the expressed reservation, she views the distinction between borrowed and switched forms as crucial for the matrix language frame model because borrowed forms, and not switched forms, are projected by lemmas tagged for the matrix language (Myers-Scotton 2002: 41).

A discussion of recurrence as a determinant of (established) borrowings would be incomplete without mention of the pioneering work by Poplack et al. (1988). The authors perform a large-scale analysis of lone other-language items in a vast corpus of French-English bilingual speech, of a size unmatched by any other bilingual corpus to date. The usage frequency of a borrowing is related to its social integration and defined by the number of speakers using it. Established loans – corresponding to "borrowings" in the terminology adopted in this section – are thus distinguished by *diffusion* in the speech community. Specifically, Poplack et al. (1988: 55) handle borrowed items as widespread loans if they are uttered by more than ten speakers whose speech is represented in the corpus. The authors

[13] Note that the difference in the approaches is also manifested in the grammatical domains where integration is investigated: Myers-Scotton (1993) and Boumans (1998) refer to morphological integration alone, whereas Poplack and associates (e.g., Sankoff et al. 1990, Meechan & Poplack 1995) proceed from integration in the morphosyntax of a language. Furthermore, whilst the former regard integration as a matter of degree, the latter view it as an abrupt process.

acknowledge that the threshold for this criterion, i.e., ten people, is the result of an "arbitrary, though rather severe" decision (Poplack et al. 1988: 100). To avoid a terminological confusion, it is necessary to emphasise that Poplack et al.'s use of the term "frequency of use" deviates from Myers-Scotton's considerably: for Myers-Scotton, frequency of use is the frequency with which a borrowed item occurs in a corpus of bilingual speech, regardless of the number of people who utter it. Both conceptions of frequency as a criterion for distinguishing borrowing from mixing have been criticised in a number of studies (e.g., Haust 1995, Boumans 1998, Muhamedowa 2006, Stammers & Deuchar 2012).

Whilst one problem with word frequency that Haust (1995: 49), Boumans (1998: 57) and Muhamedowa (2006: 46) identify is that coincidental circumstances such as the conversation topic and the speech style disparage frequency distribution in smaller corpora – a common source of material in the field – a more serious objection is raised by Boumans (1998: 57) and Backus (2013), who assert that a borrowed word in a specific community can be a switched item for a subset of speakers of this community. Beyond that, Backus (2013: 29) takes an extreme position and states that the mixing-borrowing "debate is misguided, because a foreign-origin word can be both: borrowing and codeswitching [=code-mixing] are not mutually exclusive like that". These considerations make Backus (1996: 96–97) and Boumans (1998: 58–60) abandon the distinction between these phenomena as hopeless. Nonetheless, other researchers, for example, Muhamedowa (2006), maintain this distinction and rely on listedness as the factor determining the status of loanwords.

Listedness as a distinctive feature of borrowing is introduced in Muysken (1995). Muysken (1995) defines borrowing as a process "which involves the incorporation of lexical elements from one language in the lexicon of another language" (p. 189). Hence, he views code-mixing as a supra-lexical phenomenon and refers to borrowing, whether momentary or established, as a sublexical phenomenon. Muysken's approach resembles that of Poplack and associates. However, in Muysken (1995, 2000), the differentiation between established and nonce borrowings is based on the criterion "listedness", i.e., being part of a memorised list. This means that an (established) loanword has to be listed in the lexicon of a speaker after the corresponding speech community has accepted it. According to Muysken (1995, 2000), conventionalisation may also affect code-mixing. Namely, specific multiword sequences in the source language can appear in the recipient language discourse on a regular basis. multiword loans are common in monolingual use: for instance, numerous Latin expressions such as *ex post*, *ad hoc*, *tabula rasa*, *anno domini*, *et caetera* have been adopted by many European languages.

Lantto (2015) examines conventionalised word sequences in a situation of bilingualism. Using data from the Spanish Basque Country she shows that Spanish multiword discourse markers are conventionalised to such an extent that they infiltrate most bilingual conversations. When multiword strings, represented in the mental lexicon as units, acquire a high degree of diffusion in the community, they can behave like established borrowings. Thus, care should be taken in analyses of code-mixing regarding multiword strings because these strings are also possible candidates for conventionalisation.

Muhamedowa (2006) and Stammers & Deuchar (2012) use the criterion "listedness" to differentiate between code-mixing and borrowing. An operationalisation of this criterion includes a reinterpretation of listedness as a property of the mental lexicon: the authors reinterpret it as dictionary attestation. This operationalisation proves useful especially in the context of long-standing language contact, such as the situations of bilingualism between Welsh and English (Stammers & Deuchar 2012) and Kazakh and Russian (Muhamedowa 2006). However, such operationalisation is problematic for languages and varieties spoken in immigrant settings in view of the absence of lexicographic sources. In the aforementioned study, Poplack et al. (1988) also considered dictionary attestation as a possible determinant of social integration of English lexical items in the French-speaking community. In doing so, they systematically sought the examined English lexical items in numerous dictionaries of Canadian and European French. The result of their analysis (Poplack et al. 1988: 58–59) is that dictionary attestation is not a reliable predictor of loanword status: 18% of frequent (i.e., used by more than ten speakers) English words in French discourse from their corpus were not attested in the corresponding dictionaries. At the same time, some words listed in those dictionaries had the status of momentary borrowings in their data. In contrast to this study, Stammers & Deuchar (2012) test listedness, operationalised as dictionary attestation, and token frequency as predictors of full morphological integration of English verbs in Welsh and come to the opposite conclusion. In Welsh, the initial consonants of verbs employ soft mutation depending on the lexico-grammatical context. According to the authors, borrowed verbs achieve full morphosyntactic integration only if they use consonant mutation and thus completely assimilate into the Welsh system. Their analysis shows that only verbs attested in the corresponding dictionary rely on soft mutation to a degree comparable with native Welsh verbs. High-frequency verbs that are not attested in the corresponding dictionary do not employ this morphophonological alternation. Thus only "listed" verbs are considered *bona fide* established loans. My interpretation of this result is twofold. For one, Stammers & Deuchar's result can be possibly explained by the duration of language contact. In a cross-

linguistic study of foreign lexeme integration, Nortier & Schatz (1988, quoted in Boumans 1998: 53) argue that the integration of foreign lexemes can be related to the duration of language contact. They find that in a situation of long-established contact, as that between Spanish and Quechua in South America, lexeme integration is higher than in a situation of recent contact, such as that between Dutch and Moroccan Arabic in The Netherlands. We can assume that English verbs attested in the Welsh dictionary are subject to the examined morphophonological alternation because they entered Welsh earlier than current high-frequency borrowings, which are still not affected by this process. In other words, the lexical items may correspond to different historical layers of the vocabulary and can hardly be comparable. Another explanation of the restricted applicability of Welsh soft mutation to English verbs frequently occurring in Welsh lies in the nature of stem alternation. Haspelmath & Sims (2010: 216) state that "the effects of morphophonological alternations need not be found in loanwords", even if a particular stem alternation is productive (e.g., Turkish *k*/*ğ* and Indonesian Nasal Substitution, ibid.: 219). The examined alternation, like Nasal Substitution in Indonesian, might be no longer productive when applied to novel words as such, which is why unattested frequent English items remain unaffected. Crucially, determining the productivity of the alternation examined by Stammers & Deuchar (2012) will require obtaining psycholinguistic evidence.

The overview of various approaches to borrowing confirms that there are good reasons to differentiate between mixed and borrowed forms in a synchronic analysis of code-mixing. One reason is that borrowed items, whether at the sublexical or supra-lexical level, apparently have the same status in the mental lexicon as native items. Hence, ignoring this distinction would result first in a skewed database and eventually in a misleading analysis of code-mixing. The overview of the studies tackling the problem of identifying borrowing demonstrates that there is no single criterion applicable to all contact situations and all databases. Dictionary attestation is impossible for languages spoken in immigrant communities, for example in the Russian-speaking community in Germany. Control over diffusion of foreign lexemes in the community, a correlate of their social integration, sets the most demanding requirement for compiling vast corpora of bilingual speech.[14]

In this situation, the only feasible criterion is recurrence, operationalised as token frequency in absolute and relative values. The research reported in the subsequent chapters will thus consider usage frequency of items at both the sublexical and the supra-lexical levels. That is, I will examine single words, but also

[14]This is one reason why the findings by Poplack et al. (1988) have not been validated in other corpora to date.

multiword sequences in the bilingual corpus with regard to their usage frequencies since all of them may be possible candidates for social integration in the speech community.

1.7 Conclusion

In this chapter, I have presented current approaches to the linguistic patterning in bilingual speech. According to the widely accepted typology articulated by Muysken, the major types of language mixing include insertion, alternation, and congruent lexicalisation. Following Muysken, I have argued that there is a correlation between a specific type of mixing in a given community and a particular constellation of linguistic, psychological and social factors, although the social conditions of language use take precedence over other factors. For example, I showed that in immigrant settings the dominant type of mixing depends on the bilingual individual's generational membership in an immigrant community, which is in its turn related to the individual's bilingual proficiency and, in the end, to their linguistic experience as such. The speech of intermediate-generation immigrants, who are usually fluent in both the community and the country language, exhibits variable patterns of insertional mixing, with minimal and maximal insertions alternating with each other.

Minimal insertion, described as a process whereby content morphemes from another language are combined with recipient language markers, is a pervasive process of word incorporation in language contact. There is general agreement among scholars that this is a default mode of insertional mixing. At the same time, views vary widely on the nature of maximal insertion, or the occurrence of embedded-language islands, and the factors contributing to their emergence in bilingual speech. While the proponents of the Matrix Language Frame model argue that maximal insertions arise because there is a lack of overlap between the lexico-grammatical aspects of the equivalent structures in two languages, the cognitive-linguistics approach to bilingual speech proposed by Backus assumes maximal insertions to correspond to lexical units in the mental lexicon/grammar. The unit status makes maximal insertions highly accessible to bilingual individuals in online speech production and hence contributes to their ubiquity in bilingual speech. I have argued that although this suggestion seems perfectly plausible, it requires further research and systematic evaluation.

The chapter has also discussed insertion and lexical borrowing as distinct phenomena of bilingual speech. Despite adverse views aired on this dichotomy, most researchers agree that these are different, though related phenomena and should

thus be distinguished in analyses of bilingual speech. However, controversy exists over the criteria indicative of lexical borrowing. In light of this debate, I have presented arguments in favour of the view that recurrence, or frequency of use, may be operationalised as a reliable diagnostic feature of lexical borrowing.

2 Usage-based approaches to grammar and variation

The last two decades have witnessed a remarkable advance in empiricist approaches to language, whose hallmark is a commitment to psychological realism in explaining and modelling language and human linguistic behaviour. Among these approaches, usage-based linguistics is regarded as one of the most fruitful and rapidly expanding research programmes. In a usage-based view, linguistic representations emerge and change because cognitive processes, commonly referred to as "domain-general psychological mechanisms", operate in language use. These general and basic capacities of the human brain such as memory, categorisation, chunking, analogy, cross-modal associations, to mention only the most relevant ones, affect linguistic representations and patterns of language structure observed in human linguistic behaviour. For my account of Russian-German code-mixing, I adopt a usage-based perspective on language and language variation. This chapter presents the main assumptions underlying usage-based theories and thus lays the theoretical foundation for the studies reported in Chapters 4–6. As it is beyond the scope of this chapter to give a detailed review of all the premises, findings and insights gleaned from recent usage-based work in various branches of linguistics, I restrict myself to introducing those tenets and concepts which are basic to the usage-based approach to code-mixing as being developed in this thesis.

To set the stage, I start by giving a brief description of linguistic representations posited by usage-based exemplar models. In light of the focus of this book on multimorphemic words and multiword sequences (henceforth, multimorphemic elements), the following section will scrutinise aspects of their acquisition, comprehension and production. The final section introduces usage-based approaches to language variation, which is viewed as emergent from competitions between functional linguistic units and communication strategies in interaction. Additionally, to supply the reader with an illustration of a motive for competition, I will elaborate on the cognitive process of priming and discuss it in the light of usage-based models.

2.1 Rich memory for language: Exemplars, networks and constructions

Usage-based models posit that the language user's memories of specific language experiences are stored in the brain as exemplars. A specific language experience leads to the activation of numerous exemplars along multiple dimensions. This section will show how exemplars are organised and how new linguistic experience updates networks of exemplars. Further, it will elaborate on the emergence of abstract structure, most notably of schematic constructions, through generalisations over exemplars as an emergent consequence of network organisation.

The usage-based position builds on the fact that the human brain has a very large mental storage capacity (Sherwood 2012: 126). The brain keeps track of a human being's experience through storing detailed information about her specific experiences as memories. With regard to language, these memories comprise "phonetic detail, including redundant and variable features, the lexical items and constructions used, the meaning, inferences made from this meaning and from the context, and properties of the social, physical and linguistic context" (Bybee 2010: 14; cf. Langacker 1987, 2000). In other words, long-term memory representations abound in fine-grained information about experience with language.

Usage-based theory holds that linguistic memories are organised as exemplars, which are representations formed of individual tokens of linguistic experience (Johnson 1997, Bybee 2001, 2010, Pierrehumbert 2001). Exemplars and their clusters, or clouds, emerge in the process of categorisation, which is a cognitive mechanism whereby a new token of experience is compared against an established exemplar with the purpose of determining whether it belongs to the same category. This process is carried out as follows: If the organism considers a new token of experience and an existing exemplar the same, the new token strengthens the existing exemplar. If a token of experience differs from the existing exemplars in some dimension, it may either fade away or, if reinforced by later experiences, form the basis for a new exemplar. Similar exemplars are stored near one another to constitute clusters or categories.[1] In Johnson's (1997: 146) terms, categorisation is based on "sums of similarity over each category". This implies that a single usage event, or a token of experience, affects multiple exemplar clusters. In a nutshell, the exemplar model holds that every token of experience, or use, affects cognitive representation (cf. Bybee & Beckner 2009), although it should be kept in mind that "an individual exemplar – which is a detailed perceptual

[1]Sherwood (2012: 127) explains the fact that long-term memory traces are stored with memories of the same type by the need to maintain memory stores searchable for future retrievals.

memory – does not correspond to a single perceptual experience, but rather to an equivalence class of perceptual experiences" (Pierrehumbert 2001: 141).

Exemplar models were developed to capture similarity and frequency in perception (cf. Pierrehumbert 2002). However, already Johnson (1997) articulated the idea that an exemplar model is capable of explaining the production-perception link. He argues that exemplars include not only auditory properties, but also articulatory properties. According to Pierrehumbert (2002), a goal for the current production corresponds to the average properties of the exemplars in a cluster. In order to produce a certain linguistic item, activation spreads to a group of exemplars in an area of the perceptual map. Hence, this type of model can successfully account for the effects of language exposure on production.

One of the consequences of modelling linguistic representations in terms of exemplars is that "[e]xemplar representations provide a natural way to allow frequency of use to determine the strength of exemplars" (Bybee 2006: 717). The idea that usage events, which are inherently repetitive, can shape the cognitive representation of language is relatively recent in modern linguistics, but it was upheld as early as 1880 by Hermann Paul in his *Principles of the History of Language* (cf. Auer 2015). The following passage from Paul conveys the gist of his version of an exemplar model:

> […] every act of speaking, listening, or thinking adds something new. Even by an exact repetition of a previous act, some forces of the already existing mental grammar are strengthened. And even if somebody can look back on a rich linguistic life there is always the opportunity for something new: Aside from the introduction of things that were previously unusual in the language, at least a new variant to old elements may be added. […B]oth the attenuation and the strengthening of the old elements as well as the addition of new ones constantly shift the associations within this system of representations. (Quoted in Auer & Murray 2015: 31)

We can conclude from the above that the perception-production feedback loop enables the system to capture the dynamic character of experience by constantly updating the existing exemplars. Another idea contained in the passage concerns the fact that, if not reinforced, linguistic memories, just as non-linguistic memories (Sherwood 2012: 127), may decay (Pierrehumbert 2001) and thereby trigger a reorganisation of the network (cf. Bybee & Beckner 2009). This scenario is particularly common in a situation of an ongoing language change (Bybee 2006: 718).

In an exemplar model, a word in a speaker's lexicon corresponds to a cluster, or cloud, of phonetic exemplars which represent the word's phonetic variants with information about their linguistic and social contexts. The meaning of the word is stored in a cluster of semantic exemplars which represent information about the meaning and context of each token of a word. Together the cloud of phonetic exemplars and the cluster of semantic exemplars are considered a unit (for an overview and further details, see Bybee & Beckner 2009, Bybee 2010). Pierrehumbert (2001) points out that the same remembered tokens may be categorised according to several schemes. This idea can be illustrated by the phrase *Stand clear of the closing doors, please*, whose recollection can involve such categories as the words and phonemes in the phrase, 'underground rapid transit' and, in some speakers, 'male voice' and 'the New York City Subway'.

Usage-based models assume that words and even word strings are linked to other words and word strings. These relations arise from similarities among words along phonetic and semantic dimensions (Kruszewski 1995: 101 [1883]), and are represented in networks of exemplars. The exemplar space, or networks, naturally provide the capability for generalisation based on similarity (cf. Pierrehumbert 2002). Generalisations emerge on multiple levels. For example, Bybee (1985, 2010, 2002b) argues that both morphological relations and morphemes arise from similarity relations among words grounded in shared semantic and phonetic features. Specifically, she shows that the internal structure of a word, such as *revitalise*, is a result of its relations with other words, including *vitalise*, *revitalisation*, *generalise*, *vital*, *redo* and the like.

Above the word level, word strings are also represented in a network of relations. Storage is required for word strings which exhibit idiosyncrasies of meaning, such as *red herring* or *make waves*, and those which are compositional in form and meaning but represent conventionalised expressions, such as *happy birthday* – and not *joyous birthday* – or *take a decision* – and not *seize a decision*. In the former case, the necessity to register the string in memory arises from the unpredictability of its meaning, whereas in the latter case, it results from the use of a string as a conventional sign for a concept. By virtue of a network representation, a speaker may retrieve the sequence as a whole and yet maintain associations between the sequence and its component parts as independent words, i.e., the words *red*, *herring*, *make* and *waves* in the respective examples (cf. Bybee 2010: 25).

With regard to the storage of complex words and word strings, a usage-based approach is not concerned with the question whether a complex unit is represented or not, it rather handles questions concerning "the strength of the representation and the strength of its association with other representations, both

paradigmatic and syntagmatic, all of which are variable" (Bybee 2010: 24; cf. also Lieven 2010). The strengths of representations and associations, corresponding to particular units, reflect a language user's experience with these units and has been found to correlate with the (frequency) distributions of these units in her usage. Specifically, the strength of representation of a linguistic unit, often referred to as "entrenchment" (Langacker 1987, Croft 2001, Tomasello 2003, Blumenthal-Dramé 2012), depends on the token frequency of this structure, or its frequency of occurrence. There is a general agreement among usage-based linguists (e.g., Langacker 1987: 59; Bybee & Scheibman 1999: 581; Tomasello 2003: 106–107; Bybee 2007: 283; Blumenthal-Dramé 2012: 68) that repetitive sequences are entrenched to become units. However, even more important is the acknowledgement that unit status is a matter of degree, since a sharp distinction between units and non-units is psychologically implausible (Langacker 1987: 59).

As stated above, storing exemplars of words and word sequences which are similar on one or several dimensions in close proximity to each other in a network enables the emergence of exemplar clusters, or categories. The network organisation has as a consequence that exemplar clusters develop association relations on the basis of similarity and co-activation patterns. These association relations in the exemplar space underlie generalisations over exemplar clusters. Usage-based theories hold that the network organisation of the exemplar space gives rise to representation of abstract structure, including phonological, morphological and syntactic structure. It must be noted however that the emergence of abstract knowledge in consequence of generalisation over specific items still needs further empirical support and theoretical elaboration. Below I will exemplify a possible path of emergence of such structure with a case of a syntactic construction.

Usage-based linguistics regards constructions, defined as "direct form-meaning parings that range from the very specific (words and idioms) to the more general (passive construction, ditransitive construction), and from very small units (words with affixes, *walked*) to clause-level and even discourse-level units" (The "Five Graces Group" (2009: 5), cf. Croft 2001, Goldberg 2003, 2006), as basic units of grammar.

Usage-based analyses (e.g., Boas 2003, Lieven et al. 2003, Goldberg et al. 2004, Dąbrowska & Lieven 2005, Bybee & Eddington 2006, Boyd & Goldberg 2011) have shown that when similar words differing in some aspect appear in the otherwise same sequence in the input, exemplar categories for both the invariable and varying items emerge. While the group of varying exemplars represents a schematic position in a construction, the group of exemplars corresponding to the invariable elements form the fixed slots in a construction. For example, the word se-

quences in (1) have the word *hour* as a fixed element, whereas the words adjacent to it alternate: the items in the initial position are numbers and the items in the final position denote various acts of travelling from one place to another.

(1) *thirty*-hour *ride*
half-hour *drive*
four-hour *flight*
two-hour *trip*
three-hour *journey*
two-hour *hop*
three-hour *slog* (data from Hoey 2005: 16–17)

The range of nouns that occur in the final position is limited, and the semantic category they belong to is constrained: they all refer to some act of travelling. My analysis of these nouns in the British National Corpus (BNC, Davies 2004–) revealed that they occur with the word *-hour* with the following frequencies: *journey* (35), *drive* (26), *flight* (26), *trip* (12), *ride* (4), *hop* (0), *slog* (0). We could conclude from these figures that the word *hour* combines with the word *journey* more often than with the other words. Following Bybee (2010), we can regard the sequence *-hour journey* as a prefab, which is defined as a highly frequent exemplar that "represents the conventional way of expressing an idea" (p. 81). Prefabs provide the basis for the emergence of semantic categories. Semantic restrictions apply to even more schematic categories (cf. Bybee 2010: 81), such as the category constituted by the words occupying the first position in the examined sequences, all of which are numbers. The construction that arises from the specific instances in (1) in the process of generalisation could be described as NUMBER-*hour* ACT-OF-TRAVELLING, where the fixed slot is surrounded by two schematic slots (it is noteworthy that both slots exhibit differing degrees of schematicity). Even the word occurring in the second position in the examined sequences may alternate. Possible alternatives include *day*, *week*, *month* and *year*. Together with the word *hour*, they form the schematic category TIME-UNIT. A broader generalisation results in the construction NUMBER-TIME-UNIT ACT-OF-TRAVELLING.

As has been shown above, under the exemplar model the representation of a construction, which may be a fairly broad generalisation, is linked to the exemplars of all specific words experienced in a certain position in that construction. These item-general links may be maintained even when specific word sequences instantiating these constructions grammaticalise (Torres Cacoullos & Walker 2009). Importantly, the view of constructions as generalisations over specific exemplars does not mean that items previously unexperienced in a partic-

ular construction may not be used with it in production. On the contrary, constructions are potentially productive (Goldberg 2006, Lieven 2010). Their productivity depends on the semantic – and sometimes even formal – specifications of their schematic slots (Bybee 2013: 57–59; see also Zeschel 2010). In other words, a speaker may extend the use of a construction to new items if their semantic properties are compatible with the semantic properties of the words that support the corresponding generalisation. To illustrate this point, I again refer to the NUMBER-TIME-UNIT ACT-OF-TRAVELLING construction. Hoey (2005) argues that thanks to this generalisation (his term is "semantic association"), language users can produce and interpret such sequences as *27-hour meander*, *27-week flight* and *multi-month odyssey*, although they may have never experienced them. The fact that the knowledge of language encompasses not only simplex words but also schematic constructions, their specific instances and combinations of schematic and concrete pieces of language means that language elements are simultaneously represented in the mind at various degrees of granularity (Goldberg 2006, Bybee 2006, Tomasello 2003; see Langacker 1987: 63–76 for the view of grammar as a structured inventory of linguistic units). This means that a language user may represent a given multiword sequence or multimorphemic word in her mind as a concrete expression and as an instantiation of some more schematic construction or constructions.

To summarise, an exemplar model posits that linguistic structure is represented in the language user's mind as memories of specific language experiences, which are organised as exemplars, and simultaneously in form of generalisations over these memories. Exemplars of concrete linguistic expressions are stored in clusters. Newly experienced instances of language are subject to categorisation. If they correspond to any of the existing exemplars, they exert an accumulative effect on this exemplar. If they are not identical but similar to one or several existing exemplars, they are stored in close proximity to them and may form a new cluster. Exemplar clusters are organised in networks on the basis of various associations between them. These associations are due to either simultaneous activation or similarity. Schematic constructions may emerge from these networks as generalisations over exemplars representing concrete items. A linguistic expression may be simultaneously represented in the language user's mind at multiple levels of granularity. Namely, the language user may retain the memory of the expression as a whole, the voice that pronounced it and its intonation pattern, but also the situation in which the expression was uttered. Furthermore, she is most likely to activate and possibly reinforce the exemplars corresponding to the components of this expression. These components include specific items – i.e., chunks, collocations and single words – and instantiations of more schematic

syntactic and morphological constructions as well as the phonetic features and phonological patterns underlying single words and longer word sequences. Representing structures of varying degrees of specificity/abstractness is possible because different processes are involved; namely, representations of specific items rely on patterns of co-activation in language use, whereas the representation of abstract structure emerges in consequence of similarity detection and categorisation.

One of the consequences of an exemplar model is that it relates the strength of a mental representation directly to language use. In order to be able to store linguistic structure, an individual has to encounter and use it on a regular basis. This is of particular relevance to multimorphemic elements since the language user can only remember those strings that appear in the input frequently enough. In the next section, I will show that while some multimorphemic elements may be memorised as unanalysed holistic amalgams in the process of language acquisition, representations of other strings may emerge in the mind through chunking, whereby the existing representations of smaller linguistic items corresponding to parts of frequently experienced strings are repeatedly co-activated in sequence to give rise to larger processing units representing the respective sequential information. I will first describe the learning of multimorphemic elements by rote in first-language acquisition and will then elaborate on the process of chunking as the other source for representing these structures in the mind.

2.2 Recurrent multimorphemic elements in language acquisition and use

A body of evidence has emerged in last two decades suggesting that everyday language involves a wide array of fixed multiword, or formulaic, sequences (e.g., Corrigan et al. 2009, Schmitt 2004, Wray 2002, 2008, Taylor 2012). However, this idea was already expressed as early as 1974 by Bolinger.[2] He asserted, "Speakers do at least as much remembering as they do putting together" (1976; quoted in Erman & Warren 2000: 29), and argued – in contrast to the views then current – that the language user stores a large number of complex items. Sinclair (1991) cast the same idea into the idiom principle and the open-choice principle. He defines the idiom principle in the following way:

[2] According to Tremblay et al. (2011) the idea of the unintentional fusion of two or more linguistic signs into a single unit goes back to Ferdinand de Saussure (1959 [1916], quoted in Tremblay et al. 2011: 571), for whom fusion was a particular type of agglutination.

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (p. 110)

The idiom principle is contrasted with the open-choice principle, positing that almost every position in the syntactic structure allows an open choice, i.e., virtually any word can fill a slot. Sinclair relates the open-choice principle to the so-called "slot-and-filler models" of grammar (p. 109), such as traditional structuralist (and generative) grammars. These models, according to Pawley & Syder (1983), cannot account for native-like, i.e., idiomatic, selection and native-like fluency in spontaneous conversation. The authors emphasise that speakers should store abundant units of clause length ("lexicalized sentence stems" in their terminology) in order to minimise "the amount of clause-internal encoding work [...and] to attend to other tasks in talk-exchange, including the planning of larger units of discourse" (192). Although the ideas expressed by Bolinger (1976), Pawley & Syder (1983), and Sinclair (1991) lack explicit links to usage-based accounts of language, they are highly compatible with the usage-based view of grammar (cf. The "Five Graces Group" et al. 2009). For example, one of the possible factors that Sinclair considers responsible for the emergence of the idiom principle is "the recurrence of similar situations in human affairs" (p. 110), which is obviously a usage-based motivation.[3] And Pawley & Syder (1983) point out that a large proportion of fluent stretches in the conversations that they analyse are familiar, memorized clauses and clause-sequences (p. 208). This is in line with the usage-based tenet that a human being's memory for language is rich.

The observation that large portions of language are available to language users as prefabricated, or formulaic, units has stimulated research into the nature of these units and their usage. Extensive studies of formulaic expressions have been conducted in corpus linguistics and applied linguistics. Whereas studies of spoken language investigate conversational routines (Aijmer 1996, Altenberg 1990, 1998), analyses of large corpora of written texts often centre around the identification of prefabricated phrases (e.g., Erman & Warren 2000, Gries 2010, Evert 2005) and their use in various registers (e.g., Biber & Conrad 1999, Gries & Mukherjee 2010). Prefabricated phrases are naturally of particular relevance to linguists interested in second language acquisition (e.g., Granger 1998, Ellis et al. 2008a, Biber et al. 2004, Schmitt 2004). And recently, psycholinguistic studies have gathered considerable evidence that language users – both children and adults – have

[3]The other two motivations suggested by Sinclair are "a natural tendency to economy of effort", and "the exigencies of real-time conversation" (p. 110).

allocated representations for strings of concrete linguistic structures and heavily rely on them in language comprehension and production.

A central tenet of the usage-based approach to multiword sequences and multimorphemic words is that, given a sufficient frequency in a language user's experience, a complex expression tends to be entrenched in her mind so that it can be easily accessed and fluently executed. Some sequences are learnt as holistic units already during language acquisition, while other sequences may be chunked (i.e., gain an independent representation) later, as experience with language grows.

Following these observations, the next section first reports findings from the literature on the use of multimorphemic elements by children learning a language, then it introduces chunking as a principle of memory organisation, briefly indicating some of its consequences for language organisation and use. The final sub-section outlines some pertinent comprehension and production studies investigating the processing of multimorphemic elements in adults.

2.2.1 Recurrent multimorphemic elements in first-language acquisition

A continuous strand of research from as early as 1970s suggests that language development begins at a very concrete level (see in particular Bowerman 1973, Clark 1974, Braine 1976, Tomasello 1992, Pine & Lieven 1993, Dąbrowska & Lieven 2005). Considerable evidence has been adduced to date that language acquisition begins with learning multiword and multimorphemic adult expressions as unanalysed units (Lieven et al. 1997, Dąbrowska 2004, Bannard & Matthews 2008, Lieven et al. 2009, Arnon & Clark 2011). One reason why children store more than individual words in their memory is that "they do not hear demarcated words in the input; words and phrases run into one another and must be detected in the speech stream" (Bannard & Matthews 2008).

In a diary study of his daughter's speech, Michael Tomasello (1992) observes that her first word combinations correspond to unanalysed adult expressions. He finds that one of the earliest multiword expressions in her speech is *whereda-bottle*, which the child produced at the age of around 15 months, aiming at the adult expression *Where is the bottle?* (p. 45). Another example is the expression *get-it*, which the child started to use around two months later to request objects that were in sight but inaccessible (p. 72–73). This expression is again modelled on her parents' specific constructions, including *I'll get it*, *Go get it* and *You can get it*. But unlike wholly rote-learnt utterances such as *whereda-bottle*, this expression draws only on the invariable parts of the frequently experienced con-

structions. Tomasello refers to this type of expressions as limited-scope formulae (p. 22).

In a series of corpus studies, Elena Lieven and her colleagues (Lieven et al. 1997, Dąbrowska & Lieven 2005, Lieven et al. 2009) demonstrate that productivity in young children's utterances is limited, and up to a half of children's first 300 multiword utterances represent fixed strings. The authors conclude that children are learning both single words and word strings, which are learnt as "big words". Such word strings are reported to exhibit a high degree of repetitiveness in the input and are therefore highly learnable. While learnt word strings are usually subject to a subsequent internal analysis, sufficiently entrenched strings may remain in the memory unanalysed and be represented as fully lexically specific strings. Experimental evidence in support of this latter claim has been obtained by Bannard & Matthews (2008), who tested young children's memory for familiar word sequences. The authors extracted frequently-occurring chunks in a corpus of child-directed speech, such as *sit in your chair*, and matched them to infrequent sequences, e.g., *sit in your truck*. Preschoolers' ability to produce these sequences was tested in a sentence repetition task. Two and three-year-olds were significantly more likely to correctly repeat frequent sequences than infrequent sequences. The authors conclude that children have allocated representations for sequences of more than one or two words.

Inbal Arnon (2011) argues that children learn word sequences not only because the latter are highly repetitive in the input, but also because children naturally attend to larger phrases. She regards word strings as good candidates for a Gestalt process, defined as "the move from large unanalyzed units to the identification and analysis of smaller more structured ones" (p. 167). She provides evidence for this claim in a experimental study conducted jointly with Eve Clark (2011). This work investigates four to six-year-olds' production of irregular plurals in English, depending on the context in which the irregulars occurred. The given context involved either a lexically specific frequent frame, e.g., *Three blind* – or a general question *What are these?* The authors found that children produced 72 per cent of correct irregular plurals after lexically specific frequent frames and only 32 per cent after a general question. These findings imply that children represent sequences of words, which reflect their language use.

With regard to sequences of lexical and grammatical morphemes, such as inflected forms, substantial evidence has been presented that young children learn them as unanalysed amalgams (see Tomasello 2003: 118–119; and Clark 2009: 190–193, for reviews). For example, Pizzuto & Caselli (1992) observe this strategy in three Italian-speaking children (from 1;4 to 3;0 years of age) learning the verbal

morphology. Specifically, they find that of six possible inflected forms, the children used only one form with 47 per cent of all verbs, and two or three forms with 40 per cent, while four or more forms were observed only with 13 per cent of all verbs, which included highly frequent, irregular verbs, learnt by rote. One implication of these findings is that children at this age do not use inflectional suffixes productively, but rather learn combinations of specific verbal stems with particular inflectional suffixes. In other words, before abstracting the verb paradigm children start storing those inflected forms whose frequency in the input is high.

In the domain of nominal inflection, Dąbrowska (2004) shows that in the speech of Polish-learning children, the first multimorphemic forms consisting of a stem and an inflectional suffix of the genitive case appear only with particular lexical items before they begin to be used productively with other lexical items some six months later. A similar situation is observed by Pine & Lieven (1997) in young children learning English determiners. The authors demonstrate that specific determiners in children's utterances are closely associated with particular nouns. This result is interpreted as evidence that children store in their lexicon/grammar specific sequences consisting of determiners and nouns before they develop the category of determiner.

Putting the matter in a nutshell, studies of child language have amassed indications that child language acquisition involves rote learning of many highly specific multimorphemic words and word combinations. However, as pointed out above, representations of such structures may also emerge in language use as a consequence of chunking.

2.2.2 Chunking and the mental representation of multiword sequences and multimorphemic words

Chunking is the cognitive mechanism whereby permanent sets of associative connections are developed in the long-term memory (Ellis 1996). This mechanism is responsible for a fast and fluent performance of sequenced actions (Melton 1963). With increasing practice, or frequency, sequences of actions are performed faster because the sequences are processed as units (Miller 1956, Anderson 1982). Newell (1990) regards chunking as the overarching principle of human cognition:

> A chunk is a unit of memory organization, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit. Chunking implies the ability to build up such structures recursively, this leading to a hierarchical organization of memory. Chunking appears to be a ubiquitous feature of human memory. Conceivably, it could form the basis for an equally ubiquitous law of practice. (p. 7)

He points out that people chunk at a constant rate: Every time they get more experience, they build additional chunks.

The observation that new behavioural routines emerge as experience grows has been applied to language only recently. Researchers who argue for the relevance of chunking to language are Nick Ellis (1996), Joan Bybee (2002a), and Christiansen and Chater (2016). Their proposal is that the frequency with which linguistic units are perceived, or produced, in sequence strengthens associations between them in the long-term memory. Another mechanism is suggested by Diessel (2016), who argues that the process behind the emergence of units corresponding to sequential elements is automatisation, which is closely related but not identical to chunking. He describes automatisation in the following way:

> Automatization is the cognitive mechanism whereby controlled processes are transformed into automatic processes. Almost all sequential activities start off as controlled processes, but are then often transformed into automatic processes through repetition or practice. This is a very common cognitive phenomenon involved in many everyday tasks. Automatization enables people to perform complex sequential activities with little effort [...] (p. 19)

It appears that chunking and automatisation are interrelated as both refer to aspects of complex sequential activities. Diessel asserts that these processes complement each other: Automatisation is the process responsible for the processing of sequential elements and their transformation to units, whereas chunking is the process behind their storage and organisation in memory (p. 20). However, whether it is beneficial to contrast these complementary processes is a practical and theoretical question. On the one hand, the distinction between the procedural and representational aspects of sequential actions may provide us with additional insight into specific properties of these processes.[4] On the other hand, if we adopt Melton's (1963) conception of memory as a continuum of short-term memory and long-term memory, we have to conclude that a separation of these processes is problematic because every execution of sequential actions leads not only to automaticity in performance but also irresistibly strengthens the corresponding trace in the long-term memory (cf. Hay 2001: 1046). For this reason I will not treat chunking and automatisation as essentially different processes in

[4]Diessel (2016) notes that chunking and automatisation are not always distinguished in linguistic literature.

this work.[5]

In the domain of second-language acquisition, Ellis (1996) argues that language learning is concerned with the acquisition of memorised sequences of language on various levels of linguistic structure. He reviews experimental studies documenting a reciprocal relationship between short-term memory and long-term memory in language learning and observes that short-term representation and repetition are responsible for the development of long-term sequence information, whereas long-term sequence representations enable the chunking of working memory contents (p. 115). The latter circumstance could be interpreted as the basis for the development of hierarchical relations between elements in the process of learning.

The emergence of hierarchical relations in syntax is the focus of Bybee (2002a). The author reports evidence from the literature that frequently repeated word combinations undergo fusion, or are chunked, sometimes in violation of traditional notions of constituent structure. This process is captured by the Linear Fusion Hypothesis: "Items that are used together fuse together" (p. 112; see also Bybee & Scheibman 1999). In a study conducted in a corpus of spoken American English, Bybee observes a general correspondence between patterns of sequential co-occurrence and constituent structure in the domain of the noun phrase and comes to the conclusion that "the hierarchical structure of language is derivable from the more basic sequential nature of language".

In a similar vein, but in an more general perspective, integrating language evolution, acquisition and processing, Christiansen & Chater (2016) view chunking as an essential tool for processing language because it allows the brain to "compress and record its input into "chunks" as rapidly as possible" (p. 15) and to pass it to a higher level of representation. Listeners have to act in this way, as Christiansen & Chater argue, because they are confronted with a massive flow of linguistic information in face-to-face interaction and memory is limited at every level of representation. In production, the process is reversed as the memory constrains require that "only a few chunks are kept in memory at any given level of linguistic representation" (p. 102). Crucially, this model directly links chunking and the multiple levels of linguistic representations in the mind. In other words, chunking is fundamental to language processing and thus to the organisation of language.

[5]Although Bybee (2002a) does not define chunking and automatisation ("automation" in her terminology) explicitly, her treatment of these processes rests on their inseparability, as is reflected in her account of fluent production of recurrent word combinations: "[…] repeated sequences become fluent because they become automated into a single chunk that can be accessed and executed as a unit" (p. 316).

To summarise, individual items processed in sequence are grouped together into chunks. With increasing practice, the associative links between the representations of the individual items become strong, so that the sequence can be accessed and executed as a unit. The process of chunking, whereby such sequential units emerge, is crucial not only to fluency and the fusion of sequential elements but also to hierarchical relations in language. The following section reviews studies which investigate the processing of recurrent multimorphemic words and multiword sequences and thus presents compelling evidence for their psychological reality.

2.2.3 Processing of multimorphemic elements

2.2.3.1 Comprehension studies

The impact of usage frequency on the processing of multimorphemic words and multiword sequences has been the focus of almost four decades of psycholinguistic research. Starting from the early 1970s, experimental psycholinguistics has richly documented a facilitatory effect of lexical frequency on the recognition of multimorphemic, or complex, words, but it is only recently that researchers have developed an interest in the processing of multiword sequences (this is probably due to the late rise of corpus-based studies). Generally speaking, psycholinguistic studies confirm the observation that both frequent multimorphemic words and frequent multiword sequences are processed in a different way than their less frequent counterparts.

Early studies of lexical access to multimorphemic words, which include inflected, derived and compound words, report a correlation between the speed of access to a multimorphemic word and its usage frequency (Morton 1969, Taft & Forster 1976, Bradley 1981, Stemberger & MacWhinney 1986; see also Giraudo & Grainger 2003, Janssen et al. 2008 as examples of some recent studies). But already Taft (1979) presents evidence that the recognition time for multimorphemic words is sensitive to both the word's surface frequency and the frequency of the word's base (cf. Burani & Caramazza 1987, Colé et al. 1989, Alegre & Gordon 1999, Meunier & Segui 1999). To account for the competition between a complex word and its base morpheme in lexical access, many current models of morphological representation postulate two access routes for multimorphemic words: a decomposed route and a non-decomposed route. In the latter case, a complex word is accessed supralexically, i.e., directly, whereas in the former case, access is mediated by contact to the decomposed parts of the word, its base in the first place (cf. Caramazza et al. 1988, Baayen 1992, Frauenfelder & Schreuder 1992, Baayen &

Schreuder 1999, Hay 2001, Blumenthal-Dramé 2012; but see Bien et al. 2011 for an alternative approach). These models cogently accommodate the effect of a word's surface frequency by stipulating that holistic storage of a multimorphemic word is a function of its surface frequency relative to the frequency of its base. In other words, the higher the frequency of the multimorphemic word relative to the frequency of its base, the higher the likelihood for this word to be accessed directly and to have an allocated autonomous representation in the mental lexicon. Most studies investigating lexical access to multimorphemic words provide support for this frequency effect, even though the overall reported results appear somewhat inconclusive. This confusion owes in part to differences in the types and functions of the analysed morphemes (e.g., inflectional vs. derivational suffixes), the typological differences between the investigated languages, but also, as Hay (2001) notes, some methodological caveats (p. 1063–1066).

To give an example, several studies which investigate processing of singulars and plurals[6] by using lexical decision tasks present conflicting evidence for English (Sereno & Jongman 1997, New et al. 2004), on the one hand, and Dutch (Baayen et al. 1997) and French (New et al. 2004), on the other hand. Whereas in all languages the effect of surface frequency holds for plural forms of plural-dominant words, i.e., words whose plurals are more frequent than singulars, the processing of singular forms behaves according to the surface frequency in English (Sereno & Jongman 1997, New et al. 2004), but neither in Dutch (Baayen et al. 1997) nor French (New et al. 2004), where it predominantly depends on the base frequency. The competition between the base and the whole inflected word was interpreted as an indication of parallel activation of both the storage route and the decomposed route.

Findings from studies which explore the processing of multiword sequences are more straightforward, although all of them are based on English data and the examined multiword sequences vary in length, structure and complexity. This research relies on diverse on-line methods such as the phrasal decision (or recognition) task (Bod 2000, Arnon & Snider 2010), the word monitoring task (Sosa & MacFarlane 2002, Kapatsinski & Radicke 2009), the self-paced reading paradigm (MacDonald 1993, Reali & Christiansen 2007, Bannard & Ramscar 2007, Tremblay et al. 2009, 2011) and eye-tracking methods (McDonald & Shillcock 2003, Siyanova-Chanturia et al. 2011).

According to Jurafsky (2003: 51), MacDonald (1993) was one of the first to show that the frequency with which two words appear together influences the reading

[6] I report the findings on the processing of singulars and plurals here because they are relevant for Chapter 6, which explores the use of German plurals in otherwise Russian sentences.

time of these words in sequence.[7] She used a self-paced reading task to investigate the factors that determine the reading time of sequences consisting of nouns and words ambiguous between nouns and verbs, such as *stores*. The interpretation of the second word as a noun is likely to be evoked if the noun noun pair is frequent, such as *grocery stores*, but not in the case of infrequent noun noun pairs, such as *warehouse fires*. The finding that an inverse relationship exists between the frequency of a word-pair (bigram) and its recognition was extended by Bod (2000) to three-word sequences. He demonstrated in a phrasal decision experiment that frequent three-word sequences, which corresponded to subject-verb-object sentences, such as *I like it*, were recognised faster than their infrequent counterparts, e.g., *I keep it*, when other factors including word frequency, word complexity, syntactic structure and plausibility are controlled. A reason for this result, as pointed out by Jurafsky (2003: 62), may be a lack of control over sub-string bi-gram frequencies. Interestingly, Bannard & Ramscar (2007) could replicate Bod's (2000) result for sequences of between 4 and 7 words under control of the aforementioned factors as well as sub-string bi-gram frequencies.

I will not give a concise account of recent research but only briefly raise the point that the multiword sequences examined in the reported studies vary in terms of their structure. For instance, Reali & Christiansen (2007) and Arnon & Snider (2010) as well as the aforementioned studies investigate the recognition of multiword sequences which exhibit an overlap with the phrase structure. The examined sequences include such items as *a lot of places* and *I have to pay* in Arnon and Snider's (2010) investigation, and *I*-verb chunks, e.g., *I met* and *I liked*, in Reali and Christiansen's (2007) work. The former study explores the processing of four-word sequences in isolation, whereas the latter study investigates the comprehension of pronominal object relative clauses, of which the target pronoun-verb combinations are part. Following Tremblay & Baayen (2010), I refer to multiword sequences of this kind as phrasal multiword sequences. The processing of non-phrasal multiword sequences is the focus of Tremblay et al. (2011). The authors adopt Biber et al.'s (1999) term "lexical bundle" to refer to multiword strings: "A lexical bundle (LB) is a relatively common continuous multiword sequence that may span phrasal boundaries". Although the definition implies that a lexical bundle crosses syntactic boundaries only optionally, many of the items examined by Tremblay et al. do not match the phrase structure, e.g., *in the middle of the*. We may therefore refer to them as non-phrasal multiword sequences. It should be noted, however, that each of the studied items was embedded in a sentential context of two words to the left and to the right, e.g., *I sat* ***in the***

[7] Another study which Jurafsky mentions is Trueswell et al. (1993).

middle of the *bullet train*. Unfortunately, no comprehension studies are known to me which pursue a comparison between phrasal and non-phrasal multiword sequences as well as their processing in isolation versus in a context.

Another aspect of multiword sequences that I would like to highlight is the competition between sequences forming linguistic units and their parts. Kapatsinski & Radicke (2009) argue that although a high-frequency multiword sequence seems to form a linguistic unit of its own, access to it may be mediated by competition with its parts. The authors conducted a monitoring experiment in which participants had to respond whenever they detected the particle *up* in a verb-particle combination such as *get up*. Reaction times sped up with the frequency of the verb-particle collocation increasing. But the particle detection decelerated when the collocation frequencies reached the highest values. This result confirms Sosa and MacFarlane's (2002) finding that detection of *of* was slower in collocations in the highest frequency bin such as *kind of*. Kapatsinski & Radicke interpret these results as evidence for the storage of only high-frequency phrases as units in the lexicon and the competition between the part and the whole for collocations which are not stored in the lexicon. In such sequences, detectability of the particle improves with its predictability growing, given the sequence. These findings are reminiscent of the outlined parallel activation account of multimorphemic words, which posits a competition between the storage route and the decomposed route (Baayen et al. 1997, Hay 2001).

In sum, despite the indicated differences in the methodology, the utilized material, and the theoretical interpretations, all the outlined studies suggest that regular multiword sequences, whether phrasal or non-phrasal, leave memory traces in the brain. This result is consistent with the evidence laid out above that regular multimorphemic words, whether inflected, derived, or compounded, may be stored in memory holistically.

2.2.3.2 Production studies

This section outlines psycholinguistic research which established the influence of frequency on the production of multimorphemic words and multiword sequences. In each of these domains, I will first present some relevant findings from acoustic-phonetic corpus studies and then provide a brief outline of most relevant experimental research.

Studies which investigate the production of multimorphemic words in spoken language corpora are based on the observation that lexical frequency determines the production of words in spontaneous speech. As as rule, high-frequency words are subject to phonetic reduction (Bybee 2000) and tend to be pronounced shorter

than low-frequency words (Jurafsky et al. 2001). Studies focusing on the production of complex, multimorphemic words not only aim to pinpoint factors responsible for phonetic reduction, but also to scrutinise the (possible) consequences of this process for the morphological structure of wordsd.

For example, Keune et al. (2005) examine the variation in the reduction of Dutch words with the suffix *-lijk* such as *natuurlijk* 'of course', *moeilijk* 'difficult' and *uiteindelijk* 'finally', and attribute it to socio-geographic and linguistic factors. The analysis of these multimorphemic words in a corpus of spoken Dutch reveals that the degree of reduction is the largest for the high-frequency words *natuurlijk* 'of course', *mogelijk* 'possible' and *eigenlijk* 'actually', whose reduced forms are [tyk], [mok], [ɛɪk], respectively.[8] The authors conclude that these words undergo a process of erosion, which is marked by a loss of morphological structure and a development towards monosyllabic forms. Alongside frequency, the variation in the reduction of words ending in *-lijk* is explained as an interplay among such factors as socio-geographic origin, speech rate, the word's position in the sentence and its contextual predictability, measured by mutual information, which is a frequency-based probability measure of the likelihood that a word would occur given the preceding word. Schäfer's (2014) analysis of Icelandic adverbs with the suffix *-lega* confirms Keune et al.'s result. The main finding of his work is the effect of lexical frequency on the phonetic reduction of adverbs in *-lega*, in the absence of any effect on the part of metrical rhythm.

Bergmann (2012) uses electropalatography, an articulatory method, to investigate the influence of lexical frequency and prosodic structure on articulatory reduction of n#g-sequences in German compounds such as *Zinn#krie-ger* 'tin warrior' and particle verbs, e.g., *ein#geben* 'to enter'. This research focuses on the effects of word frequency, accentuation and vowel quantity in the first part of the multimorphemic word. Bergmann's findings suggest that reduction of the alveolar nasal is likely in items which are (i) distinguished by high frequency, (ii) contain long vowels and (iii) occur in an unaccented position. The results were more straightforward for particle verbs than for compounds. Taken together, the reported findings suggest that lexical frequency is an important determinant in lexical production (but see Jurafsky 2003: for another interpretation). Specifically, it influences the duration of a word, the degree of its reduction and finally its prosodic and morphological structure.

A number of studies of both spontaneous and elicited speech have recently presented evidence for the role of frequency in the production of multiword se-

[8]Keune et al. emphasise that in addition to these highly reduced forms, other, less reduced forms also exist.

quences. For example, words may be subject to reduction depending on the specific linguistic context in which they are used. Bybee & Scheibman (1999) find that the reduction of the contraction *don't* – as analysed in a corpus of naturally occurring conversation recorded by Scheibman in Albuquerque, New Mexico – is influenced by the frequency of multiword sequences of which *don't* is part. The greatest degree of reduction was observed in sequences whose usage frequency was especially high in the data; these phrases include *I don't know*, *I don't think*, *I don't have (to)*, *I don't want* and *I don't care*. The authors argue that the reduction of the middle element in these multiword sequences is possible because the high frequency of these sequences grants them unit status in the mental lexicon/grammar. Corroborative evidence for Bybee & Scheibman' findings comes from Bell et al.'s (2003) study, which shows that speakers tend to reduce frequent function words, such as *I* and *the*, in recurrent multiword sequences, or as the authors put it, in positions in which the word is predictable from neighbouring words. Unlike Bybee & Scheibman (1999), who interpret vowel reduction in recurrent multiword sequences as conditioned by the holistic representations of these sequences, Bell et al. argue that this process is driven by the probabilistic relations between words, they contend that "[w]hile some of this reduction may be due to lexicalization of multiword phrases, some of it is due to the mental representation of some kind of probabilistic links between words, since the effects are not limited to frequent collocations" (p. 1021). As is evident from the above, the two views on the representational status of recurrent multiword sequences, namely the localist and the distributed view, were already articulated in the early literature on the issue.

Especially interesting in regard to the representational status of multiword sequences are studies investigating repair and the overall distribution of disfluencies in spontaneous conversation (e.g., Schneider 2014). For instance, Kapatsinski (2005) reports that how much is repeated in a repair is influenced by the distributional information beyond individual words. The repetition in a repair usually involves the last word, as in (2a), but sometimes two or even more words, cf. (2b).

(2) a. I really appreciated [*the*, +**the**] whole, uh, English class
b. The crime level is not as high as it is in other areas [*of the*, +**of the**] city (Kapatsinski 2005: 481)

The extent of the recycle in repetition repairs is investigated as determined by syntactic constituency and frequency-based probabilities. The results suggest that how much speakers recycle depends largely on the constituent structure, i.e., speakers start to repeat from the nearest syntactic boundary. But they tend

to cross that boundary given a high-probability transition (as in 2b). The results are interpreted as evidence for the representational status of probabilistic links between words. At the same time, we could entertain the possibility that a high transitional probability may indicate that the repeated material, whether a single word or a multiword sequence, exhibits a high degree of cohesion (for the effect of frequency on interruptibility of words, see Kapatsinski 2010). This interpretation would be in line with the observation that cohesive units play an important role in speech production, as has been shown in analyses of disfluencies (for reviews, see Kapatsinski 2010: 74–75, and Schneider 2014). Outlining this research, Kapatsinski (2010) states that "[...] interruption is sensitive to cohesion: speech production is more likely to be interrupted at the boundary between cohesive units than within a cohesive unit" (p. 75). Schneider (2014) exploits this observation and conducts a large-scale analysis of hesitation placement in three syntactic contexts. Her results confirm the fact that cohesive multiword sequences repel disfluencies, but the effect could be observed for multiword sequences in both the high-frequency and the mid frequency range. This finding tallies with the observation from language comprehension studies (cf. Reali & Christiansen 2007, Arnon & Snider 2010) that the chunk frequency effect is gradual in nature. Against this background, it is impossible to definitely answer the question of whether the extent of the recycle in a repair as well as disfluency placement are conditioned by transitional probabilities between words or whether they are determined by cohesive production units.

2.2.3.3 Representational status of multiword sequences: speeded computation vs. holistic retrieval

In this context, experimental studies of language production focusing on the representational status of multiword sequences appear particularly promising, since they allow one to examine the production of multiword sequences when their frequencies and the frequencies of their parts are controlled for. Other relevant factors include meaningfulness of a sequence and syntactic constituency.

In one such study, Janssen & Barber (2012) test whether the frequency of a multiword phrase, such as a noun-adjective combination, determines naming latencies in a language production task. In the noun-adjective condition, Spanish and French native speakers were asked to name objects in one of ten colours, using a standard noun phrase in Spanish or French, respectively, i.e., object followed by colour. The second condition of the experiment was varied between the two groups: whereas the Spanish participants were told to name objects corresponding to noun-noun combinations, the French participants were asked to

produce a set of determiner-noun-adjective phrases. The latter condition enabled a comparison between two-word and three-word sequences, and allowed the experimenters to test the hypothesis that naming latencies for determiner-noun-adjective phrases are sensitive solely to the phrase frequency and are not influenced by substring frequencies. On the basis of these studies, the authors demonstrate that naming latencies for all the examined phrase types were determined by their respective phrase frequencies. They report the effect of whole-string, or phrase, frequency in the absence of any effect of the frequency of the component parts, including the substring frequency (see Tremblay et al. 2011: for a similar result observed in language comprehension). On the one hand, we can consider these result to provide indirect support for the localist view on the representation of multiword sequences, i.e., their retrieval from memory as holistic units. On the other hand, Janssen & Barber explicitly do not rule out other possible explanations, admitting that "[…] phrase frequency effect might reflect transitional probabilities between individually stored words, [or] the connection weights between low-level input and higher level output representations […]" (p. 10).

Interestingly, Janssen & Barber's (2012) results disconfirm those of Tremblay & Baayen (2010), who demonstrate the relevance of both the frequency of a string and the frequencies of its parts to language production. In an immediate free recall task, Tremblay & Baayen (2010) investigated the production of four-word sequences by native Canadian English speakers. In their experiment, participants were exposed to four-word sequences, thereafter they were asked to recall as many sequences as possible. The analysis of correctly and incorrectly recalled four-word sequences revealed that recall was modulated by whole-string probability of occurrence, based on varying frequencies of the four words, given the preceding tri-grams. This result supports the conclusion that four-word sequences are stored as both wholes and parts (cf. the findings of the aforementioned study by Kapatsinski & Radicke 2009). Crucially, Tremblay & Baayen acknowledge the fact that it is impossible, by using behavioural data, to answer the question of whether the whole-string probability effect reflects speeded computation or holistic retrieval. Therefore, by resorting to electroencephalogram (EEG), they collected electrophysiological data, which shed light on this controversial issue. The recorded data indeed revealed that the amplitudes of P1 and N1 waves during the production of the high-frequency four-word sequences under scrutiny are comparable with the amplitudes reported for single word processing. This finding suggests that "it is most unlikely that *four words* can be accessed, let alone stringed together, within this time frame" (p. 171). In other words, Tremblay & Baayen deliver corroborative evidence that recurrent four-word sequences are retrieved as holistic units in speech production. This result confirms the local-

ist view, according to which a high-frequency sequence develops an allocated representation in the long-term memory as a consequence of its repeated activation, and the frequency with which a sequence is accessed strengthens the representation of that sequence (cf. Bybee 2010, Hay 2001, Reali & Christiansen 2007, Siyanova-Chanturia et al. 2011). However, further experimental evidence is needed to support this view.

In addition to the empirical results challenging the distributional explanation of the chunk frequency effect, reported for the recognition and production of recurrent multiword sequences, some scholars have expressed theoretical reservations against this explanation. Recall that the assumption underlying the distributional explanation is that frequent activation of associations between individually stored words leads to the speeded online computation of these words as a sequence (cf. Jurafsky et al. 2001, McDonald & Shillcock 2003). That is, associative links between the items of a sequence become stronger and directly reflect the frequency-based probability of their co-occurrence. The distributed explanation of the chunk frequency effect accords well with the more general and widely held but possibly idealised conception of language users as unconscious statisticians. Blumenthal-Dramé (2012) raises a number of points of criticism regarding this view, which basically equates knowledge of language and knowledge of the statistical structure of language (p. 41, see p. 36–44 for the critique). Following Bley-Vroman (2002), she regards the statistical structure of language and frequency effects as secondary by-products of the semantic and pragmatic dimensions of language and advocates a functional view, according to which the distribution of a construction, regardless of its specificity, is determined by its functional role (cf. Goldberg 2006). At the same time, she acknowledges the challenge of envisaging experiments which could effectively investigate frequency and function as orthogonal factors in the processing of multimorphemic words.

In the domain of multiword sequences, however, there exists some encouraging evidence that function, or meaning, outperforms frequency in the recognition of idiomatic and compositional multiword units. By using a reaction time task, Jolsvai et al. (2013) investigate the role of a multiword chunk's relative meaningfulness independently of the chunk's compositional status. They find that highly meaningful multiword sequences, exhibiting both idiomatic and compositional semantics, are processed faster than less meaningful non-phrasal multiword sequences.[9] The authors show that although chunk frequency is also predictive

[9]Note that while I reserve the term "non-phrasal multiword sequence", following Tremblay & Baayen (2010), Jolsvai et al. refer to these multiword sequences as "fragments". Another term for this phenomenon is a "non-phrasal lexical bundle" (Biber et al. 1999, Tremblay et al. 2011).

of the processing latencies, the meaningfulness of different chunks is the most important factor determining them. They conclude that "the meaningfulness of multiword chunks may be as important to their processing as their distributional properties" (p. 696). Furthermore, their results offer corroborative evidence for the usage-based hypothesis that both idiomatic and compositional sequences are stored as form-meaning pairings in the brain (Bybee 2010, Goldberg 2003, 2006).

Despite the evidence offered that phrasal multiword sequences have processing advantages over non-phrasal multiword sequences, we may attribute this effect to syntactic constituency (though this interpretation is at odds with Arnon & Cohen Priva 2013, who found no effect of constituency on production latencies for multiword sequences). Syntactic constituency was also the focus of the aforementioned study by Tremblay & Baayen (2010). An analysis of their electrophysiological data reveals that phrasal multiword sequences leave memory traces in both the centro-parietal and the occipito-parietal pathway, whereas non-phrasal multiword sequences leave memory traces only in the occipito-parietal pathway. The authors explain this effect in terms of the meaningfulness of a multiword sequence and not by constituent structure. They attribute this effect to the fact that "phrases instantiate (relatively) complete concepts compared to non-phrases" (152). This explanation ties in with the observation reported by Schmitt et al. (2004) that semantic/functional transparency of multiword sequences – exemplified by *I don't know what to do* and *go away*, on the one hand, and *in the same way as* and *aim of the study*, on the other hand – is responsible for higher performance scores in their production, as was found in a dictation test (in the absence of any correlation between the phrase frequency and performance).

Taken together, these results suggest that (i) frequent multimorphemic words as well as idiomatic and compositional multiword sequences tend to be processed in a holistic manner rather than computed online, (ii) this frequency effect is gradual in nature and (iii) driven by meaning/function. Beyond the general mechanisms underlying the representation of recurrent multiword sequences in the mind, as discussed so far, are the individual differences in their processing and use. Studies such as McCauley & Christiansen (2015) report substantial inter-subject variability in online processing of multiword sequences and attribute it to individual differences in chunking, i.e., the ability of sequence learning, as well as to the subjects' linguistic experience (see also Christiansen & Chater 2016: 192–194). Crucially, Verhagen et al. (2018) demonstrate that the variation in knowledge of multiword sequences is systematic and results from differing degrees of familiarity with these sequences. According to the authors, knowledge of specific multiword sequences varies between social groups as well as between individuals within these groups and reflects their experience with the specific items. To put

these findings into a broader perspective, I will outline usage-based approaches to language variation in the next section.

2.3 Language variation as competition

In this chapter so far, I have dealt with representational aspects of language as posited by usage-based models. Now I will focus on usage-based approaches to language variation, which may be regarded as reflecting competition, or interference, between representations at specific levels of language representation owing to individual differences in linguistic experience as well as contingencies of communication.

Usage-based theories of language view variation as inseparable from language use (The "Five Graces Group" et al. 2009: 9). Linguistic variation is considered to result from the process of replication of linguistic structures in human communication. Every time speakers engage in joint actions with each other in a community, they replicate, in their utterances, the linguistic structures conventionalised in their community. But since the communicative process is indeterminate, replication is never exact and results in variation (cf. Paul 1920[1880], Croft 2000, The "Five Graces Group" et al. 2009, Poplack & Torres Cacoullos 2014). Yet, in spite of the acknowledgment of the role of linguistic variation in language use and language change, an explicit focus on variation in studies taking the usage-based perspective on language is rare (cf. Poplack & Torres Cacoullos 2014). This may be explained by the fact that within this framework, the role of specific variable patterns of use is often restricted to either indicating linguistic representations or contributing to change (cf. The "Five Graces Group" et al. 2009: 7). Hence, although most of the aforementioned corpus-based studies give due consideration to the variability of linguistic structures (e.g., Bybee & Scheibman 1999, Kapatsinski 2005, Schäfer 2014, Schneider 2014), only few studies focus on its social correlates (e.g., Keune et al. 2005; see also Lorenz 2014, Zenner et al. 2015, Verhagen et al. 2018). These studies endeavour to account for the variability of linguistic functional units by attributing it to both cognitive and socio-cultural factors. An analysis of variation is thus viewed as an analysis of competitions between motivations. The challenge of such an analysis lies in the identification of relevant competing factors and proper functional units involved in the competition.

The social nature of language determines its duality: language is viewed as simultaneously existing in individuals and the community of users. Under a usage-based view, these two levels, despite their seeming separability, are highly interrelated: "An idiolect is emergent from an individual's language use through social

interactions with other individuals in the communal language, whereas a communal language is emergent as the result of the interaction of the idiolects" (The "Five Graces Group" et al. 2009: 15). Although individual idiolects exhibit considerable inherent variability, a large amount of their heterogeneity is orderly. Patterned variation pertains to both language use (Weinreich et al. 1968) and the internal organisation and representation of idiolects (Dąbrowska 1997). Proponents of usage-based approaches to language attribute patterns of linguistic variation to interactions of representations of specific as well as schematic constructions and the general cognitive abilities underlying their acquisition (The "Five Graces Group" et al. 2009: 15). At the same time, they emphasise the role of social structure in language variation and change, admitting that linguistic interactions are inevitably determined by social networks (Milroy 1980, Eckert 2000). This background allows one to consider language variation as a generalisation based on the behaviours of different individual speakers.

Usage-based theories hold that language use is a continuous decision-making process in which speaker and hearer produce and comprehend each other's utterances, deploying particular grammatical structures and functional strategies, in order to achieve their communicative goals (Bates & MacWhinney 1989, The "Five Graces Group" et al. 2009, Bybee 2010, Du Bois 2014, MacWhinney 2014, Christiansen & Chater 2016). As Diessel (2011) puts it, "[t]he sequential decision-making is at the heart of language use; it determines the language user's linguistic behavior and the development of linguistic structure over time" (p. 841). In other words, the processes of selection, adaptation and emergence, observable in usage, originate in the decision-making process (Du Bois 2014: 264–265), and the speaker's behaviour is thus a result of competition between different pressures, or motives (Bates & MacWhinney 1989, MacWhinney et al. 2014). Just as the decision-making process itself, they operate within an individual, either speaker or hearer, and partly within the current interactional context.

Brian MacWhinney (2014) classifies motives in terms of the dynamic systems in which they operate (p. 368–370). He distinguishes between four dynamic systems: processing, memory, spatiotemporal processes of social interactions, and environment. Two of the dynamic systems, processing and memory, have already been discussed in this chapter. According to MacWhinney, the dynamic systems interact and feed each other within particular functional domains, or "arenas" – e.g.„ word production, word comprehension, sentence production, sentence comprehension, interactional maintenance, group membership and so on – or between these domains. Competition in these domains usually involves functional units, i.e.„ words and constructions of various degrees of specificity, within the

same functional niche, determined by the communicative context, the communicators' goals and the grammatical alternatives at their disposal (Du Bois 2014). For example, the decision to code-switch in a conversation would result from competition in one or several of these domains: (i) the speaker's goals (for instance, a wish to create an atmosphere of intimacy), (ii) the appropriateness of the interactional setting, (iii) the current context, including activation from previous lexemes, (iv) conversational cues produced by the co-participant, (v) estimating the co-participant's bilingual ability, and (vi) gaps in the speaker's lexicon (cf. MacWhinney 2005: 72).

John Du Bois (2014) refers to specific drives, or motives, for competition as fitness criteria. These criteria "define the fitness landscape for language, determining what counts as success in the utterance arena" (p. 278). The list of fitness criteria he distinguishes is reproduced in Table 2.1. Although fitness criteria in this approach are assigned to categories, which include meaning, cognition, evolvability, sociality and aesthetics, this is chiefly done for convenience. As such, the criteria are multifaceted and may pertain to more than one category. Du Bois emphasises the necessity to discriminate between fitness criteria and competing motivations. Competing motivations include those fitness criteria which enter competitions in the utterance arena. The outcome of these competitions determines the use of particular functional units and the corresponding communicative strategies in specified contexts. But not all fitness criteria enter competitions directly: "the timeless factors that forever frame the terms of [...] competitions – like clarity and economy – remain unchanged, persisting long after the winners and losers [among the involved functional units, or communicative strategies] have been evaluated" (p. 273). Further, an analysis of variation from this perspective considers the impact of frequency as a factor that influences the outcome of competitions between motivations, since motivations and usage are directly related. Specifically, "[t]he link is forged in the utterance arena where real competitions play out, leaving myriad marks on vast utterance populations" (p. 276). I will not elaborate further on Du Bois's approach to competition, rather, in order to demonstrate how one of the fitness criteria outlined by Du Bois motivates competitions between functional units, I will briefly introduce the fitness criterion recency, or priming, particularly because it has recently entered into the focus of attention of usage-based theories of language (cf. Gries 2005, Abramowicz 2007, Jäger & Rosenbach 2008, Diessel 2011, Bybee & Beckner 2015, Torres Cacoullos 2015).

Priming is a cognitive process, whereby using a given item increases the likelihood of using it again in the subsequent discourse within a short period of time.[10]

[10]Other most common terms which refer to this process include "recency" (Abramowicz 2007)

Table 2.1: Conflicting fitness criteria drive competing motivations (adopted from Du Bois 2014: 272).

Meaning	Cognition	Evolvability
Clarity	Economy	Transmissibility
Transparency	Simplicity	Fidelity/Heritability
Iconicity	Ease	Recognizance
Analogy	Efficiency	Learnability
Expressivity	Priming	Variability
Informativity	Memorability	Recombination
Generalisation	Distinctiveness	Viability
Individuation	Compositionality	Plasticity
Grounding/Indexicality	Reduction	Adaptivity
Monosemy	Unification	Weak linkage/
Polysemy	Binding	Double articulation
Pith/Density	Arbitrarinesss/Opacity	Fertility
		Population dynamics
		Mindshare
Sociality	**Aesthetics**	
Intersubjectivity	Beauty	
Cooperation	Symmetry	
Normativity	Resonance	
Affiliation	Affect	
Identity	Creativity	
Power/Prestige	Extravagance	
Autonomy	Authority	
Evaluation	Ritual	
	Play	

According to Bybee & Beckner (2015), the finding that a recently activated item is easy to activate again originated in experimental studies of lexical access (Forbach et al. 1974, Meyer & Schvaneveldt 1971). Levelt & Kelter (1982: 78) account for priming in the following way: "Reusing recent materials may [...] be more economical than regenerating speech anew from a semantic base, and thus contribute to fluency". In the light of exemplar theory, the chance of using an item depends on the speaker's experience with this item, which concerns all of the speaker's encounters with it – i.e., the item's cumulative frequency – and the occurrences thereof in the current context – i.e., its recency (Bybee & Beckner 2015, Pierrehumbert 2001).

Priming effects have also been observed in discourse. While repetition in discourse has been the focus of research in the discourse analytic and conversation-analytical tradition (e.g., Halliday & Hasan 1976, Tannen 1989)[11], the scant variation studies (Poplack & Tagliamonte 1996, Weiner & Labov 1983, Poplack 1980a) as Szmrecsanyi (2006)[12] puts it, "have stumbled across the phenomenon rather accidentally when parallelism in surface structure turned out to be a highly efficient predictor of the linguistic choices that speakers make" (p. 28). For example, Poplack (1980a) examines the retention or deletion of the plural marker *-s* in Puerto Rican Spanish and finds that its usage depends on recency. She shows that the retention of the plural marker is likely if the preceding word is overtly marked for plural, and the absence of the plural marker on the word is highly predictable when the previous token lacks the plural marker.

As with studies investigating frequency effects in the domain of morphosyntax, the challenge of appropriately identifying the functional unit of analysis, which I mentioned in passing above, pertains to corpus studies of priming, as well. A number of studies investigate priming effects at the level of phonologically and semantically specific symbolic units, such as inflectional morphemes (e.g., Puerto Rican Spanish plural morphemes, Poplack 1980a; past tense verb phrase marking in Nigerian Pidgin English, Poplack & Tagliamonte 1996; English comparative markers, Szmrecsanyi 2006) and functional words (e.g., English future markers, Szmrecsanyi 2006; Columbian-Spanish subject pronouns, Travis

and "persistence" (Szmrecsanyi 2006). Peter Auer (p.c.) has pointed out to me that the term "priming" may imply that a language user is exposed to some external prime, which may not always be the case in naturally occurring discourse. However, owing to the pervasiveness of this term in the literature, I will use it interchangeably with "recency".

[11]This research shows that repetition of words, forms or constructions may be motivated functionally (cf. Haiman 2014). For example, repetition is used to establish textual coherence (Halliday & Hasan 1976). This is not the perspective which I take in the present work.

[12]Szmrecsanyi (2006: 9–42) offers a comprehensive review of the literature on priming effects at the crossroads of the existing research traditions.

2005), while other studies focus on phonologically unspecified schematic units, i.e., syntactic patterns (e.g., particle placement in English, Gries 2005, and Szmrecsanyi 2006), or partially phonologically specified schematic units, i.e., morphosyntactic patterns involving phonologically specific forms (e.g., the passive-active alternation in English, Weiner & Labov 1983; the dative alternation, Gries 2005, and the genitive alternation, Szmrecsanyi 2006). The strength of the effect seems to change depending on the examined unit's level of specificity: the more specific the item, the stronger the effect (Szmrecsanyi 2006: 181–182). Consequently, when examining linguistic choices, caution should be exercised in the identification of the relevant and psychologically real functional unit of analysis. This becomes particularly evident when we contrast the traditional research of syntactic priming with more recent research. In earlier work, scholars have often attempted to find confirmation for the view that structural priming operates on the level of highly schematic syntactic representations and overlooked the relevance of functional units at the lexical-specific level (Bock 1986, Szmrecsanyi 2006). This research stands in stark opposition to recent experimental findings which provide evidence for the lexical nature of syntactic priming (Pickering & Branigan 1998, 1999, Melinger & Dobel 2005).[13] This evidence strongly supports the usage-based view of grammar as being lexically specific in nature (Lieven et al. 1997, Goldberg 2003, Tomasello 2003, Bybee 2010, Diessel 2011).

To summarise, usage-based approaches to language hold that linguistic variation is intrinsic to language because language exists in both individuals and communities of language users. A speaker's choice to use, in a specific interactional context, a particular functional unit – such as a word or construction – or a specific communicative strategy results from a decision-making process, during which several functional units or communicative strategies compete for selection.

2.4 Conclusion

In this chapter I presented a theory of language as emergent from language use, shaped by cognitive processing and grounded in social interaction. The usage-based theory rests on the fact that a human brain stores detailed information about individual experience with language. The brain stores elements of linguistic structure in the mental lexicon/grammar on multiple levels simultaneously, as concrete sequences of words and more abstract constructions. Thus, a language user represents in her mind not only words and their parts but recurrent

[13] This result is corroborated by corpus studies (Gries 2005).

multimorphemic words as well as multiword sequences, regardless of their semantic structure; that is, their meaning may be fully compositional or idiomatic. These complex functional units are either learnt by rote in the process of language acquisition, or they emerge later through the process of chunking, and ultimately through repetition while using them in interactions. Extensive literature on the processing of multiword sequences and multimorphemic words suggests that they are activated and retrieved as whole, although their constituent parts may also be activated at the same time, as a consequence of parallel activation of both the storage route and the decomposed route. As repetition is never exactly the same, and an individual's experience with language is unique, both linguistic representations and linguistic structure are intrinsically variable. Another source of language variation are the cognitive and socio-cultural pressures operating in face-to-face interactions. Hence, a usage-based approach to variation aims at explaining the variability of linguistic functional units by attributing it to both cognitive and socio-cultural factors. The subsequent chapters of this book demonstrate how this approach can be usefully applied to study bilingual speech (see also Backus 2015, Hakimov 2017 and the articles in the *Journal of Language Contact* special issue "Usage-based contact linguistics: Effects of frequency and similarity in language contact" Hakimov & Backus n.d.(b)). Specifically, I will show below that a usage-based approach allows one to analyse the structure of code-mixing in terms of a tug of war between various factors operating in online language production, such as competition between holistically stored composite forms and their parts, recency in discourse as well as perceived similarities and differences in the structure of the two languages.

3 Introducing the research participants and the corpus

As laid out in Chapter 1, code-mixing/switching is typically observed in everyday peer-to-peer conversations in bilingual communities. Data collection in code-mixing/switching studies thus involve the collecting of recorded samples of bilingual day-to-day speech[1] as well as the building of a corpus of these recordings and their transcripts (cf. Adamou 2016). According to Backus (1996: 42), the vast majority of studies analysing bilingual speech are corpus-based although other data collection procedures are also utilised in the field (see e.g., Kootstra et al. 2012, Gullberg et al. 2009). Backus (1996) has noted that all of the then current studies into code-switchin/mixing were "at least partially based on an actual corpus of spoken language data" (p. 42). This observation may well be still valid today although a bulk of experimental data have been gathered since then. The studies reported in this book are no exceptions to this tendency and also draw on a corpus of Russian-German bilingual speech, which was collected amongst Russian-German communities in Germany.

This chapter introduces the Russian-German bilinguals who participated in the data collection and explains why the speech of Russian German youths and young adults became the focus of the current research. The chapter describes Russian Germans as an ethnic group by providing information on their sociolinguistic history and their situation after repatriation to Germany. I will argue that in a situation of repatriation, a purely generational approach to the repatriates' bilingual language use may fail if patterns of their bilingual acquisition are not taken into consideration. I will demonstrate that although the immigration generation may be a reliable predictor of the bilingual ability, it is often overridden by specific paths of language development, particularly prior to immigration. I will therefore describe the respondents as a group and individually.

I will start by embedding Russian-German repatriation in a context of immigration to Germany from the Soviet Union and its successor states. I will then lay out the sociolinguistic history of Russian Germans in the twentieth century,

[1]Some scholars, e.g., Travis & Torres Cacoullos (2016), refer to this type of data as "spontaneous", or "naturally occurring", bilingual speech.

focusing on the development of their languages and the roles thereof in the community. I will also outline the considerations involved in selecting this group for data collection and particularly its members of the so-called intermediate-generation. The subsequent section will be concerned with the Russian-German youths and young adults who participated in the research as respondents and will then dwell on the respondent selection criteria and the process of respondent recruitment. I will then present the participants as a group and describe each of them in relation to the social networks, in which the recorded conversations took place. These recorded speech samples formed the corpus that was subject to analyses reported in the subsequent chapters of this book. The final section will close this chapter with a conclusion summarising the main aspects of the research participants.

3.1 Immigration to Germany from the Soviet Union and its successor states

The political and economic changes in Eastern Europe and Central Asia in the 1980s and 1990s, including perestroika, the dissolution of the Soviet Union as well as the economic policies of liberalisation in the post-Soviet successor states, led to social instability and economic hardships. These conditions sparked massive migration flows within and from the former Soviet Union. Most, but not all migration flows of the time belong to diaspora migration (Heleniak 2003). The beginning of emigration in the perestroika period was marked by liberalisation of the Soviet emigration policy in 1987, which had as a consequence that particularly representatives of the German and the Jewish diaspora started leaving the Soviet Union for permanent residence abroad. This process was facilitated by the immigration policies of the countries of their historical origin, which granted them a privileged status for immigration (cf. de Tinguy 2003). Crucially, in addition to the economic instability of perestroika, members of these groups faced political and economic discrimination in the pre-perestroika period. Along with these two groups, representatives of other ethnicities also contributed to permanent emigration from the Soviet Union and the post-Soviet successor states in the wake of socio-economic uncertainties of the 1990s. Most of them settled in the countries of Northern and Western Europe, North America and Oceania. Among the immigration countries, Germany experienced the largest influx of former Soviet citizens; its extent was paralleled, but not equalled by Israel.[2] Overall, three

[2]The number of ethnic Germans who immigrated to Germany from the Soviet Union and the post-Soviet successor states between 1988 and 2000 amounts to approximately two million

groups of individuals contributed to the immigration from the Soviet Union and the post-Soviet successor states: ethnic Germans, persons of Jewish origin and members of other ethnic groups (cf. Brehmer 2007).

Ethnic Germans from the Soviet Union and its successor states constitute the largest group of immigrants in Germany. The strong migration flow can be explained by the (West) German policy to regard all ethnic Germans living in Central and Eastern Europe as potential citizens. This policy was based on the (West) German Expellees' and Refugees' Law (*Bundesvertriebenen- und Flüchtlingsgesetz*) of 1953 as well as on an extensive interpretation of the Constitution (*Grundgesetz*) of 1949 (Article 116), which defines as citizens not only former citizens of Germany (within the borders of 1937), but also ethnic German refugees on the West German territories (Münz 2003). In 1957, ethnic Germans in the Soviet Union, but also Albania, Bulgaria, China, Czechoslovakia, Hungary, Poland, Romania and Yugoslavia were officially proclaimed German nationals and began to be referred to as *Aussiedler* 'repatriates'. Upon being granted, the *Aussiedler* status guaranteed them the same access to benefits as had been conferred on post-war expellees as well as the right to acquire German citizenship immediately after arrival to (West) Germany (Münz 2003). Already between 1950 and 1988, 62,023 persons of German descent left the Soviet Union for residence in Germany, but the mass emigration of Germans from the Soviet Union began after the liberalisation of the Soviet emigration policy in 1987. The following decade saw an unprecedented immigration flow of approximately two million *russlanddeutsche Aussiedler* ('German repatriates from Russia') and accompanying family members without German ancestry (Lederer 1997). After a change in the legislation, repatriates who emigrated to Germany after January 1st, 1993 began to be officially called *Spätaussiedler* 'late repatriates'. A more widely used term is *Russlanddeutsche* 'Russian Germans' (with *russkie nemcy* being its Russian equivalent, cf. Meng & Protassova 2017). Because this term is a self-designated name, I will use it throughout this thesis.

Although the term "Russian Germans" may imply that members of this group have a full command of the German language, including German dialects spoken in the former Soviet Union, this may not always be the case. As has been widely reported, Russian Germans exhibit varied patterns of bilingual language acquisition and use prior to emigration (Berend 1998, Meng 2001, Riehl 2017, Worbs et al. 2013). In this regard, a differentiation ought to be made between the pre-war generations and the post-war generations: while the pre-war generations were

people (Lederer 1997), whereas Israel received around one million of Jewish immigrants from these countries in the same time period (Tolts 2009).

German-dominant at the time of emigration, the post-war generations' primary or only language was in most cases Russian (see section §3.2 for further details). Notably, the post-war generations made up the lion's share of Germany's Russian Germans in 2011[3] (Worbs et al. 2013: 41). This fact allows to conclude that the majority of Russian Germans can be well considered part of Germany's Russian-speaking community (cf. Meng 2001, Roll 2003, Brehmer 2007).

Jewish immigrants constitute the second group of Russian-speaking immigrants in Germany. The resolution of the Conference of the Interior Ministers of the Federal States (*Innenministerkonferenz*) from January 9th, 1991, added persons of Jewish descent from the former Soviet Union and their family members to the list of quota refugees. This measure, in connection with the Law Relating to Humanitarian Aid for Refugees (*Gesetz über Maßnahmen für im Rahmen humanitärer Hilfsaktionen aufgenommenen Flüchtlinge, HumHiG*), enabled persons of Jewish descent from the post-Soviet states, excepting the three Baltic states, to immigrate to Germany. Hence, 209,134 Jewish immigrants and their family members from the post-Soviet states came to Germany between 1993 and 2018 (BAMF 2020: 96). The addition of 8,535 persons of Jewish origin who immigrated to Germany before 1993 results in a total influx of 217,669 people (BAMF 2020). As Russian is either the first or the primary language of Jewish immigrants in Germany (cf. Brehmer 2007: 166), they are considered part of Germany's Russian-speaking community in its full amount (for an overview and further literature on Jewish immigrants in Germany, see Remennick 2017).

The last group of Russian-speaking immigrants includes citizens of the post-Soviet successor states who are living in Germany. Whereas the majority of them are foreign nationals, some of them have in the meanwhile acquired German citizenship. These individuals include au pairs, labourers, students, scientists, spouses in mixed marriages, etc. The varied legal and social status of these migrants challenges their systematic quantification (cf. Brehmer 2007: 166). Table 3.1 reports the numbers of citizens of the post-Soviet states living in Germany in 2018. As can be seen from the table, Russian citizens make up the largest part of migrants from these countries. As in the case of the aforementioned Jewish community, Russian is the first or the second language for Russian citizens (cf. Marten et al. 2015). Regarding the citizens of the other states listed in Table 3.1, we cannot rule out the possibility that citizens of Belarus, Ukraine and Kazakhstan have a fluent command of Russian and citizens of Moldova and the Transcaucasian states have at least some knowledge of Russian (Gasimov 2012b).[4] The rea-

[3] 2011 was the second year of the data collection in this project.

[4] The volume edited by Gasimov (2012a) offers an outline of Russification processes in the Russian empire and the Soviet Union, covering their sociohistorical aspects.

son for this assumption is the Soviet language policy to promote Russian as "a single language in the formation of a unified, industrialized nation state" (Grenoble 2003: 1). According to the 1989 Soviet census, 81 per cent of the population reported fluency in Russian as either their first or their second language, although Russians comprised only one half of the population (cf. Grenoble 2003: 2). We could thus assert that at least for some part of non-Russian immigrants from the post-Soviet successor states, Russian may be a lingua franca for communication between fellow immigrants (cf. Levkovych 2012). Yet, as Brehmer (2007: 167) correctly notes, it is hardly possible to draw valid conclusions about language use from governmental statistics, such as in Table 3.1, he therefore considers nationals of the post-Soviet states living in Germany, with the exception of Russian citizens, as potential speakers of Russian.

Table 3.1: Nationals of the post-Soviet states (without the Baltic states) living in Germany in 2018 (adopted from BAMF 2020: 276–278).

Country	Number of individuals
Russia	254 325
Ukraine	141 350
Kazakhstan	46 740
Armenia	27 275
Azerbaijan	26 270
Georgia	25 775
Belarus	22 980
Moldova	20 375
Total	565 090

Overall, in the period between 1952 and 2018, at least 2.8 million Russian speakers migrated from the Soviet Union and its successor states to Germany, of whom only 2.5 per cent arrived in Germany before 1988. These migrants include 2.5 million of *(Spät-)Aussiedler*, 17.5 thousand of Jewish immigrants and 254 thousand of Russian nationals living in Germany. (Together with the citizens of Ukraine, Kazakhstan and Belarus registered in Germany, a total of at least 3 million people were living in Germany in 2018 who were likely to have a fluent command of Russian.[5]) Russian speakers thus constitute one of the largest linguistic minorities in Germany. However, it should be kept in mind that Germany's Russian-speaking

[5]This count could not consider the former nationals of the Soviet Union and its successor states who have acquired the German citizenship as well as the deceased persons.

community is highly heterogeneous, consisting of persons with diverse social and ethnic backgrounds (although being united by similar cultural experiences in the Soviet Union and its successor states, cf. Gasimov 2012b, Levkovych 2012). On this account, the case studies reported in the subsequent chapters investigate the speech of only one group, i.e., *russlanddeutsche (Spät-)Aussiedler*, being by far the largest group among Germany's Russian speakers.

3.2 Russian Germans and their languages prior to emigration

Most Russian Germans arrived in Germany through 1990s and the beginning of 2000s. Prior to their emigration, their linguistic situation was similar to that of other linguistic minorities in the Soviet Union and its successor states. Berend (1998) describes it as stable German-Russian bilingualism, involving regular code-switching and code-mixing (p. 3). In a language proficiency survey involving 130 ethnic Germans from the post-Soviet states, she finds that most of them were proficient in Russian and German, usually a German dialect, but on rare occasions standard German. However, as was mentioned in the previous section, their language repertoires varied considerably, depending on the generation to which they belonged. Several studies (e.g., Dietz & Hilkes 1993, Berend 1998, Meng 2001, Riehl 2017) have noted that Russian Germans born before World War II (WWII) reported and demonstrated the ability to speak a Russian German dialect and Russian, whereas the post-war generations had no or a very limited command of Russian German, and some knowledge of Standard German, which they learnt at school. Berend (1998) observes a correlation between her respondents' age and their multilingual competence: the younger the respondents, the higher their proficiency in Russian and Standard German and the lower their knowledge of a German dialect (p. 55, cf. Table 3.2). The speakers' competence in Standard German also positively correlated with the duration of education (pp. 56–57). It must be noted in this regard that 8-year secondary education had been extended to 10-year education throughout the country by the 1960s, and we can certainly assume that in areas with large proportions of ethnic Germans, they learnt Standard German as a foreign language at school. However, considering the fact that foreign language teaching in the Soviet Union was aimed at teaching only basic comprehension skills, school leavers were barely fluent in standard German (cf. Ivanova & Tivyaeva 2015: 309).

Two main factors contributed to the pre-war generation's ability to speak at least one variety of German: for one thing, most Russian Germans belonging to

these generations grew up in German-speaking cities and settlements (cf. Berend 1998); for another thing, the Soviet language policy of the 1920s allowed and even prescribed the use of German as the language of administration and a means of instruction in the public schools of the country's German-speaking areas (such as the Volga German Autonomous Soviet Socialist Republic), though sometimes both German and Russian were used in these domains (cf. Meng 2001, Riehl 2017). However, the situation changed radically already in the late 1920s and early 1930s (cf. Mukhina 2007): German clergy fell victim to the Soviet policy of elimination of religion, and some wealthy German farmers and well-to-do peasants, who were then referred to as kulaks and considered class enemies, were prosecuted and banished to Kazakhstan or Siberia in consequence of the dekulakisation policy. Further repressions followed: already before WWII, German communities in the European part of the Soviet Union, being accused of collaborating with Nazi Germany, were transported to Siberia and Kazakhstan and forced to live there in special guarded settlements (*Kommandaturaufsicht*); during WWII, many of the deportees, mostly males but also females, were conscripted into labour columns (*Trudarmee* or *Trudarmija*) and sent to labour camps in the north of Russia and the Ural region. Although the police surveillance in the guarded settlements was abandoned in 1955, it is not until 1964 that Soviet Germans were partially rehabilitated and allowed to return to their former homes. Since the deportations of the 1930s, Russian began to oust German from the public sphere. Hostile attitudes toward Germans and all things German, including the language, persisted in the Soviet Union for decades, long after the end of WWII, and resulted in the socio-economic discrimination against the German minority. Although the situation eased slightly during the 1960s, the German language, being still stigmatised, was relegated to the private domain (Berend 1998: 49; Blankenhorn 2003: 21).

Another factor which facilitated the shift in language dominance in the afterwar generations is urbanisation. Berend & Riehl (2008) report that 85 per cent of Russian Germans were living in rural areas in 1926, and the proportion sank to 50 per cent in 1979 (p. 23). This statistic is corroborated by Meng's (2001: 83) study of Russian Germans' linguistic biographies: of 42 interviewed Russian Germans born between 1948 and 1972, only one person reported growing up in a German settlement, 19 interviewees grew up in cities, and 22 respondents spent their childhood in multilingual settlements. Hence, the aforementioned historical events, particularly the dissolution of the established German-speaking settlements, and socio-economic factors directly affected Soviet Germans' language use: after WWII most of them were living outside German settlements and thus "in an unstable linguistic situation, which [in many cases] led to a gradual attrition of German and to its eventual loss" (Berend 1998: 20, my translation).

The correlation between the generation to which a Russian-German person belongs and a specific language acquisition pattern holds for all the recently reported surveys of Russian Germans' language use, even despite some heterogeneity across the studies with regard to two aspects. Some researchers (e.g., Berend 1998, Meng 2001) collected data in Germany, whereas other scholars (e.g., Blankenhorn 2003, Riehl 2017) gathered data from respondents in Russia. Moreover, studies vary slightly as to the classification of generations, or age groups, particularly with regard to the parameter "birth year". For example, Meng (2001) identifies six age groups and specifically differentiates between two groups, or generations, of children: "children of preschool age" (*Vorschulkinder*), born between 1984 and 1992, and "school children" (*Schulkinder*), born between 1976 and 1986. In contrast, Riehl (2017) distinguishes between four generations, the youngest of which includes Russian Germans born after 1975. In Table 3.2, I reproduce Riehl's (2017) approach to relating the four generations of the interviewed Russian Germans to reported language acquisition patterns.

Table 3.2: Language competence across generations in the Russian German diaspora in Russia (adopted from Riehl 2017).

	1. Generation born before 1932	2. Generation born between 1932 and 1952	3. Generation born between 1952 and 1975	4. Generation born after 1975
L1	Russian German dialect	Russian German dialect with markers of attrition; Russian	Russian; (passive competence of German dialect)	Russian
L1	Standard German	Standard German (rudimentary)		
L2	Russian (partly as an interlanguage)		Standard German as an inter- language	Standard German as an inter- language at various levels

The gradual shift from German to Russian, which is characteristic of the Russian German post-war generations, is easily identifiable in the table. Interestingly, although the generation born between 1932, namely before WWII, and 1952 is distinguished by simultaneous bilingualism, they, in their turn, "did not

actively transmit German to their children" (Riehl 2017: 22). The next generation, i.e., generation three, roughly corresponds to Meng's (2001) generation four, called "young parents, born between 1948 and 1972, mainly between 1955 and 1969" ("Junge Eltern, geboren zwischen 1948 und 1972, meist zwischen 1955 und 1969", p. 20). Albeit half of this generation reports German (usually a dialect) to be their first language, and one third claims to have simultaneously acquired German and Russian in their early childhood, 93 per cent of Russian Germans belonging to this generation asserted a better command of Russian than German prior to emigration (Meng 2001: 36). It therefore comes as no surprise that they spoke almost exclusively Russian with their children (Meng 2001: 35). Their children – they correspond to Meng's (2001) group two ("schoolchildren") and Riehl's (2017) generation four, born after 1975 – as is shown by these studies, almost invariably acquired only Russian as their first language, albeit they may have been exposed to either a German dialect or Standard German in their childhood (cf. Meng 2001: 35). Indeed, four of my research participants who repatriated at the age of seven or older have reported remembering their speaking German (probably a dialect) in their families. It is impossible to say though what their proficiency in that language, or its dialect, was and how the input they received could be described in terms of its quality and quantity. Additionally, more than a half of my research participants attended secondary school in the Soviet Union or in any of the post-Soviet successor states prior to their emigration to Germany. Therefore, they might have learnt German as a foreign language and may thus have had (limited) exposure to standard German. However, considering the duration of the secondary education that they received in their countries of origin and the aforementioned fact that the focus in foreign language teaching up to the mid 1990s had been largely on passive skills, we may hardly expect that they were fluent in standard German. The members of this group were either children or adolescents when they resettled in Germany with their parents. In the next section, my focus will be on these speakers as a specific immigrant group.

3.3 Research participants: Russian-German youths and young adults

3.3.1 1.5-generation Russian-German immigrants

As outlined in the previous section, young Russian Germans, born in the late 1970s, 1980s and early 1990s, were socialised in Russian (cf. Meng 2001: 106) and integrated into the Russian-speaking community (cf. Roll 2003: 275). As most

young Russian Germans were fully accepted in the social contexts prior to emigration, in the majority they had "neither any knowledge of German language and culture nor of modern or even postmodern German society" (Roll 2003: 272; for further details, see Dietz & Roll 1998). Learning the German language was thus the main challenge that they faced upon emigration and a key factor influencing their socioeconomic status and prospects. This challenge was even greater for children, who had not yet developed a full command of Russian, their first language (cf. Meng 2001: 106–152). This demanding situation is typical for immigrants of the intermediate, or 1.5, generation. These are individuals who moved to a host country as children or adolescents (between the ages of 5 to 18), usually following their parents and/or other family members. (Backus 1999b, 2006 uses the term "intermediate generation" to refer to these individuals, cf. chapter (1.3.1), whereas Remennick 2017, employs the shortcut "1.5".) According to Remennick (2017), no agreement exists among scholars on the age bracket defining the 1.5-generation, but most researchers acknowledge the unique character of immigrant experience typical of this generation: it differs from the experience of both the first generation (their parents) and the second generation (children born in the host country).

1.5-generation immigrants are well suited to collecting bilingual speech samples because most of them are proficient in both languages (cf. Backus 1992: 43; Halmari 1997: 36–38; Boumans 1998: 160–167). For example, Turkish 1.5-generation immigrants in The Netherlands have been found to have neither a preference for Turkish, the origin country language, nor a preference for Dutch, the host country language (Backus 2006: 201). This view is very much in line with research on nativelike attainment, according to which nativelikeness, defined as L2 learners' performance that corresponds to the range observed with native subjects, is found even among learners whose age at immigration is in the late teens (Birdsong 2009: 121). Another argument in favour of collecting speech of 1.5-generation immigrants for investigating code-mixing is a relatively high frequency of code-mixing in their speech. For example, in a 1999 paper Backus reports that of the three generations of Turkish immigrants in The Netherlands, representatives of the 1.5-generation produce twice as many mixed sentences as the second-generation immigrants and about three times as many as the first-generation immigrants (p. 263).[6]

These observations motivated the data collection for building a bilingual Russian-German corpus in the first place. The goal was to record samples of natu-

[6]Goldbach (2005) shows that the speech of the first-generation immigrants (Russian and Ukrainian nationals without German ancestry living in Berlin) may also involve intensive code-switching and code-mixing.

rally occurring bilingual speech produced by Russian Germans of the intermediate generation. The first phase in the process of gathering these data was the recruitment and selection of eligible individuals.

3.3.2 Recruiting and selecting research participants

The selection of participants was guided by four parameters: (i) immigrant generation, (ii) use of Russian on daily basis, (iii) social integration in Germany, and (iv) duration of residence. While the two former parameters introduced indirect means for control of proficiency in Russian, the latter two parameters sought to provide indirect means for control of their knowledge of German.

The key parameter for selecting suitable bilingual speakers was the generation of immigration: Russian Germans belonging to the 1.5 generation of immigrants were identified as eligible for participating in this research. This consideration rested on the foregoing observation that 1.5-generation immigrants are usually fluent in the origin country language. However, as origin country languages may be affected by attrition (on the development of Russian in Russian-German immigrants of the 1.5 generation, see Meng 2001, Meng & Protassova 2017), I introduced a further requirement, namely, regular use of Russian in daily life.

The third criterion pertained to the immigrants' social integration. The underlying assumption here was that participation in the host country's major social institutions, such as the labour force and educational institutions, enhances host country language proficiency. For example, in a study of the social integration of Russian Jewish immigrants in Israel, Remennick (2003) found that Hebrew competence correlated with the immigrant's occupation type: "respondents with skilled or professional jobs reported good/very good Hebrew three times more often than did respondents with unskilled jobs and *five times more often than did unemployed or retired respondents*" (p.32, my emphasis). Therefore, only if Russian German youths and young adults were attending or had graduated from German general secondary or vocational education and training institutions, they qualified as eligible participants in this research. Virtually all 1.5-generation Russian Germans automatically satisfy this criterion as the lion's share of them received general secondary education and another portion was involved in vocational education. Finally, to additionally ensure the respondents' proficiency in German, a further condition was introduced as a selection criterion: respondents were expected to have stayed in Germany for a period of approximately ten years, or longer. In a nutshell, in order to meet the selection criteria of this research, an ideal prospective participant was supposed to be an intermediate-generation Russian-German immigrant who had been staying in Germany for a period of

at least ten years, was receiving or had received secondary, and possibly vocational, or higher, education there and used Russian in their daily communication at home or in their social network.

In order to ensure the greatest possible degree of authenticity and naturalness in the recorded interactions, the respondents were selected as members of either existing social networks or naturally occurring social groups. This criterion was given primacy over the language biographical parameters introduced above when applied to informal multi-party conversations. Whenever the majority of bilinguals engaging in such a conversation satisfied the aforementioned language biographical criteria, the recording was included in the corpus.

The corpus falls into two parts, each corresponding to a different kind of setting in which the audio-recorded interactions took place. One type of interaction subsumed informal conversations recorded by recruited young adults in their social networks, whereas the other type covered peer group interactions among youths in a school setting. I will first dwell on recruiting the respondents who recorded naturally occurring conversations with their friends and family members and will then elaborate on selecting younger respondents, whose conversations were recorded in school.

Each individual from the first group received a written invitation to participate in the study. The invitation contained general information about the project and instructions for audio-recording of conversations. It specifically asked the respondents to record themselves and their communication partners in spontaneous everyday interaction, so as to emphasise the importance of recording conversations that reflect natural language use. Moreover, the invitation recommended the respondent to engage in conversation with their peers, i.e., friends and/or relatives of a similar age and with the same Russian-German background. It also instructed the participant to inform beforehand the conversation partners about the audio-recording and to obtain their consent on it. Finally, it guaranteed all the persons involved that the audio-recordings will be used only for research purposes and in an anonymous format.

A first attempt to recruit Russian German young adults consisted in sending the invitation to the Youth and Student Association of Germans from Russia (*Jugend- und Studentenring der Deutschen aus Russland e.V.*), which is the youth organisation of the Homeland Association of Germans from Russia (*Landsmannschaft von Deutschen aus Russland e.V.*), and the Youth Migration Service of the Caritas Association of Freiburg (*Jugendmigrationsdienst des Caritasverbandes Freiburg-Stadt e.V.*).[7] Whilst my request addressed to the former organisation

[7]The mentioned organisations are all incorporated associations (*eingetragene Vereine, e.V.*).

remained without response, the latter organisation kindly provided me with a list of persons willing to participate in the study. Two respondents, Svetlana and Elena (here and below I use fictitious names), could be recruited in this way. Another participant, Tatyana, was recruited from my personal Hanover network. Notwithstanding our previous acquaintance, she remained unaware of my specific interest in gathering speech samples with code-mixing until the end of the data collection.

My efforts to recruit young Russian Germans in Lahr, a city in Baden with a large proportion of Russian Germans (cf. Roll 2003), were futile, despite the mediation of the local group of the Homeland Association of Germans from Russia. The contacted persons either felt reluctant to participate in the research study or, being second generation immigrants, demonstrated only a limited command of Russian and thus did not meet the study's selection criteria. However, Lahr's local group played a major role in establishing contact with Marina, an respondent from Villingen-Schwenningen. Finally, I distributed posters publicising the project on the University of Freiburg campus. Two students with the Russian-German background could be recruited thereby: Irina and Olga. They were both living in the aforementioned city of Lahr and were strongly rooted in the local Russian-German networks. With their help, overall six Russian-German young adults volunteered for data collection. Irina and Olga were involved in at least one audio-recorded conversation, which they had held with members of their social networks. A speech sample of another Russian-German bilingual, Julia, was taken from the Regensburg corpus of Slavic-German bilinguals (*Das Regensburger Korpus slavisch-deutscher Bilingualer*); in this corpus, collected by Grillborzer & Meyer (2008-2009), each speech sample is supplemented by relevant sociolinguistic information about the speaker.

As regards the other group, i.e., Russian-German youths, they were all senior students of *Die Kaufmännischen Schulen/das Integrierte Berufliche Gymnasium Lahr* ('the commercial schools/vocational secondary school of Lahr'). At this level of education, completion of a second foreign-language programme is compulsory. In order to pass this course successfully or with little effort, some Russian German students, although they may have acquired Russian as an L1, chose Russian as their second foreign language subject. Yet, other Russian German students elected the Russian course for further reasons. Some sought, for example, to acquire literacy in Russian and to thus enhance their language proficiency. The identified motivations should however not be conceived of as mutually exclusive. Although many Russian German students had acquired Russian as an L1, their proficiencies in Russian differed. This variability is presumably determined by a shift, at least in some of the individuals, to German as a primary language.

The director of the school and the Russian teacher allowed me to establish contact with the students and engage them in an elicitation task. The task was to be carried out in freely formed discussion groups of three or four students. Free choice of partners allowed the students to consider the natural bonds and affiliations existing between them while forming groups. This measure was intended to facilitate lively and natural group discussions. After the groups were formed, the students were confronted with the task, which consisted in comparing and contrasting aspects of Russian and German cultures. The group discussions, stimulated by a set of prepared questions, were recorded for subsequent analysis. The questions contained mixed utterances such as *Kak možno ocharakterisovat' Russlanddeutsche doma im Vergleich zu den einheimischen Deutschen?* 'How can Russian Germans be characterised at home when compared to indigenous Germans?' (the first three words of the question as well as the word *doma* 'at home' are Russian), or *Silvester i Weihnachten, was wisst Ihr von den Traditionen in Russland und Deutschland? Kak prasdnujut Russlanddeutsche simnie prasdniki?* 'New Year's Eve and Christmas, what do you know about the traditions in Russia and Germany? How do Russian Germans celebrate the winter holidays?' (the first sentence is in German except for the Russian coordinator *i* 'and', the second sentence is in Russian except for the autonymic ethnonym Russian Germans).

The analysis revealed that six of the thirteen students had high fluency in Russian and frequently switched the languages during the group discussions. All of them had learnt Russian in early childhood, although the extent of their literacy was often limited to the input received from the Russian class at school. With the teacher's permission, I made several visits to the school, during which I engaged with these selected students in group conversations by asking them questions but leaving room for extensive peer interaction. Crucially, my speech, like theirs, was distinguished by extensive code-switching and code-mixing. In the course of the conversations, the students frequently left the given subjects and started to talk about other things. Hence, these interactions are comparable with conversations recorded by young adults in their social networks.

Upon the recordings, respondents were asked to complete a paper self-report questionnaire, which was designed in order to gather basic sociolinguistic information. The requested information concerned the participants' migration and language acquisition histories, their language preference as well as language use in their social networks. Since the questionnaire data reported below are based on self-report, as any other self-report based data, they should be regarded with caution. Although self-report measures assume that people answer the questions honestly, respondents may sometimes be inclined to give incorrect responses, wanting to make a good impression (cf. Kassin et al. 2012: 208). For example,

already in her (1980b) pioneering paper, Poplack showed that bilingual speakers occasionally over- and underrated their language proficiency, even though the great majority of the respondents provided self-report estimates which were compatible with the objectively observed proficiency rates. However, because the focus of the present research is on explaining patterns of Russian-German code-mixing by relating them to facts of usage, cognitive processing and structural regularities, rather than social factors as such, the reported survey results are intended as a demonstration of the respondents' sociolinguistic backgrounds, which serves as a basis for comparing the investigated group with other bilingual communities, as to their preferences for specific patterns of bilingual speech, in general, and code-mixing, in particular.

The remainder of this chapter entails a presentation of the questionnaire results. It first characterises the study participants as a group and then introduces them as members of their social networks.

3.4 Research participants as a group

This section presents and compares the respondents in terms of their age, migration history and linguistic memories about their language development. All the respondents' names are fictitious. It is furthermore necessary to note that all the respondents, with one exception, are female. This situation was not intended and owes to the strenuousness of the recruitment process.

3.4.1 Age and migration history

As detailed in Table 3.3, the respondents were youths and young adults, with their age ranging between 18 and 35 years (mean 25 years). Their age at immigration varied from 4 to 21 years. Those individuals who were 7 to 18 years of age at the time of immigration qualify as immigrants of the intermediate generation. These respondents constitute the majority of the bilinguals sampled in the corpus. Among the few exceptions to this tendency are three first-generation and two second-generation immigrants. The former group comprises three participants who moved to Germany at the age of 19 to 21 years. The decision to include these speakers in the corpus rested on two considerations: First, the period of their residence in Germany had exceeded ten years by the time of recording. Second, all of them were fluent in German: while Valentina and Marina reported growing up bilingual, Inna developed her proficiency in German after the immigration and that enabled her to enrol at the University of Freiburg. The latter group includes two second-generation immigrants, Larisa and Alex. Larisa was born in

Germany into a Russian German family, whereas Alex arrived in Germany being a four-year-old. Their speech samples were included in the corpus because, being rooted in the Russian-German community of Lahr, they still maintained Russian as a community language and regularly used it in their networks. Regarding residence duration, one respondent, Rita, did not satisfy the study's residence requirement, according to which each respondent was supposed to have stayed in Germany for approximately 10 years. The reasons for including her in the corpus are discussed in conjunction with her sociolinguistic profile in §3.5.

Table 3.3: Research participants by age at immigration.

Participant's fictitious name	Age at recording	Age at immigration	Duration of residence	Place of living
Larisa	21	local-born	21	Lahr
Alex	20	4	16	Lahr
Olga	24	7	17	Lahr
Alina	24	8	16	Hanover
Vika	20	8	12	Lahr
Vera	19	8	11	Lahr
Svetlana	27	10	17	Freiburg
Nataša	24	10	12	Freiburg
Elena	28	11	17	Freiburg
Ira	27	11	16	Freiburg
Nadja	19	11	8	Lahr
Julia	21	12	9	Regensburg
Tanja	30	14	16	Hanover
Rita	18	15	3	Lahr
Irina	27	17	10	Lahr
Olesja	30	18	12	Lahr
Inna	29	19	10	Lahr
Valentina	31	20	11	Lahr
Marina	35	21	14	Villingen-Schwenningen

As to the geographical distribution of the bilinguals sampled in the corpus, it follows from the table that more than half of them were living in Lahr, a city with a high proportion of Russian German repatriates. The other speakers represented in the corpus were residents of other German cities.

3.4.2 Language acquisition history

The study participants report learning Russian as their first language or one of their first languages. As to German, the situations and ages of its acquisition varied across the participants. These aspects of the acquisition of German are given in Table 3.4. The participants were asked whether they remembered speaking German by the age of seven years. Of 19 participants, five individuals reported having an oral proficiency in German at that age, most probably in a Russian-German dialect. They learnt German through the natural interaction with their parents (referred to in the table as "family") and can thus be considered early bilinguals. The others had no memories of speaking German at the age of seven years, although five respondents (marked by asterisks in the table) remembered hearing single German words, sayings or songs. One respondent reported learning German before immigration through interactions with her relatives and friends ("environment" in the table).

The respondents who were living in their home countries at the age of eleven or older were learning German as a foreign language in the classroom ("GFL" in the table). After migration to Germany, the participants either acquired German through natural interactions with speakers in their environment or through visiting specific language courses ("GFL" in the table). A comparison of Tables 3.3 and 3.4 makes two tendencies visible: First, individuals who moved to Germany either at a very early age or at least 16 years before the study took place, i.e., in the early 1990s, manifest the tendency to acquiring German through natural interactions. Second, individuals who immigrated to Germany at the age of seven years or older and less than 16 years before the study took place, i.e., in the late 1990s, tend to have learnt German in classroom context.

In a nutshell, the participants of the study manifest the following acquisition patterns for German: five respondents learnt German in the family and named it one of their first languages, three respondents learned German in an informal context beyond the family, and ten respondents acquired German in a primarily formal context. However, as we can see in Table 3.4, the identified patterns of acquisition sometimes overlap and the exposure to German in the early childhood may also play a role in the later acquisition of the language. I therefore present detailed sociolinguistic backgrounds of the individual respondents below.

3.5 Participant subgroups

This section draws a portrait of each of the 19 respondents by making use of the sociolinguistic information collected by the aforementioned questionnaire.

Table 3.4: Research participants' acquisition of German by situation types prior and upon immigration. (The typical situations in which the respondents learnt German include the family [in early childhood], the natural environment and classes of German as a foreign language (GFL) at school in their country of origin, or in a language school in Germany. The asterisk marks respondents with memories of fragmentary exposure to German in early childhood. The ReBiSlav corpus did not provide information on Julia's linguistic development.)

Participant's fictitious name	Age at immigration	Acquisition prior to immigration	Acquisition upon immigration
Larisa	local born	family	
Alex	4	family	family
Olga	7		environment
Alina	8	*	environment
Vika	8		GFL
Vera	8		GFL
Svetlana	9		environment
Nataša	10	family	family
Elena	11	environment, GFL	environment
Ira	11	GFL	GFL
Nadja	11	GFL	GFL
Julia	12	GFL	?
Tanja	14	GFL*	environment
Rita	15	GFL	GFL
Irina	17	GFL*	GFL
Olesja	18	GFL*	GFL
Inna	19	GFL*	environment
Valentina	20	family	
Marina	21	family	

3.5.1 Irina and Olga

Irina and Olga responded to the announcement poster distributed at the university campus Freiburg. The young women knew each other prior to the recordings. At the time of data gathering they were enrolled in the Slavic Department of the University of Freiburg. This fact may be interpreted as a sign of their affinity towards the Russian language and culture. Some of their audio-recordings contained conversations between each other, whereas the others included conversations with members of their individual social networks.

Both Irina and Olga were born into mixed families: their fathers are Russian Germans, their mothers are Russian. As would be expected in this case, their parents spoke Russian to them prior to emigration. Although Irina reported being exposed to fragmentary German input in form of chunks, i.e., individual lexical items, phrases, and songs, she remembers speaking only Russian before the age of seven. Olga, in her turn, asserted that before the age of seven she spoke only Russian. Although Irina and Olga's early language biographies parallel each other to a large extent, their migration histories vary. Olga left Russia in 1995 at the age of seven years, whereas Irina moved to Germany in 2002 at the age of 17 years. Hence, upon arrival Olga learnt German from friends, relatives and at school, i.e., in her natural environment, whereas Irina had to take a German course before attending public school. Irina considered herself more proficient in Russian than German, whereas Olga's estimation was exactly the opposite. However, in daily-life situations it was easier for both of them to interact in German. Both Irina and Olga were living in Lahr. Being rooted in the large local Russian-speaking community, each of them audio-recorded several hours of conversation within their social networks.

Irina's Russian is close to standard, and her German is fluent. She has high language awareness and actually tries to avoid language mixing, obviously due to the need to speak standard Russian at the university. Although mixing and switching do occur in her speech, especially when her conversation partners mix languages, she sometimes pauses in the middle of a sentence and looks for a translation equivalent. Olga's language development differs from Irina's. Her German is native-like, but her Russian is also fluent. Her bilingual speech is characterised by frequent code-switching and code-mixing, even when she speaks to Irina.

The bulk of the recordings comprises one-on-one conversations between the two young women, in which they talk about their university and work life as well as mutual acquaintances and friends. Some of their conversations cover such topics as the Russian identity, the histories of Russia and Germany as well as

current developments in the Russian politics. One of their interactions involves one of their fellow students. In this conversation, the primary topic is university-related matters and the speakers code-mix a lot. Apart from these conversations, Irina and Olga audio-recorded several interactions in their individual networks. Irina's recordings include a group conversation with Olesja and Valentina, both of whom were also Lahr residents.

3.5.2 Olesja and Valentina

Olesja and Valentina arrived in Germany in 1999 and 2002 at the ages of 18 and 20 years respectively. Their parents are Russian Germans and they grew up in similar socio-cultural environments. Although Olesja's home town Omsk exceeded Valentina's home town Astana in population size at the time when the women were living there, both cities had a fairly equal proportion of Russian Germans. Yet, despite this fact, Olesja's parents decided to speak only Russian to her, whereas Valentina received an equal amount of input in both German and Russian from her parents and reports growing up bilingual. However, Olesja remembered being exposed to German, even though this input was fragmentary. Upon her arrival in Germany, unlike Valentina, she had to attend a German course.

From the recorded conversation, it can be concluded that Russian is the dominant language for each of the interlocutors. Throughout the conversation Russian is more frequent than German, although both languages are widely used. Olesja's and Valentina's Russian may be described as colloquial with a few non-standard sprinklings. As to German, Valentina appears to be more secure about it than Olesja. She also uses more German in the conversation than Olesja and Irina taken together. Nevertheless, Olesja's productive proficiency in German is adequate, which may be due to the fact that she had lived in Germany for twelve years before the recording took place and had been working as a sales person. The three conversation participants – Irina, Olesja and Valentina – regularly switch to German when attending to their children, who speak German most of the time.

3.5.3 Olga and Inna

Olga recorded several conversations with her close friend Inna, who was also a Lahr resident. Like Olga, Inna was born into a mixed family: her mother is a Russian German, her father is Russian. Unlike Olga, who arrived in Germany in 1995 at the age of seven years (see above), Inna immigrated as a young adult of

19 years in 2002. Therefore, they demonstrate opposing trends in language proficiency: Olga regards German as her strongest language, whereas Inna considers Russian to be her better language. Nonetheless, before the age of seven Olga and Inna report Russian to be their primary language for speaking, despite they had received a minor portion of German input.

By the time of recording, Inna had lived in Germany for ten years and identified herself as a Russian. In her daily life, the use of each language is often compartmentalised: Russian is used in the affective domain, i.e., with the family members, relatives, partner and Russian-speaking fellow students, and German is used at the university, at work, but occasionally also with Russian-speaking friends.

In the recorded conversations they discuss their university life, plans for the future and exchange opinions about mutual acquaintances and friends. Hence, the conversations abound in topic-related code-switching as well as code-mixing.

3.5.4 Tanja and Alina

I had known Tanja before conducting this study, and after I asked her to audio-record some conversations with her sister Alina, she immediately agreed. Tanja and Alina were born in Uzbekistan to Russian-German parents in 1980 and 1986, respectively. The family language was Russian, but the children also heard some German (presumably a German dialect), as used by their grandparents and other relatives. When the family left Uzbekistan in 1994, Tanja was 14 years of age and her sister was eight years old. The different ages at immigration influenced (and most likely still influence) the sisters' language use and ethnic identities. The elder sister identified herself as a Russian German, whereas the younger sister felt as a German. In her network, Tanja used both languages equally, unlike Alina, for whom the use of Russian was restricted to her parents and grandparents. With her sister, she reported speaking both languages, switching them back and forth. As for their proficiency assessment, Tanja considered herself to be equally fluent in both languages, whereas her sister rated her oral proficiency in German much higher than in Russian. However, the recorded conversations reveal extensive code-switching and code-mixing in each of the sisters' speech. They talk about household chores, flat renovation, their children and mutual acquaintances.

3.5.5 Marina

Marina is the only respondent from Villingen-Schwenningen, a city at the historical border between the regions of Baden and Swabia. Marina recorded an

interaction at the nursery school where she was working as a teacher and an informal conversation with her Russian-speaking friends. As both of her friends did not fulfil the study's requirement of residence duration, their contributions to the conversation were disregarded in the analysis. Prior to her emigration in 1999 at the age of 21 years, Marina lived in Kazakhstan, where she was born into a Russian German family. A Russian German dialect, standard German and Russian were spoken in her family. She thus grew up bilingual, acquiring Russian and standard German. At the time of the recording she considered herself a Russian German and evaluated her command of Russian as native-like and her proficiency in German as fluent. Based on the recorded conversations, her Russian may be described as a variety very close to the standard. She preferred speaking both languages without separating them as isolated codes, or, putting it in Grosjean's terms, using them in a bilingual mode (cf. Grosjean 1985). Interestingly, she used Russian with her Russian husband and spoke German with her children. In her social network, both languages were constantly in use. In the recorded conversation, she and her friends talk about their lives in Germany, their children and food.

3.5.6 Elena, Ira and Nataša

Elena's contact details was provided by the Youth Migration Service of the Caritas Association of Freiburg. Elena recorded several casual interactions with her friends Ira and Nataša. Elena's and Nataša's linguistic memories are similar in several ways: First, they were born into mixed families, with one of the parents being a Russian German and the other parent being a Russian. Second, they remembered only speaking Russian before they started school (although Nataša's mother spoke German to her daughter). Third, they were almost of the same age when they moved to Germany: Nataša was ten years old (year of immigration: 1997) and Elena was eleven years old (year of immigration: 1996). However, at the time of recording, more than fifteen years after their immigration, they manifested diverging tendencies in language use and bilingual ability: Elena reported a regular use of the two languages in her network and equal fluency in both of them, while Nataša tended to speak German slightly more frequently than Elena and regarded her command of German higher than her proficiency in Russian. As to Ira, her parents, though being Russian Germans, used only Russian as a family language. The family arrived in Germany in 1995, when she was eleven years old. Upon her immigration, Ira learnt German predominantly in a language course. According to her report, the extent to which she used Russian in her daily-life interactions prevailed the extent of German used in the same contexts. Although

Ira had lived in Germany for twenty years by the time of data collection, she still preferred Russian to German, but evaluated her proficiency levels in the two languages as equally high. In their conversation, Elena, Ira and Nataša talk about family matters and real estate prices in Freiburg. During their conversation, the friends have their children at the table receiving their meal and repeatedly attend to them.

3.5.7 Svetlana

As in the above case, I received Svetlana's contact information from the Youth Migration Service of the Caritas Association of Freiburg. At the time of the recording she was a single working mother with a preschooler and was about to marry her fiance, also a Russian German. Being pressed for time, or for any other reasons, she refused to record a casual conversation for me, but allowed me to converse with her. During our conversation, Svetlana (and I) were switching the languages back and forth; that was a natural and preferred way for her to speak Russian. Svetlana's family moved to Germany in 1994, when she was 10 years of age. During her childhood in Russia, her parents spoke to her Russian and German, and she claimed speaking ability in these languages by the age of seven years. Svetlana may thus count as a Russian-dominant bilingual. It is yet unclear whether the variety of German that she remembers speaking was closer to Standard German, or to one of the German dialects spoken in Altai Krai, Russia. At the time of the recording, Svetlana considered herself a German and evaluated her command of German as higher than her command of Russian. Although her Russian was strong, she had experienced several situations in her professional life in which she was unable to use it for professional purposes. This may have led her to think of her Russian competence as poor. In daily communication, Svetlana preferred speaking Russian and German in a bilingual mode. She reports using both languages for communication in most contexts.

As stated above, all younger participants of the study, except Julia, were senior students from Lahr. Julia's speech sample was taken from the ReBiSlav corpus (cf. Grillborzer & Meyer 2008-2009). Although the respondents from Lahr knew each other from the Russian class, each of them was linked with one of the two pre-existing groups of friends: Nadja, Rita, Vera and Vika, on the one hand, and Alex and Larisa, on the other hand. These bilingual speakers are described as members of the corresponding social network.

3.5.8 Nadja, Rita, Vera and Vika

The girls caught my attention already during the elicitation task by being conspicuously loud among their classmates while discussing the task questions. They turned out to be a tightly-knit, enclosed group who stuck together in and after school. The girls regularly met each other individually, or gathered as a group. They celebrated holidays together, and on summer holidays, some of them travelled together. In their clique, they cultivated an atmosphere of trust and emotional immediacy. As to the clique's language of interaction, Russian was unequivocally the primary language, but code-mixing was the norm. The topics of the recorded conversations range from school matters to their childhood experiences in the countries of the former Soviet Union and Germany to the Fukushima nuclear disaster.

Nadja, Vera and Vika were all born into mixed families, with one of the parents being a Russian German, and the other parent belonging to another ethnic group. In other words, they have the same ethnic background through one of their parents. By contrast, Rita's ties to Russian Germans are only through friends. Hence, she is the only respondent to contribute to the corpus who has no Russian-German ethnic background. Although Nadja, Vera and Vika spent their childhood in various places, even different countries (Nadja and Vera was born in Kazakhstan, and Vika in Russia), their language biographies and language learning memories are alike. All of them report learning Russian as their first language and not being exposed to German in their childhood. Yet, the three respondents also manifest differences. For example, their ages of arrival in Germany vary: Vera's and Vika's families moved to Germany in 2001 and 2000 respectively, when the girls were eight years old, and Nadja's family arrived in Germany in 2004, when she was eleven. This means that Nadja completed her primary education and the other two received some part of it in the countries of their origin in Russian. Upon immigration to Germany, they all attended German classes for repatriates. At the time of recording, the speakers rated their proficiencies in the two languages as follows: Nadja and Vika reported equally high proficiencies in both languages; Vera considered her German to be better than her Russian, which she rated as good. All of them stated that they preferred speaking the two languages in a bilingual mode, i.e., without separating them.

Rita was born into a Russian family, and arrived in Lahr following her mother. Rita moved to Germany in 2009 at the age of 15 years, but her first encounter with German took place some time earlier: she began learning it before leaving Russia. After English, German was the third language she learnt. By the time of recording, she had lived in Germany only for three years, but her German

was more than adequate, particularly in casual interactions. Rita's extremely fast integration into the local community contributed substantially to her fluency in German. Being a very communicative person by nature, not only she had started attending school from her first days in Lahr, but she had also became a core member of the described bilingual clique as well as a member of a sports club, and had a German-speaking partner for some time. In her after-school job, she was involved in communication with German-speaking customers and colleagues. At the time of data collection, her German was slightly dialect-coloured, partly because she had received much input from her stepfather and colleagues, who spoke the local Alemannic dialect to her. Rita's leading position in the group and her exceptional language ability were the primary reasons for including the recordings of her speech in the corpus.

3.5.9 Alex and Larisa

Alex and Larisa were classmates having a friendly relation without being close friends. They had similar language biographies but entirely different identities and attitudes towards Russian. Both of them were born into mixed families, Alex in Kazakhstan and Larisa in Germany. Both Russian and German were spoken in their families, so that the children grew up bilingual. Larisa is the only study participant who was born in Germany. Alex's family moved to Germany in 1996, when he was four. Therefore, they both count as second-generation immigrants. His parents saw no value in Russian, and German became the primary language at home. At the time time of recording, he considered himself as a German rather than a Russian German and his use of Russian was limited to the communication with his grandparents and a few friends. Unlike Alex, Larisa valued the Russian language and culture. She felt strongly attached to her Russian-German friends and partner and considered herself a Russian German. Larisa reported receiving much of her Russian input from her bilingual clique. Nevertheless, Larisa's and Alex's language proficiencies paralleled each other at the time of recording: they were evidently German-dominant and ranked their German skills higher than their Russian skills.

3.5.10 Julia

Julia is the only intermediate-generation immigrant from the ReBiSlav corpus (Grillborzer & Meyer 2008-2009). Of all the speakers sampled in the corpus, only her data could be used in this work. Julia's family moved to Germany in 1996, when she was twelve years of age. Russian was her first language, and she reports learning German in Germany. However, at the time of recording she was

a German-dominant bilingual, with a limited Russian proficiency. Nevertheless, she reported regularly using Russian in several contexts.

3.5.11 Summary

As can be seen from the linguistic portraits of the Russian-German bilinguals sampled in my corpus, their linguistic competences manifest similarities rather than differences, even though they represent immigrants of three generations: first-generation immigrants (two participants), 1.5-generation immigrants (14 participants) and second-generation immigrants (three participants). Crucially, all of the respondents demonstrated bilingual ability, being fluent in the host-country language and maintaining Russian in their day-to-day interactions. The intergenerational differences in my respondents' language competences and preferences appeared to be minor. This situation is attributed to two circumstances: the existence of tightly knit social networks and early bilingualism. Social networks and communal ties were particularly important for the two second-generation immigrants included in the corpus. Maintaining tight connections to Lahr's Russian-speaking community allowed them to be regularly exposed to Russian and to learn to use it in everyday situations. At the same time, two of the sampled first-generation immigrants reported acquiring both languages at an early age prior to their immigration to Germany. Such a case is generally rare in the context of immigration, but not uncommon in the context of repatriation, provided that the minority language was maintained. That is, the "generational" approach alone cannot account for the patterns of bilingual language use in a situation of repatriation. Therefore, an adequate description of Russian Germans' languages should consider both the generation of immigration and the individual paths of language acquisition.

3.6 Data

The data were collected in two types of setting. The lion's share of the gathered material included informal interactions audio-recorded by research participants in their social networks. These interactions occurred in the participants' everyday situations such as at home, in the street and in a restaurant as well as on train and car rides. Overall, the participants recorded 15.5 hours of naturally occurring speech. While most of the conversations of this type involve two participants, some of them are multi-party conversations; for example:

(1) (HO100712)

A: ty slyšala čto mara ma- mara maša s danikom *heiraten*?
you heard that Mara Mara, Maša with Danik get_marry

T: gde? kogda?
where when

A: v ijule zags, a v avguste *hochzeit*
in July registry_office and in August wedding

T: tak *kurzfristig*?
so short-term

A: ja sama tol'ko nedavno uslyšala. ona mne desjat' raz rasskazyvala
I myself just recently heard she me ten times told
što oni *heiraten* *dann doch* *nicht heiraten*
that they get_married then actually not get_married

T: èto oni rešili vot. mama, kak tam u maši s danikom?
PTCL they decided PTCL mother how there at Maša with Danik
kakogo čisla u nix zags? *standesamt*?
what date with them registry_office registry_office

M: čisla pervogo
date first

T: ((laughing)) *standesamt* kogda u nix?
registry_office when with them

M: aah ponjatija ne imeju. vy že mne govorili.
idea not I_have you PTCL me told

A: net.
no

A: 'Have you heard that Maša and Danik are getting married?'
T: 'Where? When?'
A: 'In the registry office in July, and their wedding is in August.'
T: 'At such a short notice.'
A: 'I've heard it myself only recently. She told me ten times that they are getting married and that they aren't.'
T: 'Well, they've decided now, you see. Mother, how is it with Maša and Danik? On what date do they have (the appointment in) the registry office?'

M: 'Around the first.'
T: 'The registry office, when do they have it?'
M: 'Ah, I have no idea. It is one of you who told me.'
T: 'No.'

This conversation begins as a two-party interaction and develops into a three-party talk. At first, Alina announces her sister Tanya that a familiar couple is getting married, and then Tanya consults their mother about the date of the marriage registration. The language of the conversation, although being basically Russian, is permeated by German single words and multi-word sequences. The speech is thus described as language mixing. At the same time, the passage contains an instance of code-switching: After using the Russian term for the registry office, *zags*, Tanya provides its German equivalent, *Standesamt*, in order for her mother to better understand what she is speaking about. This kind of reiteration is a typical example of code-switching. Crucially, the base language of each turn is Russian, and no change in the conversation's base language takes place between the turns. Even Alina's multiword switch *heiraten, dann doch nicht heiraten* 'are getting married and then they are not' at the end of her turn does not incite a shift in the language of the conversation. I therefore conceive of the language practice in this passage, just as overall in the corpus, as code-mixing, rather than code-switching.

The other type of data includes recordings of interactions made by the writer. With two exceptions, the conversations in which I was one of the participants or an observer took place in the school setting. A total of ten hours of speech was recorded in this type of participant constellation. All but one interactions recorded by myself were multi-party conversations. The following snatch of informal talk, involving Rita (Ri), Vera (Ve), Vika (Vi) and the writer (Re), illustrates this type of data.

(2) (LS110316)
Ri: èto ja nazyvaetsja ja xotela na ètju- èti *ferien* zarabotat' i
this I is_called I wanted for this holiday to_earn and
vsë *spar*-ovat'. i na *führerschein* i čë?
everything to_save and for driving_licence and what
führerschein? na jogu ščas pojdu [ha-ha *ha-und-em*] *bestellen*
driving_licence to yoga now will_go H&M order
ha-ha-ha

Vi: [ha-ha *ha-und-em*]
H&M

Ve: [ha-ha *ha-und-em*]
H&M

Ri: *alles ähnliche* a *führerschein* podoždët
all similar and driving_licence will_wait

Re: *aber jemand hat gesagt dass man gar nicht bei ha-und-em*
but someone has said that they PTCL not at H&M
einkauft.
shop

Ri: ja sebe èto toka vsjakie *görtel* ili čën'-t' takoe pokupaju ili
I myself PTCL only various belts or something similar buy or
balerinas vsë takoe kak by; čën'-t' [drugoe ne pokupaju]
all that like PTCL something different not buy

Ve: [tam] tam takie prostye veščі možno kupit', a esli vot
there there such simple things possible to_buy but if PTCL
während tam odevat'sja èto
while there to_dress this

Ri: [da] ne-ne
yes no

Vi: *ich hab mir ein kleid für fünf euro bestellt*
I have me a dress for five euros ordered

Ri: he-he

Vi: *aber nicht so einfach*; *ist einfach mit rüschchen so_n bisschen*
but not so simple is simply with frills a bit

Ri: ne esli ja sebe tam čën'-t' pokupaju to vot èto vot top
no if I myself there something buy then PTCL this PTCL
za četyre *euro* ((laughing))
for four euros

Ri: 'So much for my wish to earn (some money) during these holidays and save everything. For my driving licence. And now? The driving licence? Now I am going to take yoga classes, to order H&M articles.

Ha, ha, ha!'
Vi: 'Ha, ha, H&M.'
Ve: 'Ha, ha, H&M.'
Ri: 'All those things, and the driving licence will wait.'
Re: 'But somebody said that they never buy anything at H&M.'
Ri: 'I buy myself only stuff like belts or similar things, or balerinas, stuff like that. I don't buy other things.'
Ve: 'You can buy simple things there, but, while you cannot buy all your clothes there.'
Ri: 'Yeah, no-no.'
Vi: 'I've ordered myself a dress for five euros.'
Ri: 'He, he.'
Vi: 'But it's not that simple; it is simply with some frills.'
Ri: 'No. If I buy myself anything there, then a top like this for four euros.'

At the beginning of this excerpt from a conversation recorded in a school setting, Rita describes her plans for the upcoming holidays and mentions her incapability to save money for the driving licence because she spends it on yoga classes and H&M articles. After I expressed my surprise at the fact that despite an earlier statement, the girls buy things at that store, each of the girls downplayed this fact by naming the few articles that they consider purchasable there. While I formulated my surprise in German, Rita and Vera responded to it in Russian, and only Vika phrased her turn in German. In spite of that, Rita continues in Russian, thus sticking to the language of the conversation. Crucially, each of the turns framed in Russian contains German(-origin) words.[8] In this regard, the language consultants' speech recorded by myself is similar to the samples recorded by the recruited community members in natural settings.

The recorded conversations in both types of setting contain purely monolingual intervals. Most of the time the base language of such monolingual passages is Russian, but in some situations it is German. For example, in situations involving a German monolingual or a German dominant speaker, the interlocutors switch to German as the base language of interaction. Being extremely rare, these situations barely influence the distribution of languages in the corpus.

Table 3.5 lists the situations in which the data were collected and details the constellations of the conversation participants. Additionally, it specifies the duration of each of the recordings. All the recordings marked with the asterisk

[8]Without additional evidence, it is impossible to decide whether the German lexeme *Euro* is a switch, or a borrowing.

correspond to the situations in which the writer was one of the participants. Altogether, these recordings include ten hours of recorded conversation. The duration of the samples recorded by the language consultants in their networks amounts to 15 and a half hours. The bulk of these data (6 and a half hours) includes the interactions between Irina and Olga, university students of Russian. Crucially, including such a large portion of interactions between specific speakers to the corpus did not affect the distribution of the reported patterns of bilingual speech because the conversations between Irina and Olga contained a large number of long Russian monolingual intervals. As these passages were unsuitable for the investigation of language mixing, the scope of this book, the studies reported below draw on the data extracted only from bilingual turns.

The next step in the construction of the data sets for specific case studies involved the transcription of the recorded speech. As mentioned above, I disregarded the monolingual intervals in the conversations and transcribed only the bilingual passages. I finally extracted Russian sentences containing German lexical items and German sentences containing Russian lexical items. The great majority of the identified items included singly occurring nouns and verbs as well as adjective-modified noun phrases and prepositional phrases. These contexts, with the exception of verbs, yielded the grammatical contexts scrutinised in the remainder of this book.

3.7 Conclusion

In this chapter, I have described the research participants, who were German repatriates from the former Soviet Union and its successor states. They are traditionally referred to as Russian Germans and constitute Germany's largest group of Russian speakers. Although German, encompassing the standard language and various Russian German dialects, was the traditional community language of Russia's Germans, they have been undergoing language shift to Russian since the Second World War. Prior to the large-scale repatriation to Germany in the late 1980s and the 1990s, transmission of the minority language in the families had largely ceased. Therefore, the focus of the current research has been on the generations born between the late 1970s and the early 1990, which were labelled here as youths and young adults. These speakers demonstrate high-level proficiencies in both Russian and German, and their bilingual speech is likely to exhibit considerable variability in code-mixing.

Crucially, unlike the typical intermediate-generation immigrants, who learn the host language after moving to the host country, Russian Germans of the in-

Table 3.5: Speech situations of the recordings.

Recording	Speaker	Speech situation	Duration	Other people present
FI110801*	Svetlana	home	1 h 30 min	child
FM120811	Elena, Ira, Nataša	restaurant, lunch table	25 min	children
HO100712	Alina, Tanya	home	05 min	
HO100712:01	Alina, Tanya	home	35 min	mother, brother-in-law, children
HO100712:02	Alina, Tanya	restaurant, lunch table	30 min	friend
LA120503:01	Irina, Olga	restaurant, lunch table	14 min	
LA120503:02-05	Irina, Olga	on a train	2 h 00 min	
LA120503:06	Inna, Olga	in a car	48 min	
LA120503:09, LA120503:13	Inna, Olga	home	3 h 08 min	
LA120503:11, LA120503:12	Inna, Olga	in the street, later Inna's home	1 h 28 min	
LS101221*	Nadya, Rita, Vera, Vika	classroom	37 min	fellow students, teacher
LS110125*	Nadya, Rita	classroom	1 h 21 min	
LS110316*	Rita, Vera, Vika	schoolyard	58 min	
LS110405*	Vera, Alex	classroom	1 h 02 min	
LS110510*	Larisa, Alex	classroom	55 min	
LS110526*	Nadya, Vera	classroom	1 h 22 min	
LS110712*	Larisa	classroom	1 h 12 min	fellow student
LS110714*	Nadya, Rita, Vera	classroom	35 min	
LV120224:01-10, LV120224:13-21	Irina, Olga	on a train	4 h 02 min	
LV120224:11	Irina, Olesya, Valentina	home	40 min	their children
LV120224:12	Irina, Olga	on a train	35 min	fellow student
VS120425*	Marina	restaurant, dinner table	50 min	friends

termediate generation had often been exposed to German prior to their repatriation. A third of the research participants reported having grown bilingual, with German being one of their family, or community, languages prior to the repatriation. Other five respondents claimed that they had received some linguistic input in German in form of chunks, including sayings, songs, and the like. For this reason, the selection of the research participants was guided by their linguistic biographies, although it largely followed the traditional generational approach to immigrant languages, which relates the first, the second, and the intermediate generation to a specific pattern of bilingual language use and dominance. I have argued that careful treatment of the bilingual speaker's linguistic development in the community languages prior and upon their repatriation may ensure a high degree of homogeneity among the gathered speech samples in terms of the amount and quality of language mixing therein as well as more general patterns of bilingual language use.

Finally, I described the methods of data collection. While half of the material used for the corpus construction were informal peer interactions recorded by language consultants in natural speech situations, another part of the data included conversations recorded by the writer, in which he was one of the participants. Solely the bilingual turns of the recorded interactions were subject to transcription and further analysis. The inspection of the other-language elements in sentences framed by the base language of the interaction demonstrated that nouns and their combinations with adjectives and prepositions are among the most frequent items occurring in bilingual sentences, aside from verbs. Therefore, the remainder of this book focuses on the distributional patterns of the other-language nouns in the presented bilingual corpus.

4 Code-mixing in the adjective-modified noun phrase

This chapter[1] investigates insertional code-mixing in noun phrases with adjective modifiers. In this syntactic context, the speaker may insert an attributive adjective from one contact language into a noun phrase headed by a noun from the other contact language. Alternatively, she can insert a nominal constituent that comprises a noun and an attributive adjective into a sentence framed by the other language. Hence, two patterns of code-mixing in this context are distinguished: the switch may be placed within the modified noun phrase or at the nominal constituent's boundary. Insertions of both nouns and nominal constituents, with and without modifiers, are well documented in the literature. For example, already Poplack (1980b) offered empirical evidence suggesting that "nouns and noun phrases are frequently switched" (p. 604). Being ubiquitous in corpora of bilingual speech, nouns, nominal constituents and fully-fledged noun phrases have been the focus of extensive research along various lines (e.g., Sankoff et al. 1990, Poplack & Meechan 1995, Cantone & MacSwan 2009, Parafita Couto et al. 2015).

The aim of this chapter is threefold: First, the chapter sets out to investigate patterns of code-mixing in the modified noun phrase in Russian-German bilingual sentences. Second, it aims to construct and assess a statistical model which predicts, by using phrase frequency and noun frequency as predictive factors, the variation in the way Russian-German bilinguals code-switch in the examined syntactic context. Third, the chapter explores the potential of usage-based explanations for code-mixing patterns in the noun phrase domain and thus contributes to clarify the ongoing discussion on the nature of adjective-noun insertions.

To set the stage for the analysis, I first review the existing research on code-mixed nouns and nominal constituents, including adjective-modified noun phrases, and examine the issue of their status and nature as addressed by the major approaches. In section two, I outline the main features of adjective-noun combinations in German and Russian. The third section entails an analysis of

[1]Hakimov (n.d.) is an earlier and abridged version of this chapter.

Russian-German code-mixing in the context of the noun phrase modified by the attributive adjective. The fourth section introduces the main hypotheses underlying this chapter and investigates the factors suspected of regulating the variation in the scrutinised code-mixing patterns: switching within the adjective-modified nominal constituent and switching at its boundary. In the subsequent section, I present and assess the statistical model which predicts this variation. The final section entails a summary and a discussion of the main findings of the current study.

4.1 Insertion of nouns and nominal constituents in bilingual speech

4.1.1 Noun insertion

In language contact, nouns are involved in borrowing and in code-mixing more frequently than other word classes. Linguists who approach borrowing from a historical perspective (Whitney 1881, Moravcsik 1978, Thomason & Kaufman 1988, Matras 2009) place nouns on top of borrowability hierarchies. Numerous studies of code-mixing show at the same time that noun insertion is the most frequent pattern of code-mixing. For instance, using multilingual data from the Gambia, Haust (1995: 112, 107) finds that in her Mandinka-Wolof-English corpus, nouns constitute 55.8 per cent of all lexical morphemes inserted and integrated into the matrix language (327 of 586 tokens) and 66.5 per cent of all bare insertions (141 of 212 tokens). Researchers who regard noun insertion as borrowing (Poplack et al. 1988, Sankoff et al. 1990, van Hout & Muysken 1994, Muysken 2000)[2] also report that bilingual speakers borrow nouns extremely often, even if, on certain occasions, noun insertions can count as switches (cf. Chapter 1.6). Treffers-Daller (1994: 99–100) demonstrates by using bilingual French-Dutch data that nouns constitute the greater part of single word switches (borrowings) in her Brussels corpus. She finds that French nouns encompass 58.4 per cent of overall French single word switches in Brussels Dutch, whereas Dutch nouns cover only 23.9 per cent of Dutch single word switches in French. (In fact, inserted Dutch nouns lie just behind inserted Dutch interjections with regard to both their token and type frequencies.) In total, noun switches still clearly dominate single word switches in the corpus. In conjunction with this, the question

[2]It should be noted that Muysken (2000) acknowledges the equivocal status of single word insertion; he contends, "Lexical borrowing has been associated with insertional code-mixing, and not without reason. Nouns are the class of elements borrowed par excellence and also the prime example of insertion under categorical equivalence" (p. 75).

arises why nouns occupy such a prominent position in code-mixing and in the diachronic process of borrowing.

One reason why nouns exhibit such a high degree of borrowability is their semantic nature. According to van Hout & Muysken (1994), nouns, as prototypical content words, are borrowed because unlike function words they "have a clear link to cultural content and the latter do not" (p. 42). Moreover, nouns are notably prone to being selected in code-mixing because their meanings are highly specific (Backus 1996, 2001, Field 2002). Backus (1996: 115–131) proposes a specificity continuum, with one pole formed by proper nouns and the other pole formed by schematic, or abstract, units. One of intermediate categories includes words denoting very specific objects. Backus (1996: 116) defines these words as names for those concepts for which no other terms are available; these words are often referred to as cultural borrowings (cf. Myers-Scotton 1993: 165). If we consider the word classes to which most proper names and cultural borrowings belong, we will state that most of them are nouns.[3] A structural property of nouns that makes them highly borrowable is their high syntagmatic freedom (Backus 2013), which means that nouns are less bound to other words in the sentence and can thus be inserted in any sentence configuration. In language contact, especially in situations of intensive contact, the semantic and structural factors apparently interplay (cf. van Hout & Muysken 1994).

Nouns can be inserted into virtually all possible configurations of the noun phrase. Louis Boumans (1998: 221) characterises noun insertion as an unconstrained process. In his 1998 monograph The Syntax of Codeswitching: Analysing Moroccan Arabic/Dutch Conversation he offers a detailed account of patterns in which Dutch nouns appear in the context of Moroccan Arabic. He describes various types of noun determination (p. 181–191, 211–214) and modification (p. 196–198, 200–205). In the latter case, he shows that inserted Dutch nouns can be modified by prepositional adjuncts and interrogative forms, but noun modification by Moroccan-Arabic adjectives is limited to few "atypical" instances (Boumans 1998: 200–201). Furthermore, as nouns can subcategorise for complements in both Moroccan Arabic and Dutch, inserted Dutch nouns sometimes take Moroccan-Arabic prepositional and clausal complements. Unfortunately, Boumans (1998) provides little or no information about the distribution of the various structures in the corpus. Alongside single nouns, insertion of nominal constituents from another language is also commonplace in bilingual speech involving various language pairs. To this, I turn in the following section.

[3]Backus (1996) states that "Especially in early stages of contact many single EL content words are names of some sort, such as place names and personal names, but also less obvious cases" (p. 116).

4.1.2 Insertion of nominal constituents

Noun phrase insertions are commonplace in mixed clauses as well, although they are not as frequent as insertions of single nouns (cf. Poplack 1980b). For example, Treffers-Daller (1994: 205) observes that noun phrases make up 41.3 per cent of full-constituent switches in her French-Dutch data from Brussels. Haust (1995: 165) also reports that noun phrases comprise the most frequent type of constituent insertion in her Gambian corpus: 79 noun phrases make up 47.9 per cent of 165 constituent insertions observed. Although such general quantitative data are not rare in studies of code-mixing, researchers often provide little or no information on the distribution of noun phrases varying in their structure. However, it is interesting to know what types of noun phrase modifiers and complements are more or less common in code-mixing and why. The aforementioned monograph by Boumans (1998) represents an exception to this tendency insofar as it provides a comprehensive account of syntactic structures involved in Moroccan Arabic-Dutch code-mixing. Let us consider some of the findings regarding noun phrase insertion in greater detail.

Boumans (1998) shows that Dutch noun phrases inserted into Moroccan Arabic can be simple, such as *het vermogen* 'the capacity' in (1a), or expanded. The latter may contain a prepositional complement or an attributive adjective, as illustrated in (1b) and (1c).

(1) Moroccan Arabic-Dutch (Boumans 1998: 212, 199, 212)

a. ʕend-i *het vermogen* baš n.. eh ne-tbeʕ *chemie*
at-1SG DEF.N capacity COMP 1-er 1-follow chemistry(Fr)
'I have the capacity to study chemistry.'

b. kayn-in *daar-voor verklaring-en*
EXIST-PL there-for explanation-PL
'There are explanations for that.'

c. ila bġi-tu t-reyyḥ-u ʕend-i, t-šerb-u l-atay u t-neʕs-u u
if want-2PL 2-rest-PL at-1SG 2-drink-PL DEF-tea and 2-sleep-PL and
te-mši-w eh *de volgend-e dag* t-zid-u!
2-go-PL er the next-AGR day 2-go.on-PL
'If you (PL) want to stay at my place, you drink tea, you sleep, and you leave er the next day you go on.'

Fully-fledged Dutch noun phrases inserted into Moroccan-Arabic clauses are considerably less frequent than inserted nouns (cf. Boumans 1998: 210). Particularly

insertion of expanded Dutch noun phrases is low in number. For example, the noun phrase with a prepositional complement and the noun phrase modified by a relative clause appear only twice each in the Moroccan-Arabic discourse in Boumans's data. In contrast, inserted Dutch nouns, as mentioned in (4.1.1), are freely modified "by means of [Moroccan-Arabic] possessive constructions, adjunct or complement PPs, and relative clauses" (Boumans 1998: 201). Expanded Dutch nominal constituents, just like Dutch single nouns, often follow Moroccan-Arabic determiners, as shown in (2a), where the inserted Dutch noun *nadeel* 'disadvantage' follows the Moroccan-Arabic indefinite determiner *ši* and is modified by the Dutch prepositional adjunct *voor de universiteit* 'for the university'.

(2) Moroccan Arabic-Dutch (Boumans 1998: 198, 192)

a. u hadak š-ši y-kun *nadeel* *voor de universiteit*?
and DEM DEF-INDEF 3-be disadvantage for the university
'And this will be a disadvantage for the university?'

b. ʕla xaṭer ʕend-ek ši beʕḍ l-h.. *handeling-en die je doet* (..)
because at-2SG INDEF part DEF-h.. action-PL that you do
'Because you have some .. some actions that you do.'

Example (2b) illustrates the use of the Moroccan-Arabic definite article *l-* with a Dutch nominal constituent, which consists here of the Dutch noun *handelingen* 'actions' and the modifying relative clause *die je doet* 'that you do'.[4] As such, switching the language between determiner and noun has been frequently reported for other bilingual situations. According to Poplack (1980b: 604), this is a favourable locus for code-switching in her Spanish-English data. We can conclude from Boumans's analysis that insertion of fully-fledged Dutch noun phrases is less frequent than insertion of nominal constituents after Moroccan-Arabic determiners.

Noun-adjective combinations represent the most commonly inserted nominal constituents in Boumans's data (cf. Boumans 1998: 203–205). These insertions have an internal Dutch structure: first, the order of adjective and noun is A N, which is the opposite of the Moroccan-Arabic order, and secondly, the adjective exhibits the agreement marker *-e* – for all forms except for the indefinite singular neuter – as in (3) below:

[4]We can analyse the utterance in (2b) as an instance of revision: the speaker interrupts the utterance just before the switching site, i.e., at the beginning of the switched noun, so that the restart does not contain the Moroccan-Arabic determiner any more. It should be noted however that the same speaker produces smooth switches involving the Moroccan-Arabic prefix *l-* as well (cf. Boumans 1998: 185).

(3) Moroccan Arabic-Dutch (Boumans 1998: 204)
walakin daba d-derri ġadi ye-dxel *islamitisch-e school*, škun ġadi
but now DEF-child FUT 3M-enter Islamic-AGR school who FUT
ye-lqa
3M-meet
'But now the child will go to an Islamic school, whom will he meet?'

The adjective-noun combination *islamitische school* 'Islamic school' is inserted without a determiner, the omission of determiner results from either the determinerless use of the noun *school* in Dutch, as in the collocations *naar school gaan* 'go to school' and *met school beginnen* 'start school', or the tendency in Moroccan Arabic to omit the definite article before foreign stems (cf. Boumans 1998: 187). The omission of determiner with Dutch adjective-noun insertions in Moroccan Arabic discourse is as common as with Dutch single nouns. Hence, Boumans (1998: 205) states that inserted adjective-noun combinations behave just like inserted nouns. Another point that needs emphasis here is that, unlike Dutch attributive adjectives, Moroccan-Arabic adjectives are never used to modify single Dutch nouns in Moroccan-Arabic discourse (cf. 4.1.1 and Boumans 1998: 201). Furthermore, Boumans asserts that Dutch adjectives are generally infrequent in Moroccan-Arabic sentences, even in the predicative function. Possible explanations for this observation are a lack of congruence between Moroccan Arabic and Dutch in the category of adjective and a mismatch between the involved languages in the adjective-noun order. However, as I will show in the subsequent section, this is not always the case and inserted nouns may well be modified by attributive adjectives in the same language when the patterns of adjective modifications employed by the contact languages vary.

4.1.3 Inserted adjective-noun combinations

Extensive insertion of adjective-noun combinations is characteristic not only of Moroccan-Arabic-Dutch code-mixing, but of code-mixing in general. Numerous studies of code-mixing have shown that adjective-noun combinations from one language appear in the discourse of the other language on a regular basis, regardless of the adjective-noun order that both languages use. As illustrated below, adjective-noun combinations are inserted into sentences in the other language when both languages share the adjective-noun order. For instance:

(4) Turkish-Dutch (Backus 1996: 177)

… cuma günü *vaste dag* yani
Friday.day fixed day so
'…we always go on Friday, so'

(5) Croatian-English (Hlavac 2003: 231)
… to sam završi-o dva tjedn-a, *so* … i tako
that be.PRS.1SG finish-PTCP.SG.M two week-GEN.SG so and so
sam dobi-o rezultat-e *last week* i …
be.PRS.1SG receive-PTCP.SG.M result-ACC.PL and
'…I finished that two weeks ago, so …and so I received the results last week and …'

(6) German (dialect)-Hungarian (Szabó 2010: 373)
… ihr ha-t *szociális villany* ghab
2PL have-PRS.2PL social electricity[NOM.SG] have.PTCP
'…you have had a social benefit rate for electricity'

The language constellations here – Turkish-Dutch in (4), Croatian-English in (5), and German-Hungarian in (6) – all use the A N order. The inserted noun-adjective combinations *vaste dag* 'fixed day' (in 4), *last week* (in 5) and *szociális villany* 'social (benefit rate for) electricity' (in 6) all exhibit the internal structure of the corresponding language: Dutch, English and Hungarian, respectively. That is, the Dutch adjective *vast* takes the Dutch agreement marker *-e* (cf. 3), and the English adjective *last* and the Hungarian adjective *szociális* remain uninflected, as required by English and Hungarian, but neither Croatian nor German.

In contact situations with languages differing in the adjective-noun order, combinations of nouns and adjectives have also been reported among frequent insertions. This applies, for instance, to Moroccan Arabic-Dutch code-mixing discussed above (§4.1.2). Another example is code-mixing between English and Welsh, since in Welsh, unlike in English, the head noun precedes the adjectival modifier (cf. Deuchar 2005: 260). In (7), the word combination *main news* occurs in a Welsh sentence, the English adjective-noun order indicates that the internal structure of this nominal constituent is English.

(7) Welsh-English (Deuchar 2005: 261)
ar y *main news*, ar y *news* oedd o
on DET on DET be.3S.IMP PRON.3S.M
'On the main news, it was on the news'

An excellent source of further examples for this case is the (1991) study by Penelope Gardner-Chloros. The author reports numerous insertions of nominal

constituents containing adjectives in her French-Alsatian data from Strasbourg. Again, the involved languages contrast in the adjective-noun order, though not as categorically as English and Welsh. While attributive adjectives in Alsatian occupy the position preceding the head noun, the position of adjectives in French can vary: most attributive adjectives follow their head nouns, but a reasonably large group of adjectives occur pre-nominally. With regard to the adjective-noun order, both types of combinations are registered, even though combinations with the N A order seem to dominate all French noun-adjective insertions (cf. Gardner-Chloros 1991: 141). The presented data contain instances such as *gouverneur militaire* 'military governor', *groupement alsatique* 'Alsatian co-operative', *portes ouvertes* 'open day', *résidence secondaire* 'holiday home', *visites guidées* 'guided tours', *temps morts* '(cassette) blanks' (p. 148), but also combinations with the A N order, such as *jeune homme* 'young man' (p. 139). Interestingly, the bilingual speaker can realise the adjective at one point in French, as in *toupie jaune* 'yellow spinning top', and later in the same conversation in Alsatian, as in *gäls toupie* 'yellow spinning top' (Gardner-Chloros 1991: 133). The latter example illustrates the possibility of single French nouns to be modified by Alsatian adjectives, as in *e kleini attention* 'a little treat' (p. 120), *wunderbari charcuterie* 'wonderful charcuterie' (p. 124) and *e klanni note* 'a little note' (p. 140). It appears from this brief survey that despite the highly restricted congruence between Alsatian and French in the order of noun and its adjectival modifier, Alsatian-French bilingual speech exhibits considerable variability regarding code-mixing patterns in nominal constituents with attributive adjectives. These findings contrast strongly with those of Boumans (1998) on code-mixing between Moroccan Arabic and Dutch, which also use differing adjective-noun orders. As such, single Dutch nouns, unlike single French nouns in Alsatian discourse, are not modified by Moroccan-Arabic attributive adjectives (cf. 4.1.1). This comparison demonstrates that, although differing adjective-noun-order patterns may restrict adjectival modification of insertions by the other-language adjectives, as in the case of Moroccan-Arabic and Dutch code-mixing, they do not automatically rule out this possibility altogether. However, distributional properties of adjectival modification in bilingual sentences across various language pairs as well as different theoretical orientations have brought some scholars to conclude that adjective-noun combinations in the context of the other language pertain to the domain of lexicon, and others to suggest that insertion of nominal constituents operates in the domain of the grammar. The existing approaches to adjective-noun insertions are detailed in the following section. I first discuss their status as lexical or grammatical forms, and then turn to explanations of their occurrence in bilingual speech.

4.1.3.1 Status of inserted adjective-noun combinations

The status of inserted adjective-noun combinations is the focus of much current controversy: while some scholars consider them pertinent to the lexicon, others regard them as inserted nominal constituents and thus a matter of grammar. In a detailed study on Tamil-English code-mixing, Sankoff et al. (1990) analyse English adjective-noun combinations in Tamil sentences, such as in (8), as lexical borrowings.

(8) Tamil-English (Sankoff et al. 1990: 96)
Religion-uDaya *main purpose* vantu oru *supernatural being*-la oru
-GEN PTCL DET -LOC DET
belief create paNNaratu.
do.INF
'Religion's main purpose is to create a belief in a supernatural being.'

They assume noun-adjective combinations, such as *main purpose* and *supernatural being*, to have the status of compound borrowings – whether nonce, or established – because "the function words typical of English NPs *never* co-occur with them" (Sankoff et al. 1990: 80, emphasis in the original). Other compound nonce-borrowings include *snide remarks*, *serious subjects*, *educational system*, *slacks and blouses*, *Government of India Scholarship*, *Hindi songs*, *Indian women*, *arranged marriage* (Sankoff et al. 1990: 80, passim). This analysis is criticised by Muysken (2000: 78–81) for several reasons. His central argument against the nonce-borrowing analysis is that the borrowing process, for the most part, does not apply to multiword combinations at such a high rate as is the case with the material presented in Sankoff et al. (1990). Nevertheless, Muysken agrees that "complex" nouns such as *supernatural being* in (8) are borrowable per se, because they might have been lexicalised already in English (p. 79). His alternative analysis assumes that these constructions are inserted NPs and thus a matter of syntax, rather than the lexicon.

A further relevant study into the nature and status of inserted nominal constituents is Poplack & Meechan (1995). In this study of Wolof-French code-mixing, the authors examine complex French nouns that occur in otherwise Wolof sentences. Among these nouns are instantiations of the French 'N *de* N' construction, such as *langue de cuisine* 'broken language' and *conditions de vie* 'living conditions' (Poplack & Meechan 1995: 215). In (9a), a French construction *tête de liste* 'head of the list' is followed by the post-nominal Wolof determiner *bi*.

(9) Wolof-French (Poplack & Meechan 1995: 215, 228)

a. fexeel ba nekk ci *tête de liste* bi rek.
try.IMP until be PREP head of list DEF ADV
'Try to be only at the head of the list.'

b. fokk naa moom moo la *envoyer*-woon *lettre* bi *quoi*.
think I him FOC.he you send-PAST letter DEF what
'I think it's him that had send you the letter eh.'

Poplack & Meechan (1995: 215) report that 'N *de* N' constructions in Wolof sentences take the Wolof determiner *bi*, just like single French nouns in Wolof discourse (as example 9b with the noun *lettre* 'letter' demonstrates). Therefore, the authors regard the individual instances of the 'N *de* N' construction as loanwords. In doing so, they adopt the view outlined in Sankoff et al. (1990), but introduce another diagnostic feature of (nonce-)borrowing: the semantic nature of a word sequence. The authors assert that "the 'N *de* N' constructions virtually all consist of frozen or idiomatic expressions functioning as compounds" (p. 215). Hence, the lexicalised nature of the examined sequences is a further argument in favour of their analysis as borrowings.

As regards Dutch adjective-noun combinations occurring in Moroccan-Arabic sentences (cf. §4.1.2) and French noun-adjective combinations in Alsatian discourse, these combinations are reminiscent of the French 'N *de* N' constructions in the Wolof discourse and, following Poplack and Meechan's structural and semantic criteria, count as borrowings. As mentioned above, in Muysken's (2000) view, these combinations of nouns and adjectives are inserted NPs. An intermediate position in this controversy is taken, for example, by Gardner-Chloros (1991: 141), who introduces the category of "lexical switches", encompassing the noun-adjective combinations under discussion.

Another possibility to analyse adjective-noun combinations is to approach them from the Matrix Language Frame model proposed by Myers-Scotton (1993). According to this model and its extensions (see Chapter 1 for further details), noun-adjective combinations in focus would be classified as "EL islands" as they are "full constituents consisting only of Embedded Language morphemes occurring in a bilingual CP [Complementiser Phrase] that is otherwise framed by the Matrix Language" (Myers-Scotton 2002: 139). The internal structure of an embedded-language island corresponds to the norms of the language that provides the morphemes. For example, in (7), the adjective-noun combination *main news* follows English norms, whereas the rest of the sentence is in Welsh (cf. Deuchar 2005: 261). We can assert that the examples of adjective-modified nouns quoted above as well as the aforementioned instances of the French 'N *de* N' con-

struction are well-formed embedded-language islands. Whether these combinations are borrowed, or inserted, the question arises as to what factors determine the selection of the adjective modifying the inserted noun, or under what circumstances an attributive adjective comes from the same language as the inserted noun and when it comes from the other language.

4.1.3.2 Explanations for inserting adjective-noun combinations

One of the first explanations for the insertion of adjective-noun combinations is offered through the Matrix Language Framework model, which treats these combinations as embedded-language islands. As outlined in §1.4, Myers-Scotton & Jake (1995) assert that an embedded-language island may occur if two languages lack congruence in one of three aspects of "lemmas", i.e., abstract entries in the mental lexicon that underlie lexemes. These aspects include semantic/pragmatic features, predicate-argument structure and realisation patterns. However, this view neither details which of the three aspects motivates the emergence of an embedded-language island in a given context, nor is it susceptible to an empirical verification. It is apparently this consideration that has lead some scholars to restrict their analyses of embedded-language islands to a specific type of congruence. For example, Deuchar (2005) assumes that "while the lack of semantic/pragmatic equivalence between lexemes in two languages may be an important cause of code-switching [which corresponds to code-mixing here] in the first place, it is considerations of grammatical rather than semantic congruence which determine whether or not a switch can take place" (p. 258).[5] In case of nominal constituents containing an attributive adjective, congruence applies to word order within the constituent, or, in Myers-Scotton's terms, to morphological realisation patterns. As discussed above, word-order non-equivalence may account for the occurrence of embedded-language islands consisting of a noun and an attributive adjective in such language pairs as Moroccan Arabic-Dutch, Welsh-English, and possibly Alsatian-French. Nonetheless, this explanation would not hold for inserted adjective-noun combinations in language pairs with an identical noun-adjective order, namely Turkish-Dutch, Croatian-English, German-Hungarian and Tamil-English.

Another explanation of nominal constituent insertions at non-equivalence sites is adumbrated by Poplack & Meechan (1995: 215). They assert that nominal constituents, such as the 'N *de* N' constructions, can be borrowed because

[5]At the same time, Deuchar (2005) mentions one of Wei's (2001) findings that incongruence in semantic/pragmatic features or predicate-argument structure can motivate occurrences of English embedded-language islands in otherwise Chinese discourse (quoted in Deuchar 2005: 258).

they comprise "frozen or idiomatic expressions". This point recalls Backus (1996, 1999a, 2003), who elaborates the idea that idiomatic expressions are one type of multimorphemic lexical units which are inserted into a matrix language clausal frame. The "unit" hypothesis, laid out in §1.5, predicts that every embedded-language island is a unit (cf. Backus 2003: 91). Backus defines as units such multimorphemic elements that exhibit irregular morphosyntax, non-compositional semantics or high frequency. A multimorphemic element distinguished by one of these characteristics is assumed to be stored in the mental lexicon as a whole. Its parts are welded together by the irregularity of its form or the frequency of co-occurrence. However, it is possible to hypothesise that bilingual corpora may contain embedded-language islands lacking these properties. When such sequences do not count as multimorphemic units, their occurrence in bilingual clauses would remain unexplained. An example of a multimorphemic sequence lacking the characteristics of a unit is the aforementioned French combination *toupie jaune* 'yellow spinning top', inserted into an Alsatian framing clause (remember that in further discourse the speaker realises the adjective in Alsatian, i.e., *gäls toupie*, Gardner-Chloros 1991: 133). The French string is neither frequent in French, nor is its meaning non-compositional. Furthermore, adjective-noun order differences between French and Alsatian cannot be a valid explanation for the insertion of *toupie jaune* as well (in view of the French adjective-noun insertions in the Alsatian discourse). Hence, the obvious way to examine the "unit" hypothesis is to test it statistically.

A distinction between fixed expressions and ad hoc produced strings underlies the analysis of Russian nominal constituents in Kazakh discourse in Muhamedowa (2006: 77–88). Although the author claims to have adopted Backus's approach, only expressions that originate from the Russian institutional language are considered fixed. Namely, multiword strings such as *metodičeskij otdel* 'curriculum and instruction department' (p. 77) and *ministerstvo lëgkoj promyšlennosti* 'ministry of light industry' (p. 79) are analysed as fixed, whereas the strings *staryj ploščad'* 'old square'[6] in (10) and *vysotnye doma* 'high-rise houses' (p. 81) are regarded as ad hoc creations.

(10) Kazakh-Russian (Muhamedowa 2006: 82)
anau *televizor*-dan kör-gen šïğar-sïz-dar. ne-ler-i-n
that TV.set-ABL see-PERF perhaps-POL-2PL thingummy-PL-POS3-AKK
uže anau *star-yj* *ploščad'*-tï ne-ler-di
already that old-NOM.SG.M square[NOM.SG.F]-ACC thingummy-PL-ACC

[6]It is worth noting that the adjective *staryj* 'old' lacks gender agreement with the feminine head noun *ploščad'* 'square' (Muhamedowa 2006: 82).

žönde-di, anau universitet ne qïl-dï.
improve-3SG that university thingummy do-3
'Perhaps you have seen that on TV. They have renovated that Old Square and the thingummy already, and they did that thingummy at university.'

However, the adjective-noun combination *staryj ploščad'* 'old square' in (10) is obviously a place name and could thus be a fixed routine in the speaker's idiolect and a lexical unit in her mental lexicon. This may be the case because place-names are unlikely to be produced compositionally on-line. As for the string *vysotnye doma* 'high-rise houses', it is a fairly regularly occurring collocation in Russian: 17% of all occurrences of the adjective *vysotnyj* 'high-rise' in the Russian National Corpus comprise instances in which this adjective combines with the noun *dom* 'house'.[7] Consequently, the string should be viewed as a collocation and thus a lexical unit. We can conclude that Muhamedowa utilises register as an indicator of a string's fixedness. But classifying embedded-language noun-adjective combinations by virtue of register alone appears to be insufficient for determining the status of these combinations as lexical units. Furthermore, most of the examples in Muhamedowa (2006: 81–88) can well be subsumed under the category of lexical unit.

The final study to be discussed in this section is again the analysis of Dutch adjective-noun insertions in Moroccan Arabic by Boumans (1998). According to the author, in a situation in which embedded Dutch nouns are regularly modified by Dutch attributive adjectives but virtually never occur with Moroccan-Arabic adjectives, the selection of attributive adjectives, restricted by the noun being modified, results from "specific ties that bind nouns and attributive adjectives from the same language" (p. 220). By elaborating this point further, he provides support for Backus's (1996) conception: the ties between nouns and attributive adjectives are of collocational nature and are obviously not limited to idiomatic expressions (cf. Boumans 1998: 386–387). This consequence is extended to other multiword embedded-language islands. Existence of collocational ties can account not only for Dutch noun-adjective insertions in Moroccan Arabic, but also for fully-fledged phrases, like *de volgende dag* 'the following day' in (1c). As the definite determiner *de* co-occurs with the string *volgende dag* much more often than the indefinite determiner *een*, we can assume that the string *de volgende dag* is a unit, which is stored and selected as a whole in on-line production. Boumans (1998) concludes the section on collocational ties by suggesting that "[i]f the existence of collocational ties between lexical units in the mental lexicon accounts for

[7]The adjective *vysotnyj* 'high-rise' co-occurs at a higher rate only with the noun *zdanie* 'building'.

the co-occurrence of EL [embedded-language] words, the total absence of such ties may perhaps explain the observed constraints on the co-occurrence of ML [matrix language] and EL [embedded-language] lexical items" (p. 386–387). This means that an approach elucidating the emergence of embedded-language islands in bilingual speech has also the potential to clarify the intricacies of mixed constituents. Moreover, this idea provides a starting point for the subsequent study: in order to prove the existence of collocational ties, we need to compare inserted adjective-noun combinations with noun insertions modified by attributive adjectives from the matrix language.

To recapitulate, scholars suggest both structural and semantic explanations underlying embedded adjective-noun combinations in code-mixing. A structural approach considers incongruence between the involved languages in the realisation patterns of adjective-modified nominal constituents as a crucial factor (Myers-Scotton & Jake 1995, Myers-Scotton 2002). In other words, non-equivalence in the noun-adjective order can trigger the emergence of an embedded language island realised as an adjective-modified nominal constituent (Deuchar 2005). However, this explanation will fail for language constellations with an identical adjective-noun order, like German and Russian. Some scholars (e.g., Poplack & Meechan 1995) state that at least some embedded nominal constituents are idiomatic expressions. Although plausible at first sight, this explanation will not account for embedded adjective-noun combinations lacking non-compositional meanings. Finally, Backus (1996) and Boumans (1998) consider noun-adjective combinations multiword lexical units, even when their semantics is compositional. Frequently used combinations, whose meaning may be compositional, also gain unit status in this approach. Boumans (1998) argues that it is collocational ties between a noun and an adjective that determine their insertion as a combination in code-mixing. However, owing to a high variability observed in code-mixing, not every inserted adjective-noun combination would obviously qualify as a unit. Moreover, in order to obtain unequivocal evidence for the existence of collocational ties, it is insufficient to rely on code-mixing data alone; rather, examined collocations have to be investigated in both bilingual and monolingual speech. Therefore, following the report of the code-mixing patterns in the examined syntactic context as they occur in my corpus of Russian-German bilingual speech, I will analyse the adjectival modification of the obtained nouns in large monolingual corpora and will finally subject these results to statistical test. However, before turning to these issues, I will briefly introduce the patterns of adjectival modification in German and in Russian.

4.2 Adjective-noun combinations in German and Russian

In languages of the world, attributive adjectives may either precede or follow their head nouns. German and Russian employ both syntactic patterns for noun modification. However, the most common pattern is when attributive adjectives occur in the prenominal position and agree with their head nouns in gender, number and case (Švedova 2005 [1980](a): 1303; Eisenberg 1999: 232). For example:

(11) German (Zifonun et al. 1997: 1991)
ein klein-es grün-es Männchen
ART[NOM.SG.N] little-NOM.SG.N green-NOM.SG.N man(N)[NOM.SG]
'a little green man'

In (11), the adjectives *kleines* 'small' and *grünes* 'green' precede the noun *Männchen* 'little man'. Furthermore, congruence in number, gender and case is observed between the forms of the adjective, the noun and the indefinite article *ein*. This pattern of modification is also prototypical of Russian, for instance:

(12) Russian
malen'k-ij zelën-yj čeloveček
little-NOM.SG.M green-NOM.SG.M man[NOM.SG.M]
'a little green man'

The attributive adjectives *malen'kij* 'small' and *zelënyj* 'green' in (12) precede the noun *čeloveček* 'little man' and agree with it in number, gender and case.

At the same time, Russian and German adjectives may also follow the noun. Particularly in German, this pattern is restricted to special contexts and individual lexical items. German adjectives occupy the postnominal position in the following four cases: (i) adjectives are postposed as a result of topicalisation (i.e., split topicalisation), as in (13a); (ii) adjectives occur in apposition to nouns in names of products and dishes[8], as in (13b); and (iii) specific lexemes, such as *bar*

[8]The case of appositive adjectives is not limited to these contexts; apposition is traditionally characteristic of literary, poetic and folk-song texts (cf. Zifonun et al. 1997: 1991) and has recently expanded to press texts (Dürscheid 2002). In the latter case, such adjectives as *brutal*, *light* 'easy', *total* 'sheer' often appear in postposition (e.g., *Fußball brutal* 'brutal football', Dürscheid 2002: 67). However, the use of adjectives in these contexts is immaterial in the further analysis.

'cash' and *pur* 'pure', tend to follow their head nouns, as in (13c)[9].

(13) a. (Eisenberg 1999: 234)
Geld hilf-t dir nur gefälscht-es weiter
money(N)[NOM.SG] help-3SG DAT.2SG only forged-NOM.SG.N PTCL
'Money will help you, only if it is fake'

b. (Dürscheid 2002: 64, 78)
Whisky pur, Forelle blau, Schauma mild
whisky(M)[SG] straight, trout(F)[SG] blue, Schauma(N)[SG] mild
'Straight whisky, blue trout, mild Schauma'

c. (Fabricius-Hansen et al. 2009: 350; Dürscheid 2002: 67)
tausend Euro bar, Natur pur
thousand euro(M)[SG] cash, Natur(F)[SG] pur
'A thousand euros in cash, pure Nature'

As is evident from the above examples, German postnominal adjectives inflect for gender, number and case when their head nouns are topicalised (cf. 13a), whereas the postnominal adjectives *bar* 'cash' and *pur* 'pure' as well as adjectives in product names remain bare.

With regard to Russian, Zemskaja (1979: 149) reports that in the colloquial speech, attributive adjectives occur in the postnominal position more frequently than in the written language, for example:

(14) (Zemskaja 1979: 152)
Prines-i xleb-a svež-ego; A gde jubk-a
bring-IMP bread-GEN.SG.M fresh-GEN.SG.M and where skirt-NOM.SG.F
sinj-aja?
blue-NOM.SG.F
'Bring some fresh bread; And where is the blue skirt?'

Zemskaja (1979) asserts that the postposed adjectives *svežego* 'fresh' and *sinjaja* 'blue' in (14) do not carry a special information load and are thus devoid of prosodic emphasis (cf. Lapteva 1976: 208–212). According to Zemskaja (1979), prepositioned adjectives tend to carry information load and prosodic emphasis, particularly when the noun phrase appears in the clause-initial position, as in (15).

[9]This outline does not include the case of adjectival increments, also analysed as loose appositions (Auer 2007b: 654). For example, the postnominal adjectives *bayrische* 'Bavarian' and *geschnitten* 'cut' in the following instances are distinguished by an appositive relation to their head nouns: *eine mode eine bayrische* 'a fashion, a Bavarian one' (Schröder 1997: 103, quoted in Schwitalla 2006: 142) and *[...] und möhrchen brauch ich klein geschnitten* 'and I need carrots, cut in small pieces' (Auer 2007b: 654).

(15) (Zemskaja 1979: 153)
Krasiv-ye cvet-y ja segodnja kupi-l-a
beautiful-ACC.PL.M flower-ACC.PL.M NOM.1SG today buy-PST-SG.F
'What beautiful flowers have I bought today!'

However, Lapteva (1976: 208, 211, 221–222) considers both templates as functionally equal variants.

A further criterion for a syntactic description of Russian attributive adjectives is the distance between adjectives and their noun heads. Unlike in German, where attributive adjectives occupy only the positions immediately adjacent to their head nouns, adjectives in Russian may take positions distant from their noun heads. For instance:

(16) a. (Miller & Weinert 1998: 165)
interesn-uju prines-i mne knig-u
interesting-ACC.SG.F bring-IMP DAT.1SG book-ACC.SG.F
'Bring me an interesting book.'

b. (Lapteva 1976: 213)
Pojd-u prines-u stul'čik sebe
go[PERF.PRS]-1SG bring[PFV.PRS]-1SG little.chair[ACC.SG.M] REFL.DAT
malen'k-ij
small-ACC.SG.M
'I'll go and get myself a small chair.'

While the adjective of the "split" noun-phrase *interesnuju – knigu* 'interesting book' in (16a) appears in the prenominal position, the adjective *malen'kij* 'small' in (16b) occurs in the postnominal position. The distant prenominal position of the adjective *interesnuju* 'interesting' in (16a) is explained by a special information purpose, namely highlighting the adjective (Zemskaja 1979: 153). Miller & Weinert (1998) state that distant prenominal adjectives "are equal to or more important than the noun with respect to information load" (p. 167). As regards postnominal adjectives, as in (16b), Lapteva (1976: 213) and Zemskaja (1979: 153) describe one of their functions as elaboration. In this function, the adjectives are weak semantically and prosodically. Although the question of how syntactic configurations of attributive adjectives combine with intonation patterns to express various discursive meanings is of theoretical and practical interest, it is beyond the scope of this work and will not be discussed further.

The subsequent analysis of Russian-German bilingual data requires us to distinguish between the aforementioned case of distant postnominal adjectives, as in (16b), and the case of adjectival increments, as in (17).

(17) (Zemskaja 1979: 155)
Ona *zvoni-l-a* *tët-e* *Oksan-e* *čtoby*
NOM.3.SG.F call-PST-SG.F aunt-DAT.SG.F Oksana-DAT.SG.F so.that
ona… *kupi-l-a* *tort* *èt-ot* *sam-yj*,
NOM.3.SG.F buy-PST-SG.F cake[ACC.SG.M] that-ACC.SG.M very-ACC.SG.M
s *proslojk-ami*, *vafel'n-yj*
with layer-INSTR.SG.F wafery-ACC.SG.M
'She called aunt Oksana to ask her to buy that cake, with layers, a wafer one.'

The prepositional phrase *s proslojkami* 'with layers' and the adjective *vafel'nyj* 'wafery' are placed at the end of the utterance, after the word *samyj*, which is marked by a falling tone. According to Zemskaja (1979: 155–156), the function of these structures is elaboration. The author discusses them in conjunction with the principle of "associative adjoinment", which enables the speaker to add any additional detail to the end of the utterance. Nonetheless, she is silent about the difference between distant postnominal adjectives, as in (16b), and adjectival increments, such as *vafel'nyj* 'wafery' in (17). The morphological criterion, which is adopted to distinguish postnominal adjectives from adjectival increments in German (recall that German postnominal adjectives in the case of split topicalisation are inflected for gender, number and case, whereas adjectival increments, which are in apposition to their head nouns, are uninflected, see footnote 8), does not apply to Russian adjectives in the postnominal position: both types of adjectives inflect for gender, number and case. On formal grounds, I regard as increments only those postposed adjectives which are produced after words marked by the falling terminal tone. Being clause-peripheral elements, adjectival increments are irrelevant to insertional code-mixing, and will thus not be considered further below.

The last group of adjectives that need to be discussed in this outline are adjectives for which postposition is the only possible option. The use of such adjectives in Russian is restricted to two cases: (i) adjectives as part of classification terms and product names (Švedova 2005 [1980][b]: 203)[10], as in (18a), and (ii) morphologically unintegrated borrowings (Švedova 2005 [1980][a]: 556), as in (18b).

(18) a. (Švedova 2005 [1980][b]: 203)
šalfej *lekarstvenn-yj*, *šalfej*
sage[NOM.SG.M] gardenly-NOM.SG.M sage[NOM.SG.M]

[10]Rijkhoff (2009) uses the term "classifying modifier" to refer to this case.

lugov-oj; *marmelad* *jabločn-yj*,
meadow-NOM.SG.M gumdrop[NOM.SG.M] apple-NOM.SG.M
šokolad *soev-yj*
chocolate[NOM.SG.M] soya-NOM.SG.M
'garden sage (*salvia officinalis*), meadow sage (*salvia pratensis*); apple gumdrop, soya chocolate'

b. (Švedova 2005 [1980][a]: 540)
cvet *bordo*, *brjuk-i* *klëš*,
colour[NOM.SG.M] claret trouser-NOM.PL.M bell-bottom
jubk-a *plisse*
skirt-NOM.SG.F plissé
'a claret colour, bell-bottom trousers, a plissé skirt'

Interestingly, the morphologically non-integrated loans in (18b) *bordo* 'claret', *klëš* 'bell-bottom' and *plisse* 'plissé' have well-integrated counterparts: *bordovyj*, *klešënyj*/*rasklešënnyj* and *plessirovannyj*. These morphologically integrated loans, which are common in the Russian colloquial speech, may appear in the pre- or postnominal position, as in *bordovyj cvet* 'the claret colour', *klešënye brjuki* 'bell-bottom trousers' and *plessirovannaja jubka* 'a plissé skirt'.

As follows from the overview of the syntactic configurations involving attributive adjectives in German and Russian, Russian exhibits a greater number of patterns for noun modification by adjectives than German. Beyond the few lexical restrictions, pragmatics seems to be the only factor that constrains the position of attributive adjectives in Russian. In contrast, German has one canonical position for attributive adjectives, i.e., the prenominal position. The postnominal position, in which adjectives are morphologically bare, is restricted to specific lexical items and classifying modifiers in special contexts. One of such contexts, the use of adjectives in product names, is found in both languages. Although inflection for the grammatical gender, number and case generally applies to adjectives in both German and Russian, it does not apply to German postnominal adjectives (with the exception of adjectives in split noun phrases with topicalised nouns) and certain borrowed adjectives in Russian. Despite the morphological non-equivalence in the inflection of adjectives observed in these two cases, the syntactic patterns, i.e., the adjacent preposition and the adjacent postposition, are identical in both languages. What is different between German and Russian syntactic patterns of modification is the possibility of distant modification, which exists only in Russian. In other words, only in Russian inflected attributive adjectives frequently occur detached from the head noun.

On the whole, modification by attributive adjectives exhibits similarities and differences in German and Russian: adjectives appear in both the pre- and postnominal position to modify their head nouns, but in German semantic and pragmatic constraints restrict the use of adjectives in the postnominal position more rigorously than in Russian, in which the postnominal position is commonplace in the colloquial speech. Considering this outline, we can predict that in insertional code-mixing, with Russian being the matrix language, Russian adjectives may pre- or post-modify German noun insertions, whereas inserted German adjective-noun combinations would follow the German order, such that the attributive adjective precedes the noun. In order to see whether this prediction is correct, we have to investigate these two types of insertion in the bilingual corpus.

4.3 Adjective-noun combinations in the Russian-German bilingual corpus

In my Russian-German bilingual corpus, the use of attributive adjectives in mixed sentences is observed to be limited to two cases: Firstly, combinations of German attributive adjectives and their German head nouns may be inserted into Russian sentences as embedded-language islands. Secondly, Russian adjectives may modify German nouns pre- and post-nominally. Instances of modification of Russian nouns by German adjectives are absent from the corpus, although German adjective insertions occur in the corpus as (subject) predicatives (see Deuchar 2005, for a similar case in Welsh-English code-mixing).

4.3.1 German adjective-noun combinations in Russian sentences

German adjective-noun combinations appearing in Russian sentences can be analysed as instances of insertional code-mixing (see section 4.1.3). For example, the German noun phrase *gebratene Nudeln* 'fried noodles' in (19) is inserted into a Russian clause structure, namely, into a Russian prepositional phrase.

(19) (LA-1205031A)
ja xoč-u čë-nibud' s *gebraten-e* *nudel-n*
NOM.1SG want-1SG something with fried-NOM.PL noodle-PL
'I'd like something with fried noodles.'

The internal structure of the inserted constituent *gebratene Nudeln* is German: the adjective *gebraten* 'fried' agrees with its head noun in the number and takes

the inflectional suffix *-e*, which marks the nominative or the accusative case. The morphologically marked case deviates from the case projected by the preceding Russian preposition *s* 'with', namely, the instrumental case. This mismatch may be due to the absence of the instrumental case from the German case system. However, since the prepositional phrase is a constituent of a higher level than the German noun phrase, the analysis of the German adjective-noun combination as an insertion holds despite the issue of the structural non-equivalence of the two systems.

Insertions of German nominal constituents containing attributive adjectives, which can be regarded as embedded-language islands from the viewpoint of the MLF model, are the most frequent case in the corpus: a total of 71 adjective-noun combinations were identified, which correspond to 65 lexical types. These extended noun phrases almost invariably lack German determiners. For example, the German inserted noun phrases *chillige Familie* 'chilly family' in (20) and *unbekanntes Gesicht* 'unfamiliar face' in (21) would require indefinite articles *eine* and *ein*, respectively, when used in German sentences because they refer to countable indeterminate objects.

(20) (LS-110125J)
nee u menja *chillig-e* *familie*
no with GEN.1SG chilly-NOM.SG.F family(F)[SG]
'No, I have a chilly family.'

(21) (LA-12022404A)
a mnje *oh* *nee geh-t* *nicht*; vidat' to čto u menja
but DAT.1SG PTCL no go-3SG NEG perhaps because with GEN.1SG
unbekannt-es *gesicht=* da čto *ich* *komm-e* *selten oder*
unfamiliar-SG.N face(N)[SG] yes because NOM.1SG come-1SG rarely or
so= ja
so yes
'But (she says to) me, "Oh no, that's impossible"; perhaps because my face is unfamiliar (to her), yes, because I come rarely or so, yes.'

Even when the referents of the inserted noun phrases are definite objects, German articles are not used, for instance:

(22) (H-100712O)
ei smotr-i kak-ie igrušk-i; smotr-i *na?* gde
PTCL look-IMP.2SG what-NOM.PL toy-NOM.PL look-IMP.2SG PTCL where

klein-er *stern*?
small-SG.M star(M)[SG]
'Ah, look, what (nice) toys! Hey, look, where is the small star?'

(23) (LV-12022404A)
potom oni est' v *historisch-es* *seminar*
then NOM.3PL be[PRS] in historical-SG.N department(N)[SG]
'Then there are some in the department of history.'

(24) (LA-1205034V)
vo-pervyx *baltisch-e länd-er-n* otnosi-l-i-s' ne k
firstly Baltic-PL PL\country-PL-PL/DAT be.part-PST-PL-REFL NEG of
rossi-i pri petr-e perv-om
Russia-DAT.SG.F under Peter-PREP.SG.M first-PREP.SG.M
'First of all, the Baltic countries were not part of Russia under Peter the First.'

The noun phrases *kleiner Stern* 'small star' in (22), *historisches Seminar* 'historical department' in (23) and *baltische Ländern*[11] 'Baltic countries' in (24) all lack definite articles, which would be mandatory in German clauses. Furthermore, the presence of the definite article before these noun phrases in German would require the use of inflectional suffixes of the "weak" declension, i.e., *der klein-e Stern*, *das historisch-e Seminar* and *die baltisch-en Länder*.

Unlike in standard German, the vast majority of attributive adjectives in the data, just as the adjectives in the examples given above, inflect according to the "strong" declension. Of 70 German adjectives, 60 take one of the three suffixes of this declension: *-e* for feminines and plurals, *-es* for neuters and *-er* for masculines; six masculine and neuter adjectives appear with the suffix *-e* of the "weak" inflection, although the determiners requiring it are absent from the respective bilingual clauses; one feminine adjective adds the suffix *-en*, which can be attributed to either the "mixed" or "weak" inflection; and finally, three adjectives lack an agreement marker.[12] The tendency to express gender agreement with the noun by means of suffixes of the "strong" inflection may result from the

[11] The formative *-n* is added to the noun *Länder* apparently erroneously. We can interpret this form in at least two ways: the speaker either double marked the plural on the noun (cf. the standard singular *Land* vs. the standard plural *Länder*, or produced its dative plural form. The latter analysis seems yet rather implausible for two reasons: (i) the noun phrase, being the subject of the clause, should be assigned the nominative case and (ii) the adjective is not inflected for dative (cf. *baltischen Ländern*).

[12] One adjective (*Basler* in *Basler Zoo* 'Zoo Basel') was not included in this count because it does not take any agreement marker in German.

degree of syncretism of this inflectional class, which is the lowest among the three inflectional classes (Wurzel 1984). Apart from the gender, these German inflectional suffixes serve as markers of the nominative or accusative case. It is all the more striking that bilingual speakers select them even when the morphological case projected on the slot in which they are inserted together with their head nouns is neither nominative nor accusative. One instance of this tendency is the mixed prepositional phrase *s gebratene Nudeln* 'with fried noodles' in (19), which has been tackled above, another instance is the prepositional phrase *v historisches Seminar* 'in historical department' in (23). In this latter example, the noun phrase headed by the preposition would have to be marked for the dative case because the whole prepositional phrase has a spatial meaning (the equivalents of the phrase in the dative would be *historisch-em Seminar* if the adjective inflects according to the "strong" class, and *historisch-en Seminar* if it inflects according to the "weak" declension). As the dative case is not expressed morphologically, we could say that it is neutralised. At the same time, each of the inserted nominal constituents in the above examples has a coherent internal structure, which corresponds to that of the embedded language, i.e., German. This circumstance allows us to analyse such insertions as embedded-language islands. Furthermore, we can regard the pervasive use of the inflectional suffixes of the German "strong" declensional class with the adjectives in these insertions as a kind of compromise strategy, for it allows the bilingual speakers to preserve the internal structure of the nominal constituents, while eschewing the use of articles, which are non-existent in Russian but are indispensable to the German case marking system, and thus occasionally sacrificing case distinctions. In other words, structural similarity between the matrix language and the embedded language increases without the violation of the structural integrity of the inserted nominal constituents.

Beside plentiful insertions of adjective-noun combinations, the corpus contains one instance of alternation involving an adjective-modified noun phrase. In this case, see (25), the switched fragment is a fully-fledged noun phrase, consisting of an adjectival modifier and a determiner.

(25) (LA-1205031A)
on id-ët kak *der alleinig-e stark-e*
NOM.3SG.M go-3SG as DET.NOM.SG.M lone-NOM.SG.M strong-NOM.SG.M
kämpfer; *ich bin auf kein-en angewiesen.*
fighter(M)[NOM.SG] NOM.1SG be[PRS.1SG] on no.one-ACC.SG.M dependent
'He counts as a strong lone fighter: I am not dependent on anyone.'

As the well-formed noun phrase *der alleinige starke Kämpfer* 'the strong lone fighter' is part of a conjunction phrase, it could be analysed as an insertion, but it is better apprehended as an instance of alternation, because it occurs at a sight of structural equivalence, namely, after the Russian conjunction *kak* 'as'. This conjunction just as its German equivalent *als* does not govern case, the dependent noun phrase receives the case from the relevant antecedent phrase, here the nominative case from the phrase *on* 'he'. Moreover, the fact that the noun phrase is embedded in discourse, i.e., followed by a clause in the same language as the noun phrase itself is indicative of alternation (cf. Muysken 2000: 104).

A few German adjective-noun combinations in Russian discourse may be reminiscent of alternation, but are better conceived of as insertions. This is the case when a German adjective-modified noun phrase is adjacent to another German insertion. Two instances of this kind are found in the corpus, for example:

(26) (LA-1205034A)
net, èto eščë i *frage der identität* potomu čto u
no this also PTCL question DET.GEN.SG.F identity(F)[SG] because with
nas netu *identität* i oni probuj-ut čerez cerkov'
GEN.1PL no identity and NOM.3PL try-3PL through church[ACC.SG.F]
national-e identität aufbau-en; vot èto vot…
national-ACC.SG.F identity(F)[SG] build.up-INF PTCL this PTCL
'No, this is also a question of identity because we don't have an identity, and they are trying to build up a national identity by using the church; that's what I mean…'

The above utterance contains four German insertions, these include three noun phrases and an infinitive. The noun phrase *nationale Identität* 'national identity' and the infinitive *aufbauen* 'build up', which is adjacent to it, constitute one switched fragment. Since the switched fragment comprises several words and exhibits a non-nested structure (cf. Muysken 2000: 97), it may be reminiscent of alternation. Yet, the analysis as insertion is preferred because the noun phrase is syntactically integrated into the Russian clausal matrix owing to a missing article, required in German, and because the infinitive *aufbauen* 'build up' is part of the verb phrase *probujut aufbauen* 'are trying to build up'. As a rule, however, the instances of German adjective-noun insertions in my corpus represent a straightforward case.

In a few occurrences, German nominal insertions, in which German adjectives generally precede their German head nouns, are further modified. Overall, seven tokens of modification by Russian adjectives and one token of modification by a

German pronominal adjective were identified in the corpus. The latter instance is as follows:

(27) (LS-110517La)
u menja *mein* *ganz-er* *pony*
with GEN.1SG my[NOM.SG.M] whole-NOM.SG.M fringe(M)[NOM.SG]
sgore-l uže
burn.away-PST.SG.M already
'My whole fringe had already burnt away.'

Here, the extended noun phrase *mein ganzer Pony* 'my whole fringe' is inserted in the slot requiring the nominative case and is preceded by the Russian possessive construction *u menja* 'with me'. As the possessive adjective *mein* 'my' goes ahead of the adjective *ganz* 'whole', the latter receives the suffix *-er* of the "mixed" inflectional class, as is expected in German. The more common case, namely, the modification of German nominal insertions by Russian pronominal adjectives, is illustrated below:

(28) (LS-110125J)
u menja tože moj *erst-er* *freund*
with GEN.1SG also my[NOM.SG.M] first-NOM.SG.M boyfriend(M)[NOM.SG]
kogda ja pereexa-l-a nemec by-l
when NOM.1SG move-PST-SG.F German[NOM.SG.M] be-PST.SG.M
'And me, too, my first boyfriend, when I moved [here], was a German.'

In (28), the German noun *Freund* 'boyfriend' is modified by the German adjective *erster* 'first' and the Russian possessive adjective *moj* 'my'. Although the extended noun phrase is mixed, the noun and its two adjective modifiers are congruent in number, gender and case: both modifiers carry nominative singular masculine suffixes. The suffix *-er* of the adjective is attributable to either the "mixed", or the "strong" inflectional class. In parallel with the German noun phrase *mein ganzer Pony* 'my whole fringe' in (27), we can analyse the adjective *erster* as inflected according to the "mixed" type. Nevertheless, it seems more plausible to consider this adjective as "strongly" inflected because, as has been shown above, the inflectional suffixes of the "strong" class clearly preponderate over other agreement markers on German adjectives in the data.

Instances of German adjective-noun insertions preceded or followed by Russian modifiers are given in Table 4.1. While in the first five instances, modification is syntactically unconstrained, postposition is the only option for the last

modifying structure, i.e., a relative clause. With the exception of the adjective *normal'naja* 'normal', the Russian modifiers in the table are all pronominal adjectives: *moja* 'my', *kakaja-n't'* 'some', *kakie-to* 'some', *ètot* 'this', *kotoryj* 'which'. We will return to the first example later in this chapter. Suffice it to say for now with regard to the noun phrases given in the table that the grammatical gender in their parts demonstrates a high degree of incompatibility, whereas the grammatical case is entirely congruent. This may be owing to the fact that the examined noun phrases are marked by the most frequent cases: the nominative, or the accusative case, the markers of which exhibit a high degree of syncretism in German and Russian.

Table 4.1: Attributive modification of German adjective-noun insertions by Russian nominal modifiers. *Note:* The morphosyntactic glosses lack the information about the grammatical case marked on German adjectives, for in each given case the respective inflectional suffix expresses the nominative, or the accusative case.

A_R	$_G$[A	N $_G$]	A_R
normal'naja	*richtige*	*italien*	
normal:NOM.SG.F	proper:SG.F	Italy:(N)[SG]	
'normal, proper Italy'			
moja	*schlechteste*	*note*	
my:NOM.SG.F	worst:SG.F	mark:(F)[SG]	
'my worst mark'			
kakaja-n't'	*süße*	*grußkarte*	
which:NOM.SG.F-any	sweet:SG.F	greeting.card:(F)[SG]	
'some sweet greeting card'			
	kriminelle	*jugendliche*	kakie-to
	criminal:NOM.PL	youth:NOM.PL	which:NOM.PL-some
	'some criminal youths'		
	soziales	*jahr*	ètot
	social:SG.N	jahr:(N)[SG]	this:ACC.SG.M
	'this volunteer gap year with social work'		
	neue	*prüfungsordnung*	kotoryj...
	neue:SG.F	exam.regulation:(F)[SG]	which:NOM.SG.M
	'new examination regulations which...'		

In the analysed sentences Russian is the matrix language, it is therefore not surprising that Russian syntactic patterns such as post-nominal and distant modi-

fication can be applied to German attributive adjectives, although on very rare occurrences. For example, the data contain two instances of German nominal constituents in which the inflected adjectives immediately follow their head nouns, unlike in German. These instances are given below:

(29) (LA-1205034V)
ja xote-l-a dela-t' *sozial-es jahr freiwillig-es*
NOM.1SG want-PST-SG.F do-INF social-SG.N year(N)[N] voluntary-SG.N
èt-ot posle universitet-a nu posle škol-y
this-ACC.SG.M after university-GEN.SG.M PTCL after school-GEN.SG.F
'I wanted to take that volunteer gap year after my studies, I mean after school.'

(30) (LS-110714-1)
A: *das hab' ich auch überleg-t ob ich in*
that AUX.PRS.1SG NOM.1SG also think.about-PTCP whether NOM.1SG in
italien was mach-en was mach-e
Italy something do-INF something do-PRS.1SG

B: *in italien europa-park* ili *italien richtige?*
in Italy Europa-Park or Italy(N)[SG] real-NOM.SG.F

A: ne: normal'n-aja *richtig-e italien*
no normal-NOM.SG.F real-NOM.SG.F Italy(N)[SG]

A: 'I have also thought of doing something in Italy.'
B: 'The Italy in Europa-Park or the real Italy?'
A: 'No, normal real Italy.'

The mixed noun phrase in (29) contains the German noun *Jahr* 'year', two German attributive adjectives *soziales* 'social' and *freiwilliges* 'voluntary' and the Russian demonstrative adjective *ètot* 'this'. The German adjectives occupy both positions adjacent to the noun, whereas in German, these adjectives precede the noun, and the adjective *freiwilliges* is followed by the adjective *soziales*, i.e., *freiwilliges soziales Jahr* 'volunteer gap year (social)'. The relationship between the noun and the attributive adjective immediately adjacent to it, which Seiler (1976) describes as the most intimate one among the noun's semantic relationships to its attributive adjectives, obtains here as well. In other words, the noun *Jahr* 'year' is closely connected to the adjective *soziales* 'social' both syntactically and semantically. It is therefore not surprising that in the mixed phrase, this adjective occupies the same position as in the German phrase, while the adjective *freiwilliges*

'voluntary' moves to the post-nominal position. The syntactic pattern underlying this post-nominal modifier contrasts both German post-nominal constructions: the construction used for names of dishes and products, in which the noun joins the adjective in an appositive relation with zero agreement marking (cf. *Jahr freiwillig*), as well as the increment construction, whose inflected adjective is usually preceded by the determiner (cf. *Jahr, ein freiwilliges*). The usage of the inflected adjective in (29) is akin to the respective Russian pattern. As regards the internal structure of the mixed noun phrase under scrutiny, the components thereof manifest agreement in gender, number and case.

By contrast, the noun phrases *Italien richtige* 'real Italy' and *normal'naja richtige Italien* 'normal, real Italy' in (30) seem to lack agreement in gender. At first glance, the gender feature value of the German noun *Italien* 'Italy', which is neuter, contrasts with the corresponding feature value of the the adjective, which is interpreted as feminine. The analysis of the adjective suffix *-e* on the adjective *richtige* 'real, proper' as a feminine marker is based on the aforementioned observation that German attributive adjectives in inserted nominal constituents systematically inflect according to the "strong" declensional class. In this declensional class, the suffix *-e* marks either the plural or the feminine gender (cf. examples 29, 30 and Table 4.1). If we consider the neuter gender of the noun and the feminine gender of the attributive adjective, we will have to assert a lack of congruence between these parts of the noun phrase. Nevertheless, it may well be that the noun and the whole noun phrase are feminine forms. This analysis is supported by the fact that the Russian adjective *normalnaja* 'normal' of the mixed phrase in the following five is feminine. The speakers treat the modifiers and presumably the noun as feminine forms, possibly as a result of an interference with the Russian equivalent of the German noun *Italija*, a feminine noun. The use of the German feminine suffix *-e* with the adjective *richtig* may thus be explained by the transfer of the corresponding gender feature value from Russian. If the noun *Italien* counts as a feminine, just as its adjective modifiers, we may well suppose that the heads of the noun phrases in (30) as well as their phrasal constituents all manifest agreement in gender and number case, just as the majority of inserted German nominal constituents.

The syntactic pattern of the combination *Italien richtige* is identical with the pattern of *Jahr freiwilliges* in (29): the inflected adjective immediately follows its head noun. As outlined above, this syntactic pattern deviates from both post-nominal constructions available in German and is likely to be modelled on the Russian pattern of post-nominal modification. The second occurrence of the examined nominal constituent in the example is a repetition of the phrase in a rearranged order. That is, the German adjective *richtige* 'real' immediate precedes

the noun *Italien* 'Italy' and combines with the Russian adjective *normal'naja*, resulting in a mixed constituent. The syntactic pattern here complies with the canonical syntactic arrangement of German extended noun phrases.

As to the distant placement of modifiers, another feature of Russian syntax, German adjectives, on rare occurrences, may be placed non-adjacently to their German head nouns. Distant placement applies to pre- and post-nominal adjectives. Below are the only occurrences of German pre-nominal adjectives placed distantly:

(31) (LA-1205034V)
do nego on-i otnosi-l-i-s' *polen-litauisch* vot-vot *union*
before GEN.3SG.M NOM.3-PL
belong-PST-PL-REFL Poland-Lithuanian PTCL Commonwealth
'Before him they belonged [to] the Polish-Lithuanian Commonwealth.'

(32) (LA-12050313A)
...a čto èto *offen-e* by-l-a *lesung*...
but that this open-SG.F be-PST-SG.F reading[SG.F]
'...but that it was an open lecture...'

While the adjective *polen-litauisch* 'Poland-Lithuanian' in (31) is separated from its head noun by the focus particle *vot-vot*, the adjective *offene* 'open' in (32) is detached from its head by the copula *byla* 'was'. Although the pre-nominal position is the default for attributive adjectives in German, just as is the case in (31) and (32), detaching pre-nominal adjectives from their head nouns is not possible. With regard to the internal structure of the constituents, the adjective *offene* in (32) exhibits the agreement suffix *-e*, whereas the adjective *polen-litauisch* 'Poland-Lithuanian' in (31) lacks it. Another bare constituent is the verbal complement. Since the verb *otnosilis'* 'belonged' requires, as a complement, a prepositional phrase headed by the preposition *k* 'to', the prepositionless verbal complement *polen-litauisch union* may be considered bare; cf. *(k) polen-litauisch(e) Union.* Another peculiarity of the examined nominal constituent pertains to the adjective *polen-litauisch* 'Poland-Lithuanian'. This nonce formation deviates from the usual German dvandva adjective *polnisch-litauisch* 'Polish-Lithuanian', which is a habitual collocate of the noun *Union* 'Commonwealth'. The production of the nonce formation may have been triggered by the interference with the synonymous noun *Polen-Litauen* 'Polish-Lithuanian Commonwealth'. A concentration of unusual features in (31), such as bare and deviant forms, may signal processing difficulties and thus challenge the analysis of the split noun phrase *polen-litauisch*

Union as an embedded-language island. With regard to the split noun phrase *offene Lesung* 'open reading', one conspicuous problem is its lexico-semantic nature. In their conversation, the speakers discuss their academic schedule and, in order to refer to a lecture open to students of various courses of study, they use the expression *offene Lesung* instead of *offene Vorlesung* 'lecture', a more appropriate phrase to use in this context.

The corpus contains three instances of German post-nominal adjectives that are detached from their noun heads. While two of them constitute separate turns and thus coincide with turn-taking points, one post-nominal adjective is placed distantly within a continuous utterance; this instance is given below:

(33) (LV-12022413)
A: ili vsjak-ie vot èt-i vot zna-eš gm *vorspeise-n*
or all-NOM.PL PTCL this-NOM.PL PTCL know-2SG HES starter-PL.F
mne nrav-jat-sja
DAT.1SG please-3PL-REFL
V: a *vorspeise* ja ne e-l-a
PTCL starter NOM.1SG NEG eat-PST-SG.F
A: *weinblätt-er* tam *gefüllt* ili *florinis* èto *gefüllt-e*
PL\vine.leaf-PL PTCL stuffed or florinis PTCL stuffed-SG.F
paprika
sweet.pepper

A: 'I also like all sorts of well those well you know hem starters.'
V: 'Well, I haven't eaten any starter.'
A: 'Stuffed well vine leaves or florinis, it's stuffed sweet peppers.'

The above snatch of conversation contains two German adjective-noun combinations inserted into a Russian syntactic structure (lines 4–5): the adjective *gefüllt* 'stuffed' follows the noun *Weinblätter* 'vine leaves', and the same adjective immediately precedes the noun *Paprika* 'sweet pepper'. The noun phrase *Weinblätter gefüllt* 'stuffed vine leaves' instantiates a German pattern used in recipes; in this pattern adjectives follow their noun heads in the relation of apposition (cf. section 1.2). Although in (33) the adjacency between the noun and the adjective seems to be disrupted by the particle *tam* 'well', which functions as a hesitation marker, I argue that adjacency is still maintained here, since hesitation markers may appear virtually at any point of the constituent structure. We may thus consider the noun-adjective combination *Weinblätter gefüllt* 'stuffed vine leaves'

a well-formed embedded-language island. Interestingly, hesitation markers accompany one of the two distant post-nominal adjectives involved in turn-taking. This instance is reproduced as follows:

(34) (LS-110526)
A: nu tam taka-ja mam-a tože tam
well there such-NOM.SG.F mother-NOM.SG.F also PTCL
B: *modern-e*
modern-NOM.SG.F
A: *modern-e mama*
modern-NOM.SG.F mother(F)[NOM.SG]
A: 'Well, they have such a mother there.'
B: 'A modern one.'
C: 'A modern mother.'

In the conversation extract above, speaker B produces the German adjective *moderne* 'modern' in response to speaker A's utterance. Namely, speaker A signals her active word seeking by using the hesitation marker *tam* and speaker B aids her in completing her sentence. From the perspective of function, this phenomenon can be described as a self-initiated other-repair, whereas from the viewpoint of conversation organisation, it exemplifies a collaborative co-construction. Speaker B co-constructs the syntactic structure of the utterance initiated by speaker A. Coincidentally, the co-construction here involves not only a change of interlocutors but also a change of the code: speaker A's sentence is Russian, but the adjacent adjective, uttered by speaker B, is German. Speaker A accepts speaker B's repair proposal by repeating the German noun phrase with the default German word-order pattern, i.e., *moderne Mama* 'modern mother'. The bilingual homophone *mama* in the first line enables speaker B to accomplish the other speaker's utterance in German. In other words, it triggers the mixed co-construction in line two. An orthodox analysis would regard the co-constructed part as an inserted adjective, which, in its turn, would count as the only German adjective insertion in the corpus. Still, it is not impossible to view the German adjective as a modifier of a noun that exists in both languages. Then we would treat it as a part of a German noun phrase. This split noun phrase is similar to the noun phrase *Italien richtige* 'real Italy' in line three of (30) in that the syntactic patterns deviate from the German syntax in each case.

The other detached post-nominal adjective that constitutes a separate turn is reproduced in (35).

(35) (LA-12050313)
C: tut est' tol'ko odin *lesezeichen* za dv-a evro
here be[PRS] only one[NOM.M] bookmark(N)[SG] for two-ACC.M euro
D: *magnetisch-e*?
magnetic-NOM.SG.F
A: da
yes
C: 'Here there is only one bookmark for two euros.'
D: 'A magnetic one.'
C: 'Yes.'

In the above conversation fragment, speaker C is looking at products in an online shop, while her interlocutor, speaker D, is not immediately involved in the search. In line one, speaker C provides her conversation partner with details about a bookmark that she has found. In the next line, speaker D requests her for further information. Speaker C responds to the question by giving a confirming answer. In her question, containing merely the adjective *magnetische* 'magnetic', speaker D elaborates on the previous utterance, namely, the mixed extended noun phrase *odin Lesezeichen za dva evro* 'one bookmark for two euros.' The elaboration consists in adding a further modifier to the noun phrase, the German adjective *magnetische* 'magnetic'. Although the adjective as well as its head noun *Lesezeichen* 'bookmark' come from German, the use of the adjective complies with neither German syntax, nor German morphology. The adjective is detached from its head by the adverbial attribute *za dva evro* 'for two euros', as is possible in Russian but not in German (cf. the analysis of the adjective insertion in 34). The adjective receives the German inflectional suffix *-e*, which can be analysed as a feminine or plural marker of the 'strong' inflection, the typical inflectional class for German adjectives inserted into Russian sentences with their German heads as shown above. The resulting form *magnetische* does not agree in gender with the other constituents of the noun phrase: the Russian numeral *odin* 'one' is masculine, and the German noun *Lesezeichen* is neuter. In other words, the gender of the head noun is neuter, whereas the inflected phrasal components are masculine and feminine. Interestingly, the gender value of the attributive adjective coincides with that of the Russian equivalent of the German noun *Lesezeichen* 'bookmark', i.e., *zakladka*. We can therefore assert that the gender value of the Russian noun is transferred to the adjective, just as is the case with the combination *Italien richtige* 'real Italy' in (30).[13]

[13]The incongruity in gender values between the adjectives in inserted nominal constituents in

As outlined above, a crucial step in an analysis of inserted nominal constituents in bilingual sentences is to determine their word order and internal structure. These characteristics allow us to define a combination of a German noun and a German adjective in an otherwise Russian sentence as an embedded-language island or two subsequent insertions. A classification of German adjective-noun combinations based on these criteria is given in Table 4.2. As can be seen from the table, inserted adjective-noun combinations that follow German word order patterns and exhibit German internal structure make up the largest group among the observed instances. This group includes such items as *kleiner Stern* 'small star', *unbekanntes Gesicht* 'unfamiliar face', *süße Grußkarte* 'sweet greeting card', *kriminelle Jugendliche* 'criminal youths' and many others (the adjectives in these items all take suffixes of the 'strong' inflectional class). These insertions can qualify as embedded-language islands because they are well-formed nominal constituents in German.

Table 4.2: Variation in German adjective-noun insertions in Russian sentences according to their word order and internal structure.

	German word order	deviant word order
German internal structure	61	2
aberrant internal structure	5	3

The next largest group of the four possible variants includes five German adjective-noun combinations whose syntactic patterns comply with the syntactic patterns common in monolingual German but whose internal structure deviates from the German monolingual norm in terms of gender, or case agreement. Four adjectives of this group are inflected and one adjective is bare. Inflected adjectives occur in the combinations *zweite Auto* 'second car', *richtige Italien* 'real Italy', *römische Reich* 'Roman empire' and *guten Wohngegend* 'good neighbourhood'. Notably, the gender values of the adjectives in these combinations parallel the gender values of the nouns' Russian equivalents: *mašina*(F) 'car', *Italija*(F) 'Italy', *imperija*(F), the exception being *guten Wohngegend* 'good neighbourhood'. The gender values thus seem to be transferred, just as in example (35). The only bare adjective observed is part of the insertion *typisch Zwilling* 'typical Gemini'.

bilingual sentences as well as the adjectives in the same constituents in monolingual sentences is not uncommon in situations of language contact. For instance, Muhamedowa (2006) reports that in Kazakh-Russian code-mixing and in Kazakhstan Russian the gender feature is regularly neutralised in adjectives which are part of Russian nominal constituents.

The analysis of this combination as a noun phrase is not trivial. The adjective *typisch* 'typical' is non-inflected in German when it is used predicatively, as in *Das ist typisch Zwillinge* 'This is typical of Gemini'. But an assumption that *typisch* in (36) functions as a predicate is hardly tenable in view of the fact that at the higher level of organisation, the whole noun phrase serves as the complement of the preposition *u* 'with'.

(36) (LS-110316R)
u menja real'no vot kak u *typisch zwilling*
with GEN.1SG really PTCL like with typical Gemini
'With me, it's just like with a typical Gemini.'

Hence, judging by the sentence configuration, the only possibility is to analyse the adjective as a bare attribute. The speaker presumably confuses its predicative and attributive uses. Noteworthy is her consistence in handling the adjective *typisch* as a bare attribute; the corpus contains several tokens of this usage (e.g.„ U nas byl *typisch grüner Salat* 'We had a typical green salad', where the inserted nominal constituent is again not well-formed).

Adjective-noun combinations whose word order patterns and internal structure deviate from German monolingual usage involve the following structures: *Italien richtige* 'real Italy', *Lesezeichen…magnetische* 'magnetic bookmark', and *polen-litauische…Union* 'Polish-Lithuanian Commonwealth'. Finally, two tokens exhibit structural congruence as expected in monolingual German but their word-orders are modelled on Russian patterns: *offene …Lesung* 'open lecture' and *Mama …moderne* 'modern mother'. Overall, eleven adjective-noun combinations, which corresponds to barely 14% of all the instances, deviate in one of the examined features or both from the corresponding combinations in monolingual German. These instances are not counted as embedded-language islands and will not be further examined in the subsequent corpus analysis.

4.3.2 Mixed nominal constituents with German nouns and Russian adjectives

Modification of German nouns by Russian adjectives is common in my bilingual corpus, although the number of its occurrences falls behind the occurrence frequency of German adjective-noun combinations described above. As many as 41 German nouns were identified which combine with Russian attributive adjectives in the corpus. These adjectives can modify German nouns both pre- and post-nominally. With 35 instances in the corpus, pre-nominal modification represents the predominant type. It is illustrated by the following examples:

(37) (LA-120503-9G)
u neë prosto pozdn-ij *pubertät* nača-l-sja
with GEN.3SG.F simply late-NOM.SG.M puberty(F) begin-PST.SG.M-REFL
'She has just hit a late puberty.'

(38) (LS-101221J)
na sledujušč-ej *haltestelle* vy dolžn-y vylez-ti
on next-PREP.SG.F stop(F)[SG] NOM.2PL obliged-PL get.off-INF
'You have to get off at the next stop.'

In (37), the Russian adjective *pozdnij* 'late' combines with the German noun *Pubertät* 'puberty' as a pre-nominal attribute. Although the lexical gender of the inserted German noun is the feminine, the speaker treats it as a Russian masculine noun, as is evident from the inflectional suffix of the respective agreeing adjective and the zero marking on the inserted noun. The Russian attributive adjective in (38) also precedes the German head noun. In this case, the lexical gender of the inserted noun *Haltestelle* 'stop' coincides with that of its Russian equivalent, i.e., *ostanovka*. The adjective thus receives the inflectional suffix marking the feminine gender as well as the prepositional case, since the mixed noun phrase is embedded into a prepositional phrase and the preposition *na* 'na' governs the prepositional case here. However, it is impossible to determine whether the form *Haltestelle* is an uninflected noun, like in German, or whether the stem-final schwa of the German noun is reanalysed as the Russian inflectional suffix of the prepositional case, i.e., {*Haltestell-*} + {*-e*} (cf. Ždanova & Trubčaninov 2001: 276).

In rare occurrences, Russian adjectives modify German noun insertions post-nominally. As few as five instances of this modification type were identified in the corpus, two of these instances are given below:

(39) (Fr-110801-1)
A: a podružk-i u tebja russk-ie ili nemeck-ie?
and friend-NOM.PL.F with GEN.2SG Russian-NOM.PL or German-NOM.PL
B: nu *unterschiedlich*, *mischmasch* cel-yj
PTCL variable jumble(M)[SG] whole-NOM.SG.M
A: 'And are your friends Russian or German?'
B: 'Well, it varies – a whole jumble.'

Here, the German noun *Mischmasch* 'jumble' in line two is followed by the Russian attributive adjective *celyj* 'whole'. The speaker handles this insertion as a

masculine noun, as in (37), maintaining agreement marking between the inserted German noun and the Russian adjective.

The bilingual corpus contains one instance of a German noun modified by Russian adjectives pre- and post-nominally, namely:

(40) (LS-110125R)
jennifer, ona *schlampe* redkostn-aja; ona real'n-aja
name NOM.3SG.F slut(F)[SG] rare-NOM.SG.F NOM.3SG.F real-NOM.SG.F
schlampe redkostn-aja
slut(F)[SG] rare-NOM.SG.F
'Jennifer is a dirty slut, she is a real dirty slut'.

The utterance in (40) begins with a left dislocation *Jennifer* and continues with two clauses. The first clause contains the German noun *Schlampe* 'slut' and its Russian post-nominal attributive modifier *redkostnaja* 'rare'. The second clause echoes the first clause in such a way that the pre-nominal attributive modifier *real'naja* 'real' is added to the mixed noun phrase *Schlampe redkostnaja* 'a rare slut' from the previous clause.

All in all, more than 85% of Russian attributive adjectives are used in pre-position to the modified German nouns. When compared to the inserted German nominal constituents, in which the adjectives appear in pre-position in 93% of all occurrences, we attest a slight decline in using pre-modification with bilingual, or mixed, nominal constituents in favour of post-modification. As such, this is not surprising because in the case of bilingual nominal constituents, Russian, being the matrix language, provides not only the modifying adjectives but also patterns of modification. Interestingly, the finding that the vast majority of Russian adjectives in bilingual nominal constituents clearly prefer pre-position contrasts with the observation that in the colloquial Russian post-modification is as common as pre-modification (Lapteva 1976: 207; Zemskaja 1979: 148–149). In this regard, it would be promising to investigate syntactic patterns of modification in both bilingual and Russian monolingual sentences from the bilingual corpus and to compare these patterns with the patterns observed in colloquial Russian as spoken in Russia, but this undertaking is beyond the scope of the present study.

4.3.3 Frequency distribution of the structures in the data set

German adjective-modified nominal constituents and German nouns modified by Russian adjectives appear in otherwise Russian sentences in the bilingual corpus

with varying frequencies. As a rule, a specific German adjective-noun combination occurs in Russian discourse only once, but some combinations appear several times. For example, the German nominal constituent *baltische Länder* 'Baltic states' occurs three times in Russian sentences (all the occurrences are registered in a conversation passage of one minute). The German word combinations *Badische Zeitung* 'Baden Newspaper' (the title of a local newspaper covering the Black Forest region)[14], *freiwilliges Jahr* 'volunteer gap year' and *soziales Jahr* 'gap year for social work' appear in the Russian context twice each. On the whole, 60 specific German adjective-noun combinations in the data correspond to 56 various types.

As one might expect, German nouns that combine with Russian attributive adjectives appear in Russian discourse more frequently German nominal constituents, since single lexeme insertions, especially noun insertions, prevail over constituent insertions, or embedded-language islands, in number (cf. §5.2.1). The frequencies with which embedded-language nouns appear in Russian sentences are thus more variable. A substantial amount of German nouns occur in bilingual sentences sporadically, and only a relatively small portion thereof is recurrent. Of the 39 instances of noun insertion identified in the above analysis, 33 nouns correspond to different lexemes. The frequencies of these German lexemes in Russian discourse was measured by counting every token of a specific German lexeme embedded in the Russian context in the bilingual corpus. The results are reported in Table 4.3. As can be seen from the distribution of the investigated German nouns, more than a half of them occur in bilingual sentences only once. In other words, nonce items, or hapax legomena, constitute the largest group. The most frequent noun that appears ten times in the Russian discourse is *Handy* 'mobile phone'. The nouns whose frequencies range between five and ten comprise lexemes such as *LKW* 'lorry', *Mischmasch* 'jumble', *Spur* 'lane' and *Gewicht* 'weight' (the lexical items are listed in the order of decreasing frequency). Hence, the examined lexical items encompass recurrent and nonce items.[15]

An analysis of code-mixing in which nonce and recurrent insertions are treated in the same fashion might be problematic inasmuch as embedded-

[14]The inclusion of this combination in the data set may be questioned since, owing to its purely idiomatic meaning, it deviates significantly from the other investigated adjective-noun combinations and may represent an inappropriate unit of analysis. However, under the usage-based view, word combinations with both idiomatic and compositional meanings are stored and retrieved in the same fashion (for more details, see Chapter 2 and particularly §2.2.3). It may well be that the language user represents, in her mental lexicon/grammar, not only the proper name *Badische Zeitung* as a holistic unit but also its parts and the links between them.

[15]The distribution of these nouns is similar to the distribution of nouns inserted in Russian prepositional phrases, which are scrutinised in the following Chapter 5.

Table 4.3: Frequencies of German noun insertions in Russian sentences as distributed in bilingual corpus.

Word frequency		Number of lexemes	
Absolute	Relative	Absolute	%
1	0.00004	17	51.5
2	0.00008	5	15.2
3	0.00012	4	12.1
4	0.00016	2	6.1
6	0.00024	2	6.1
7	0.00028	2	6.1
10	0.00040	1	3.0
Total		33	100.0

language items recurrent in the matrix language discourse may well be (becoming) established loans (cf. Backus 2013, Myers-Scotton 1993, Poplack et al. 1988, Poplack & Dion 2012, Poplack 2018). The question of whether a given lexical item is undergoing conventionalisation in the variety of Russian spoken in Germany cannot be addressed here at length on the basis of the following considerations: First, in order to examine the conventionalisation of a lexical item a larger sample is indispensable. Secondly, frequency counts alone may not always be sufficient for determining an item's status and need to be complemented by psycholinguistic evidence (cf. Blumenthal-Dramé 2012). A conceptual hurdle concerns methods for establishing the status of a lexical item as either recurrent or nonce by measuring its frequency. Most studies which distinguish between established loans and nonce-borrowings usually conceive of these categories as discrete, and utilize threshold heuristics to identify established loans. However, as is evident in Table 4.3, the distribution of German nouns in the Russian discourse, just as any frequency distribution, is inherently gradient. Despite these cautionary notes, I will adopt the mainstream dichotomous approach to borrowing in my analysis because it is largely uncontroversial, and the task of distinguishing between established and nonce-borrowings is incidental to the main purpose of this chapter.

A practical question that arises in this context is where to draw the line between items occurring more than once and frequent items. Poplack et al. (1988) suggest an absolute frequency of ten tokens as a cut-off threshold for a word to qualify as a recurrent item and therefore a potentially established loan. The proposed solution is based on the large size of their corpus, which encompasses ap-

proximately 2.5 million words (ibid., 98). Expressed in relative terms, the threshold frequency corresponds to the value of 0.000004. An application of this threshold to a small corpus, such as mine, is not feasible because not even a single word in the corpus would count as an established loan. The frequency threshold for German noun insertions in Russian sentences was set at the relative frequency of 0.0002, which amounts to the absolute frequency of five tokens. Hence, German nouns which appear at least five times in Russian discourse were considered frequent and were removed from the data set. The excluded items comprise eight tokens of the aforementioned lexemes, which correspond to 7.9% of the data set.

To summarise, German lone nouns are inserted into Russian sentences at varying rates. Lexical items that appear in Russian discourse particularly often may represent established loans. The identification of words potentially belonging to this category involved an operationalisation of occurrence frequency as a diagnostic. In order to enhance the homogeneity of the sample, German lexical items frequently appearing in Russian discourse were removed from the data set, which is subject to factorial analysis in the subsequent sections.

4.4 Factors contributing to the variation in switch placement

The remaining chapter will investigate the role of lexical frequency and lexical chunks, or multiword units, in regulating the choices speakers make when switching their languages within an adjective-modified noun phrase or at the phrase boundary. Firstly, I will consider the individual frequencies of the adjectives and nouns involved in code-mixing. Secondly, I will address the question of whether the identified adjective-noun combinations, of which the greater part is German, represent chunks, or recurrent collocations. In order to approach this question, co-occurrence frequency will be measured and analysed, and a statistical association between the noun and the adjective will be computed by using Mutual Information (MI). Finally, a generalised linear regression model will be employed to explore the interplay between the examined factors and to assess their individual contributions to the variance of the data.

4.4.1 Frequency of the adjective

Previous research has shown that word frequency exerts a facilitatory effect in language production (Oldfield & Wingfield 1965) and operates on the lexeme level (Jeschiniak & Levelt 1994). This means that frequent words are more accessible

in production. For the examined syntactic context, I hypothesise that Russian adjectives modifying German noun insertions are highly frequent, whereas German adjectives fulfilling the same function are less frequent. As "[t]he most frequent words in the language are words with grammatical, abstract, or general meanings" (Richards 1970: 88), we can relate the hypothesis to the aforementioned specificity continuum suggested by Backus (1996, 2001). He assumes that in code-mixing, insertions tend to exhibit a high degree of semantic specificity. In other words, German adjectives inserted together with their head nouns into Russian sentences will demonstrate specific rather than general meanings and consequently tend to be infrequent, whereas Russian adjectives, which come from the more activated matrix language, will have general meanings and high token frequencies. Additionally, this hypothesis relates to the assumption that a word "may be selected and subsequently trigger the other elements, because of their collocational entrenchment" (Backus 1996: 126).

To test this hypothesis, frequencies of the adjectives in the identified noun phrases needed to be established. This task usually requires the use of large corpora. As my bilingual corpus would not suffice for such a task, an approximation of the distributions of adjective-noun combinations in German was inevitable. Hence, the frequencies of German adjectives were obtained from deWaC, a large German corpus containing around 1.6 billion words (Baroni & Kilgariff 2006). Given that this corpus is chiefly based on written language, the measured frequencies can only be considered as rough approximations of spoken language. Regrettably, no corpus of spoken German that matches deWaC in size was available. However, considering the age and education of the participants in the study, we can assume that they have received large portions of their German input from written sources as well. As for the Russian adjectives involved, their frequencies were measured in two Russian corpora: the corpus of Yevgen Matusevich (Matusevych et al. 2013) and the Russian National Corpus (RNC; 2003–2014). The Matusevich corpus contains Russian subtitles for foreign-language films, i.e., transcribed spoken texts, and therefore corresponds most closely to the variety of Russian analysed here. However, its size is not comparable with that of deWaC: it contains some 78,170 thousand words, whereas deWaC consists of 1,278 million words, i.e., deWaC has 16 times as many words as the Matusevich corpus. In order to balance frequency counts in the German and Russian corpora, the frequencies of the adjectives in the Matusevich corpus were added to the frequencies of the same adjectives in the RNC, which contains about 230 million words and is thus only 5 times smaller than deWaC. The measured word frequencies had to be normalised since the German and Russian corpora still differ in size.

As the next step, the frequencies were logarithmically transformed, as suggested in Baayen (2008: 31). Table 4.4 gives those noun phrases under investigation whose adjectives have the lowest and highest frequencies of occurrence in the corpora (the low frequency values are owing to normalisation). As can be seen from the table, most adjectives in the low-frequency range are German and most adjectives in the high-frequency range are Russian. However, both groups contain adjectives that deviate from these tendencies: the Russian adjective *pozdnij* 'late' is rare in the used Russian corpora, and the German adjective *neu* 'new' is rare in deWaC. This circumstance necessitates a large-scale comparison of the adjectives' frequency values and the languages in which they are realised.

Table 4.4: German and mixed noun phrases, ranked in order of lowest (above) and highest (below) frequencies of the adjectives involved; the values are normalised and transformed logarithmically.

Noun phrases	F_A
chillige Familie 'chilly family'	−16.364
türkise Farbe 'turquoise colour'	−15.804
standesamtliche Hochzeit 'civil marriage'	−14.634
gebratene Nudeln 'fried noodles'	−12.959
pozdnij *Pubertät* 'late puberty'	−12.942
russkij *Besitzer* 'Russian owner'	−6.770
Bekannter xorošij 'good acquaintance'	−6.745
bol'šaja *Flamme* 'big flame'	−6.613
internatsional'nye *Gerichte* 'international dishes'	−6.575
neue Prüfungsordnung 'new examination regulations'	−6.472

The relationship between the languages of the examined adjective realisations and their frequency values is represented in Figure 4.1. The horizontal axis is used for the frequency values of the adjectives under scrutiny, whereas the vertical axis is reserved for the dependent binary variable "switch placement". Its values zero and one stand for a switch within the phrase and a switch at the phrase boundary, respectively. The switch is placed within the phrase when the adjective is realised in Russian, and it is located at the phrase boundary when the language of the adjective is German. The line depicting the relationship between the two variables is a Lowess curve, which represents a function describing the deterministic part of the variation in the data and is generated by locally weighted scatterplot smoothing, a local regression method (Cleveland & Devlin

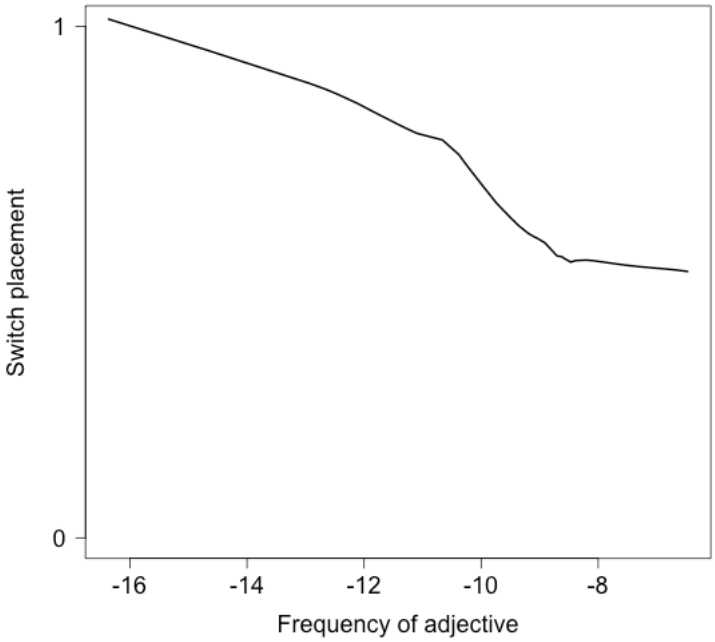

Figure 4.1: The relationship between switch placement and frequency of the adjective. The values of 0 and 1 on the *y*-axis stand for switching within and outside the noun phrase, respectively. The frequency values on the *x*-axis are on the logarithmic scale.

1988). The near straight line, which is steadily inclined downward to the right, bends at two points: at −10.68 it begins to plunge sharply until its undulation point at −8.48, where it begins to plateau. This means that with the frequency of the adjective being low, the overall propensity is to select a German adjective and thus produce a German constituent. In other words, the likelihood for using a German adjective rises with the frequency of the adjective decreasing. The opposite also holds, namely, the higher the frequency of the adjective, the higher the probability of using a Russian adjective. However, at the log frequency of −8.48 or higher, no clear preference for the language of the adjective is observed.

Interestingly, several of the examined lexemes exhibit differing frequencies when determined in deWaC and when measured in the Russian corpus, which consists of the aforementioned Matusevich *et al.* corpus and the RNC. Examples of such adjectives and their frequencies are given in Table 4.5. We can infer from the table that some adjectives, such as *normal* and *normal'nyj*, *klein* 'small' and *malen'kij* as well as *letzt* 'last' and *poslednij*, occur in the German corpus

at similar rates as in the Russian corpus, while others are used with differing frequencies. Among such adjectives we find *neu* 'new' and *novyj* as well as *gut* 'good' and *xorošij*. The word *neu* 'new' appears 1.6 times more often in deWaC than its equivalent occurs in the Russian corpus, and the item *xorošij* 'good' is used 1.4 times more often in the Russian corpus than its equivalent in deWaC. Such recurrent lexemes, realised in Russian in one instance and in German in another instance, have quite different frequencies because their occurrences are counted in two different corpora. For example, the adjective *nächst* in *nächste Woche* 'next week' has the frequency 12.92 on the logarithmic scale, whereas its Russian equivalent *sledujuščij*, used in *sledujuščaja Haltestelle* 'next stop', has the frequency of 13.6. Based on these counts, when deciding between a German adjective and its Russian equivalent, the choice would be biased towards the Russian lexeme, only because its frequency was measured in another corpus. To avoid discrepancies between the utilised corpora, which may lead to inconsistencies in frequency counts, all frequencies are obtained from one and the same corpus. As German items form the bulk of the adjectives under scrutiny, it would be conclusive to use the German corpus as the only source of frequency values for the investigated items. For Russian items, translation equivalents are employed, so that the adjective 'next', for example, has the frequency value 12.92 for each realisation: *nächst* and *sledujuščij*.

Table 4.5: Russian and German realisations of the examined lexemes and their normalised relative frequencies in the German deWaC corpus and the Russian YM&RNC corpora (i.e., the Matusevich et al. corpus and the Russian National Corpus).

Lexeme realisations in Russian and German	$F_{YM\ \&\ RNC}$	F_{deWaC}
'common' *obščij – allgemein*	3.901	2.095
'good' *xorošij – gut*	11.770	8.164
'last' *poslednij – letzt*	6.526	5.696
'new' *novyj – neu*	9.282	15.460
'next' *sledujuščij – nächst*	2.734	4.257
'normal' *normal'nyj – normal*	0.870	0.862
'small' *malen'kij – klein*	6.890	5.940

Figure 4.2 illustrates the relationship between switch placement and the frequency of the adjective (see Appendix IV for the complete list of the examined items and the corresponding values). It is conspicuous that the shapes of the

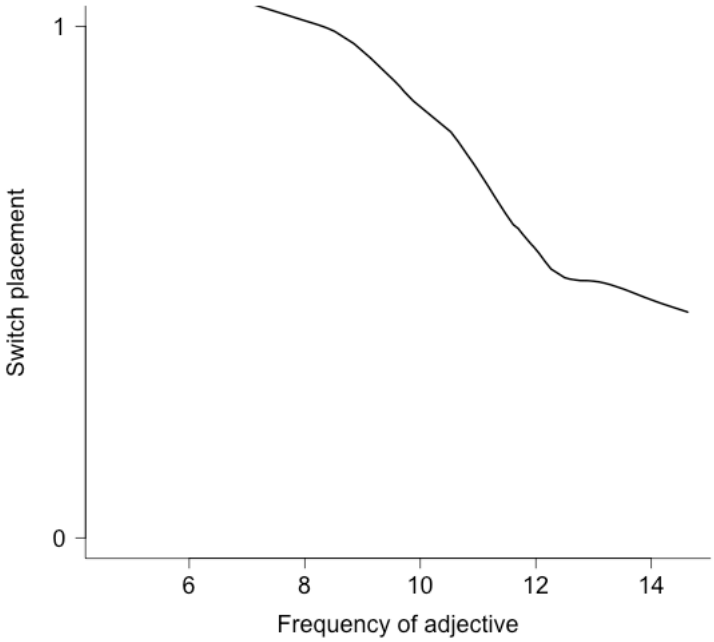

Figure 4.2: The relationship between switch placement and frequency of the adjective, as measured in deWaC. The values of 0 and 1 on the y-axis stand for switching within and outside the noun phrase, respectively. The frequency values on the x-axis are on the logarithmic scale.

lines in Figures 4.2 and 4.1 are analogous, although the slopes of the lines are different. This means that the frequency of the adjective exerts a similar effect on switch placement, regardless of the particular distribution in a given corpus. It therefore seems appropriate to use deWaC as a "benchmark" corpus. In doing so, I would be able to avoid possible inconsistencies in frequency counts owing to the differences in the given corpora and to thus ensure an adequate treatment of the adjectives by considering only the frequencies measured in the "benchmark" corpus.

To summarise, while low-frequency adjectives tend to be expressed in German, high-frequency adjectives exhibit a propensity to be realised in Russian, the matrix language. The analysis of frequency is conducted in the German and Russian corpora and shows that the frequencies with which the Russian and German realisations of several lexemes occur in the respective corpus may differ, depending on the utilised corpus. As argued above, in the subsequent statistical analysis I will consider only the frequencies obtained from the "benchmark" corpus, i.e.,

the German deWaC corpus, since the vast majority of the investigated adjectives are German.

4.4.2 Frequency of the noun

Following the fact that frequent words are more accessible in language production and assuming that interactions between lexemes seem to control syntactic patterns (MacWhinney 1997: 115), we could hypothesise that frequent words can activate lexico-grammatical patterns in which they occur faster than rare words. In other words, lexico-grammatical patterns associated with frequent words, such as collocations and more abstract syntactic constructions, may be highly accessible. In the context of this chapter, high-frequency German nouns are assumed to trigger their typical adjective collocates more often than low-frequency nouns. Hypothesis testing again involves corpus analysis. Since all the nouns involved in the noun phrases under investigation are German, their frequencies were determined in the deWaC corpus. Table 4.6 illustrates examined noun phrases whose nouns occur with the highest and lowest frequencies in deWaC. As shown in the table, both frequent and rare German nouns may combine with German adjectives on a regular basis. We can thus conclude from the few instances given that identifying a tendency in each of the groups seems to be impossible. However, a large-scale comparison may be promising.

Table 4.6: German and mixed noun phrases, ranked in order of lowest (above) and highest (below) frequencies of the nouns involved.

Noun phrases	F_N
konkretnyj *Meister[lehr]gang* 'concrete master craftsman's course'	33
gefundene Kneipentour 'invented pub-crawl'	147
ausgebildeter Polizeihund 'trained police dog'	180
Weinblätter gefüllt 'filled vine-leaves'	266
krasivyj *Saunalandschaft* 'beautiful sauna facilities'	313
normal'naja *Arbeit* 'normal work'	635,026
letzte Arbeit 'last work'	635,026
soziales Jahr 'gap year for social work'	1,034,532
freiwilleges Jahr 'volunteer gap year'	1,034,532
nächstes Jahr 'next year'	1,034,532

An examination of the relation between the frequency of the noun and switch placement is represented in Figure 4.3. The logarithmically transformed fre-

quency of the noun is on the horizontal axis, whereas the vertical axis is reserved for the dependent binary variable "switch placement", as is the case with the previously discussed factor. The values zero and one of the dependent variable stand for a switch within the phrase and a switch at the phrase boundary, respectively. The curve representing the relationship between the two variables is a Lowess curve (see 4.4.1). Although the line has several inflection points, most of it runs roughly parallel to the *x*-axis. The line begins to curve upwards only at the point of 10.2 on the log scale (i.e., the frequency of 26,903 in deWaC). In other words, noun frequency seems to influence switch placement only in the range of frequent words, which exhibit the frequency of 26,903 or higher in the given corpus. This means that recurrent nouns co-activate lexemes that regularly combine with them in German. As a result, the noun phrase is realised in German and the switch is placed at the phrase boundary. However, a reverse effect for low-frequency nouns cannot be found. In order to establish the relevance of the factor "noun frequency" for contributing to the overall variance in the data, a multifactorial analysis is conducted below.

In summary, lexical frequency does appear to play a role in code-mixing. German adjective-noun combinations seem to be inserted into Russian sentences when their adjectives exhibit low frequencies and their nouns are on the contrary high-frequency words. However, the frequencies with which the parts of such a combination appear in a German corpus may be insufficient for explaining multiword insertions. The crucial factor responsible for the occurrence of multiword insertions in code-mixing could rather be their unit status in the corresponding language.

In order for a word combination to count as a multiword unit, or a chunk, the words comprising it have to appear together with a relatively high frequency. Such recurrent word combinations exhibit not only syntagmatic stability but also semantic coherence (cf. Bybee 2010: 136). In a defined linguistic context, such as the adjective-modified noun phrase, corpus frequency of a word combination, or word string frequency, may be a reliable indicator of its unit status (Heylen & De Hertog 2014). Nevertheless, word string frequency is a controversial metric for determining multiword units. For example, based on psycholinguistic judgements of unithood, Simpson-Vlach & Ellis (2010) assert that Mutual Information (MI), a statistical measure of association, provides a better grip on semantically coherent units than word string frequency. In an attempt to underpin the nature of multiword insertions in code-mixing, subsequent analysis will consider both measures of semantic coherence: frequency of co-occurrence and mutual information.

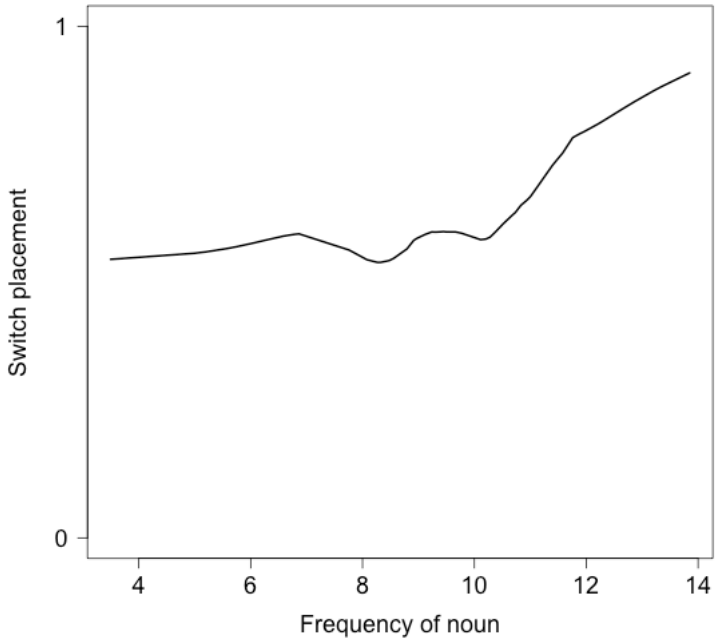

Figure 4.3: The relationship between switch placement and frequency of the noun. The values of 0 and 1 on the *y*-axis stand for switching within and outside the noun phrase, respectively. The frequency values on the *x*-axis are on the logarithmic scale.

4.4.3 Frequency of co-occurrence

Following Bybee's (2002a: 112) Linear Fusion hypothesis, which states that "items used together fuse together", switch placement in the context of the adjective-modified noun phrase can be assumed to be a function of the frequency with which specific adjectives and nouns appear together in the corresponding language. Diessel (2016) attributes this effect to automatisation, which can be defined as a cognitive mechanism whereby sequential activities become uncontrolled, automatic processes[16]. According to Diessel (2016), linguistic elements, which naturally occur in sequence, represent sequential information and are thus subject to automatisation. Repetition of strings of linguistic elements leads to the gradual emergence of processing units, or chunks of linguistic elements. The view of chunks as units emerging from usage processes needs to be complemented by the consideration that many chunks are rote learnt as units already

[16]Bybee (2010) uses the term *chunking* instead of *automatisation*

in the process of language acquisition (see Chapter 2, for details). In this vein, it could be argued that in the context of the adjective-modified noun phrase, a particular adjective and a specific noun exhibit a strong sequential link and form a unit if the frequency with which this noun is used with the adjective is high. I hypothesise that in code-mixing, a switch in such a unit would be unlikely (cf. Backus 1996: 125–131 and Boumans 1998: 386, who come up with similar suggestions but do not subject them to systematic analysis). Conversely, if the association between an adjective and a noun is weak, i.e., they appear together at a low rate, the probability of a switch between them is high.[17] Testing these hypotheses required obtaining co-occurrence frequencies of the examined adjective-noun combinations in deWaC.

The corpus analysis of German adjective-noun combinations was straightforward. The frequency of a specific adjective-noun combination was determined by counting its occurrences in deWaC, while disregarding some of the variability in its morphological forms on purpose. This variability pertains, in the first place, to the modifying adjective, since its form depends not only on the morphological case marked on the noun phrase but also on the presence/absence of preceding determiners. For example, the phrases *nächst-es Jahr* 'next year', *(das) nächst-e Jahr*, *(dem) nächst-en Jahr* involve the same lexical items *nächst* and *Jahr* but exhibit differences in the marking of the adjective.[18] Nevertheless, they are all considered instantiations of one specific underlying collocation for two reasons. First, the combination of these words is syntagmatically stable, *i.e.,* [(DET) *nächst*-AGR *Jahr*(-GEN)], and second, its overall frequency in the corpus is high. As concerns the grammatical number, items differing in this category were handled separately.[19] On this account, only the morphological variants of the underlying collocation in the singular, or in the plural were taken into consideration; their individual frequencies were added together. The same procedure applied to mixed noun phrases, headed by German nouns and modified by Russian adjectives, the only difference being in the use of German equivalents of the Russian adjectives.

[17] These hypotheses may be related to the work investigating disfluency placement in spontaneous speech (e.g., Schneider 2014).

[18] The only morphological contrast between singular noun forms is the distinction between the genitive case and the non-genitive cases. The only difference between plural noun forms is the opposition between dative and non-dative forms. Yet, case marking on the noun was also discarded.

[19] Corpus linguistic studies provide tangible evidence that different number values of a noun may frequently result in different collocation sets (e.g., Sinclair 2003: 167–172). That is, singulars and plurals may display different collocational tendencies.

The results for ten most frequent adjective-noun combinations are reported in Table 4.7. As is seen from the table, high-frequency adjective-noun combinations mostly but not invariably repel switching. The combination *sledujuščij Tag* 'next day' is the only adjective-noun combination in the table that is not produced as a holistic unit but is interrupted by a switch. However, the general tendency is in favour of the proposed hypothesis. The converse hypothesis that code-switches tend to occur in noun phrases between adjectives and nouns if these adjective-noun combinations are infrequent also seems to hold, although not all the data support it. Recall that according to this hypothesis, German low-frequency combinations will not emerge in bilingual sentences, and only mixed combinations, with more accessible Russian adjectives, will be produced instead. Adjective-noun combinations whose co-occurrence frequencies, as determined in deWaC, are in the lowest range are listed in Table 4.8. (It is important not to forget that for combinations of German nouns with Russian adjectives, German equivalents of the actual adjective realisations were used.) The table contains five adjective-noun combinations in which both words are realised in German and five combinations in which only the nouns are German. Judging by the low frequencies of these combinations, it could be argued that they represent free word combinations. Such combinations are essentially unconstrained lexically, and therefore, the adjectives in them may be expressed either in Russian, or in German. An exception to this tendency is the combination *ausgebildeter Polizeihund* 'trained police dog', which may be regarded as a lexical chunk by virtue of its high semantic coherence, and we can possibly get a grip at this chunk when another measure, such as Mutual Information, is utilised to model the strength of association between words. This will be done in the next section.

The corpus analysis revealed that some of the German noun phrases with adjective modifiers do not occur in deWaC. Among them are the following phrases: *chillige Familie* 'relaxed family', *freie Bundesländer* 'free federal states' (in reference to the highly autonomous estates of the Holy Roman Empire) and *gefundene Kneipentour* 'found pub crawl' (*gefunden* 'found' is possibly confused with *erfunden* 'devised'). Furthermore, German equivalents of the following mixed phrases are not attested in the corpus: *ogromnyj Titel* 'huge header' (G *riesig*), *sportivnye Sachen* 'sporty things' (G *sportlich*), *russkij Besitzer* 'Russian owner' (G *russisch*), *obščij Hochdeutsch* 'general Standard German' (G *allgemein*) and *konkretnyj Meistergang* (the word *Meistergang* is not found in the corpus itself, but the conversation context clarifies the use of this word: the speaker substitutes the word *Meisterlehrgang* with the given shorter form) 'concrete master craftsman's course' (G *konkret*)[20].

[20]It is noteworthy that neither was the collocation *konkreter Meisterlehrgang* attested in the cor-

Table 4.7: German and mixed noun phrases, ranked in order of their frequencies in the deWaC corpus.

Noun phrases	$F_{A\text{-}N}$
nächstes Jahr 'next year'	27785
sledujuščij *Tag* 'next day'	22927
katholische Kirche 'Catholic Church'	20856
nächste Woche 'next week'	10499
erstes Semester 'first term'	3185
gute Nacht 'good night'	3031
nationale Idenitität 'national identity'	2878
alte Bundesländer 'old federal states (of Germany)'	1445
letzte Arbeit 'last work'	1262
bares Geld 'cash money'	1148

Table 4.8: German and mixed noun phrases, ranked in order of their frequencies in the deWaC corpus.

Noun phrases	$F_{A\text{-}N}$
russische Party 'Russian party'	3
krasivyj *Saunalandschaft* 'beautiful sauna facilities'	3
ausgebildeter Polizeihund 'trained police dog'	3
bednyj *Hausmeister* 'poor caretaker'	2
real'naja *Schlampe* 'real slut'	2
Schlampe redkostnaja 'rare slut'	2
normales Klo 'normal loo'	1
lebendes Fragezeichen 'living question mark'	1
novyj *Trockner* 'new drier'	1
süße Grußkarte 'sweet greeting card'	1

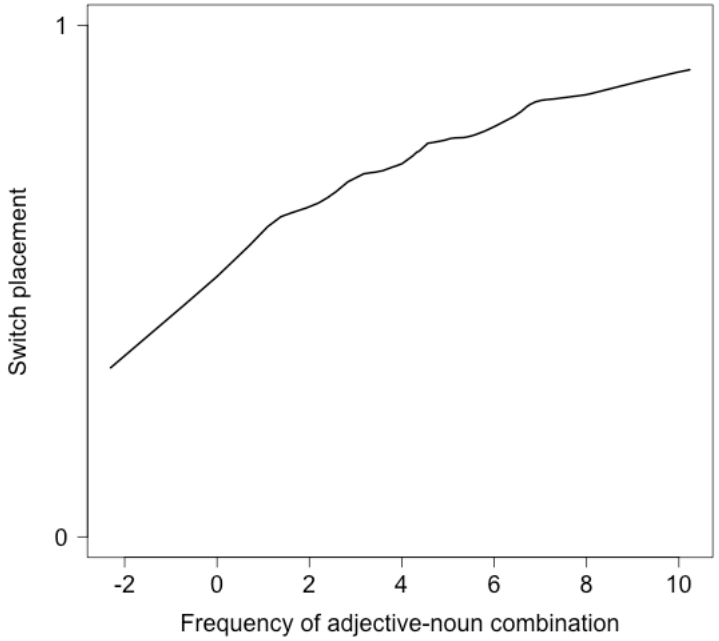

Figure 4.4: The relationship between switch placement and frequency of adjective-noun combinations in deWaC. The values of 0 and 1 on the y-axis stand for switching within and outside the noun phrase, respectively. The frequency values on the x-axis are on the logarithmic scale.

The relation between switch placement and the frequency of an adjective-noun combination, or co-occurrence frequency, is plotted in Figure 4.4. The horizontal axis is reserved for the frequency of co-occurrence, whose values are logarithmically transformed, whereas the vertical axis is used for the binary variable "switch placement". As in the previous figures representing the relationships between switch placement and frequency distributions (see, for example, Figure 4.2), the line depicting the relationship between the two variables is a Lowess curve. The curve, which is inclined upwards to the right, demonstrates that with growing co-occurrence frequency, an across-the-board gradual increase in probability for placing a switch at the phrase boundary is expected. The absence of dramatic inflection points on the curve signifies that the tendency is stable over the whole line. In other words, the higher the frequency with which a specific noun is used with an adjective, the more likely a switch at the boundary of the extended noun phrase. This finding may be viewed as a confirmation of the Linear

pus.

Fusion hypothesis; that is, if two words appear together with a high frequency, they form a multiword unit, which is activated as a whole in production. Additionally, by using the complete data for the correlation between switch placement and frequency of co-occurrence, evidence could be found in support of the reverse hypothesis, according to which a switch between an adjective and a noun is likely if the association links between them, as determined by co-occurrence frequency, are loose, or non-existent. It should be noted however that certain strings which may well count as lexical chunks, as reported above, could not be detected by the application of co-occurrence frequency as a measure of associations between words. Another method of measuring association strength, namely, Mutual Information, will be employed below, in order to get a better grip at all types of lexical chunks.

4.4.4 Mutual Information

As reported in studies on lexical production and comprehension, "mutual information is a better measure of the cohesion for two-word pairs" (Gregory et al. 1999) than co-occurrence frequency (cf. Ellis et al. 2008b). Mutual information is a statistic measure drawn from information science aimed at assessing the extent to which the words in a pair appear together more frequently than would be expected by chance (Oakes 1998, Manning & Schütze 1999, Wiechmann 2008). According to Fano (1961), if two words, x and y, have probabilities $P(x)$ and $P(y)$, then their mutual information (MI), $I(x, y)$ is defined to be

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

An informal definition is this: "mutual information compares the probability of observing x and y *together* (the joint probability) with the probability of observing x and y *independently* (chance)" (Church & Hanks 1990: 23). A higher MI score stands for a stronger cohesion between the words, while a lower score means that their co-occurrence is rather due to chance. The formula was modified as suggested in Wiechmann (2008):

$$\mathrm{MI} = \log_2 \frac{P(x, y)}{P(x)P(y)} N$$

where N signifies the number of words in the corpus. This modification does not alter the relations in the formula, but raises the scores and thus allows a more straightforward comparison. In terms of the investigated data, $P(x)$ and

$P(y)$ stand for the frequency of the adjective and the frequency of the noun, respectively, and $P(x, y)$ represents the frequency with which the adjective and the noun co-occur.[21]

Table 4.9 contains inserted German noun phrases distinguished by the highest MI scores in the data set. The words in each of the pairs from the table are strongly associated with each other: each word pair expresses a very specific meaning. The meaning of one word combination, i.e., *fauler Sack* 'lazy git', is even idiomatic. However, as evident in the table, the other word pairs exhibit compositional but still very specific meanings. Consequently, a close semantic relation between the words in a word pair appears to support the sequential link between them. In other words, the examined noun and adjective lexemes cohere on semantic grounds into multiword units. Observing the word pairs in the table, we can assert that in bilingual speech, both words in each pair are indeed realised in the same language, i.e., German.

Table 4.9: Inserted German noun phrases with highest MI scores in data set.

Noun phrases	MI
Weinblätter gefüllt 'filled vine-leaves'	15.32
standesamtliche Hochzeit 'civil wedding ceremony'	11.22
gebratene Nudeln 'fried noodles'	11.01
gefüllte Paprika 'filled pepper'	10.50
alleinerziehende Mutter 'single mother'	10.42
ausgebildeter Polizeihund 'trained police dog'	10.32
katholische Kirche 'Catholic Church'	10.30
fauler Sack 'lazy git'	10.23
gehackte Tomaten 'chopped tomatoes'	10.16
bares Geld 'cash money'	10.14

Table 4.10 lists the noun phrases which exhibit the lowest MI scores in the data set. Most of the given noun phrases are mixed constituents, though three of them are German insertions. We may thus conclude that mutual information appears to account for switch placement in a large part of the data, but not in every instance. While highly coherent phrases with high MI scores repel internal

[21]With the intention of not omitting unattested adjective-noun combinations from the data set, the frequency of a combination which is not attested in deWaC was set as 0.1, rather than 0. This allowed a differentiation between unattested and very rare combinations in deWaC.

switches, phrases with low MI scores in German apparently tend to attract them. Since this effect is gradual and best conceived as a tendency, it will be tested statistically in the remainder of this chapter.

Table 4.10: Inserted and mixed noun phrases with lowest MI scores in data set.

Noun phrases	MI
erster Freund 'first (boy)friend'	0.6626524
novyj *Trockner* 'new drier'	−0.7394403
klassnoe *Sprache* 'cool language'	−0.7835933
gute Geschäftsführung 'good management'	−1.0246024
obščij *Hochdeutsch* 'general Standard German'	−1.2734782
russkie *Begriffe* 'Russian terms'	−1.3562067
sportivnye *Sachen* 'sporty things'	−4.7012769
freie Bundesländer 'free federal states'	−6.8530364
ogromnyj *Titel* 'huge header'	−7.0364000
russkij *Besitzer* 'Russian owner'	−7.6235176

As regards the lexical items involved in the noun phrases with very low mutual information, we can observe that one of the words in a pair is often a high-frequency word. It may either be the adjective, such as *novyj* 'new' in in the combination *novyj Trockner* 'new drier' and *gut* 'good' in *gute Geschäftsführung* 'good management', or the noun, such as *Begriffe* 'terms' in *russische Begriffe* 'Russian terms' and *Sachen* 'things' in *sportivnye Sachen* 'sporty things'. Since the frequency of these lexemes is high, a chance that they combine with other words is also high. As a result, these strings exhibit low degrees of cohesion and therefore low MI scores. Of particular relevance in the discussion of mutual information is Oakes' (1998) cautionary assertion that "[t]his measure gives too much weight to rare events" (p. 171). The MI scores of some rare and unusual word co-occurrences can indeed be higher than the scores of certain frequent and semantically more coherent word combinations. Hence, the adjective-noun combinations *süße Grußkarte* 'sweet greeting card' and *schöne Saunalandschaft* 'beautiful sauna facilities' (whose adjective is realised in Russian in the bilingual corpus) have the scores of 7.142 and 4.788, whereas such word combinations as *gute Ablenkung* 'good diversion' and *(freiwilliges) soziales Jahr* '(voluntary) gap year for social work' have considerably lower MI scores, i.e., 3.239 and 0.969, respectively. For this reason, it is crucial to investigate the correlation between

mutual information and the speakers' preferences in switch placement on a large scale.

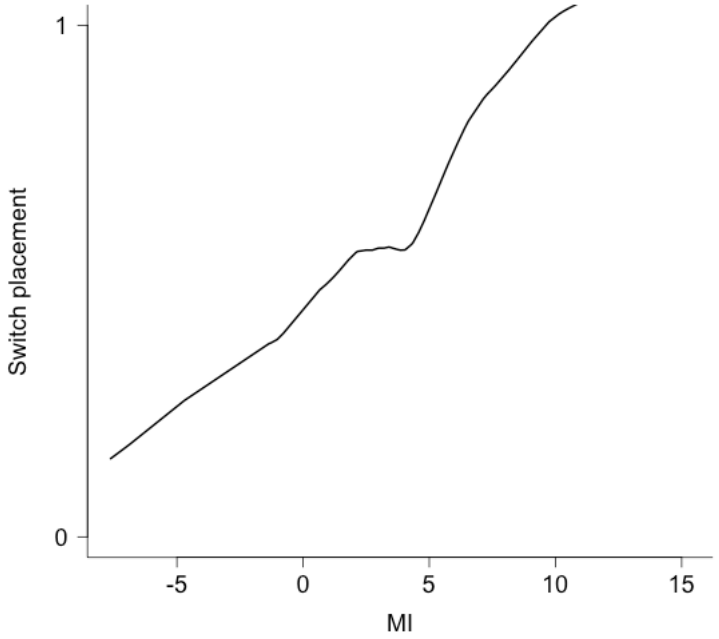

Figure 4.5: The relationship between switch placement and the mutual information (MI) of adjective-noun combinations in deWaC. The values of 0 and 1 on the *y*-axis stand for switching within and outside the noun phrase, respectively.

Figure 4.5 plots the correlation between MI and switch placement. In the graph, MI scores of the investigated co-occurrences are on the horizontal axis and switch placement is on the vertical axis. The line representing the relation between the two variables is a Lowess curve (cf. Figure 4.1). As is visible in the graph, the line displays a steep and steady upward trend for placing the switch at the phrase boundary, except for the interval between the scores of 2.14 and 4.0, where this trend levels off. In this interval, the switch may be placed either within or outside the noun phrase with a fifty-fifty chance. On the whole, however, the graph reveals a strong tendency for switching at the phrase boundary, with mutual information increasing, which appears to support the aforementioned research hypothesis. Yet, the question whether the factor MI is particularly pertinent to explaining the overall variance in the data is open to statistical investigation.

Thus, the remainder of this chapter is occupied with the statistical modelling of the investigated factors and their interplay.

4.5 Statistical prediction of switch placement

As the examined factors compete and interact with each other online in determining switch placement in noun phrases with adjective modifiers, it is crucial to account for their individual contributions as well as their interplays by using logistic regression analysis, which calculates the effect of individual predictors on a binary dependent variable under multivariate control. In terms of the investigated patterns of switch placement, this statistical method allows to determine the probability with which a switch is placed within the modified noun phrase or at its boundary given the conditioning factors discussed in the previous section as well as the importance of each of these factors. The generalised linear mixed-effects model, which I use in the present and the subsequent chapters, does this through the investigation of both fixed and random effects (see Baayen 2008: 278–84; cf. Bresnan et al. 2007).

A fixed-effect factor exhausts all of its possible levels, each of which can in principle be repeated. The fixed-effect factors in my data are measured numeric factors, such as frequency and mutual information. Random-effect factors have the property that they usually represent a random selection of the levels available in the population from a potentially infinite number of levels, or instances (Baayen 2008: 241). The individual levels of these factors are not interesting per se in a statistical analysis. The random-effect factor in my analysis is Individual. Since individual speakers contribute varying numbers of observations to the data set, the speaker may become a very influential factor (Tagliamonte & Baayen 2012), capable of distorting the effect of the fixed-effect predictors, and has therefore to be treated as a random-effect factor. According to Tagliamonte & Baayen (2012: 158), "[a]n important advantage of the mixed-effects modeling framework is that it allows the researcher to sample as many tokens from a given individual as is feasible, thereby increasing statistical power".

The statistics package R version 2.12.0 (R Core Team 2010) was used to carry out logistic regression analysis and all other statistical tests, and to generate graphical plots.

In a regression model with mixed effects, the joint contribution of all factors is computed by testing each factor individually, while the other factors remain constant. There are various search procedures for determining the model that provides the best fit to the data (cf. Baayen 2013). In selecting one of the customary heuristics, I adopt forward stepwise model selection. The general procedure

of this method consists in successive adding of potentially relevant predictors to the model specification. Applied to the factors outlined in the previous section, frequency of the adjective was the first fixed-effect factor included in the regression model, then the other factors, which include various frequency measures, were added stepwise. The models were compared in terms of Akaike's information criterion (AIC), which estimates a model's accuracy and complexity. Decreases in values of AIC indicate a better fit of the model. The minimal adequate model is rather accurate: as detailed in Table 4.11, it correctly classifies 82 per cent of all instances of switch placement in the data set, while the baseline model, which always predicts the most frequent realisation, i.e., switch placement at the phrase boundary, is only accurate in 65 per cent of cases. This shows the model's considerable increase in accuracy over the baseline model. The model delivers 86.7 per cent of correct predictions of switches placed at the phrase boundary; the prediction of switch placement within the noun phrase is more difficult since the model predicts this variant correctly in 74.2 per cent of cases. The model is reported in Table 4.12. The *C* index, estimating the probability of concordance between predicted and observed choices, is 0.903, which signals a high predictive power of the model. The performance indicator Somer's Dxy amounts to the value of 0.805, indicating a good fit. The model predictors exhibit a negligible degree of collinearity: the condition number κ, used for assessing collinearity, is 12.1 and thus below the threshold indicating medium collinearity (cf. Baayen 2008: 182). This is especially encouraging in view of the fact that the model's both predictors are frequency measures, which generally tend to be highly collinear (Baayen 2013).

Table 4.11: Model accuracy. Classification table for the model (1: switch at the phrase boundary; cut value = 0.50). (The table representation is based on Bresnan et al. 2007.)

		Predicted		% correct
		0	1	
Observed	0	23	8	74.2
	1	8	52	86.7
			Overall	82.0

Let us now inspect the predictors in the reported model. The predicted odds are for switch placement outside the modified noun phrase: positive coefficients indicate that a factor favours placing switches at the constituent's boundary, neg-

Table 4.12: Predicting switch placement in the context of the noun phrase modified by an adjective: minimal adequate generalised liner mixed model. Predicted odds are for switch placement outside the noun phrase. Significance codes: *significant at $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Factor	Est.	SE	z	Pr(> |z|)	
(Intercept)	10.237	2.795	3.662	< 0.001	***
Frequency of A	−0.919	0.233	−3.954	< 0.001	***
Co-occurrence fr.	0.374	0.109	3.422	0.001	***
Random effect:					
Speaker					
(intercept, $N = 17$, variance $= 0.999$, $\sigma = 0.999$)					
Summary statistics:					
N	91				
% correct predictions (% baseline)	82 (65)				
C index of concordance	82 (65)				
Somer's Dxy	0.805				

ative coefficients indicate that a factor attracts switch placement within the modified noun phrase, i.e., between the adjective modifier and the head noun. Hence, the frequency with which an adjective and a noun co-occur enhances the probability for a switch at the phrase boundary, whereas the frequency of the adjective reduces this probability. The coefficients are on the log scale and may be used to calculate the probability of switch placement outside the phrase in every case, as suggested in Ehret et al. (2014). The line 'Intercept' provides the log odds in the default case, which occurs when the frequency of an adjective-noun co-occurrence is average and the adjective has a low average frequency. An example from the bilingual corpus that comes closest to the default case is *türkise Farbe* 'turquoise colour', both its frequency and the frequency of its adjective are quite low in deWaC. For the default case, the model yields log odds of 10.24. We can interpret this value as odds if we reverse the natural logarithm (cf. Ehret et al. 2014). Hence, the odds are 27887.63 and the associated probability of placing the switch outside the noun phrase is 0.999. The odds for the estimated coefficients of the predictor variables are calculated in the same fashion. Thus, for co-occurrence frequency the odds ratio is 1.455, which means that while holding the other predictor constant, the likelihood for inserting a German noun accompanied by a German adjective in an otherwise Russian sentence grows by approximately 50

per cent at every one-unit increase of co-occurrence frequency on the logarithmic scale. This corresponds an increase by around 7 correctly predicted tokens on the linear scale. In other words, if the frequency with which a noun and an adjective appear together in German grows by 7 tokens, the odds for inserting an adjective-noun combination enhance by 50 per cent.

Frequency of the adjective has a lesser effect size than co-occurrence frequency. Although being a less strong predictor, frequency of the adjective is nevertheless the most important predictor of the observed variation. A model including the factor frequency of the adjective as the only fixed-effect factor has the lowest AIC in comparison to similar models testing one of the four investigated factors – MI, frequency of the adjective, frequency of the noun and co-occurrence frequency – as a single predictor. For the factor frequency of the adjective, odds were determined to be 0.39 on the linear scale. This means that increasing the frequency of the adjective by one unit lowers the expected percentage of switches placed outside the noun phrase by around 40 per cent. If the adjective is a high frequency word and the frequency of co-occurrence is particularly low, the chance for switching the language within the noun phrase is high, as in the case of *ogromnyj Trockner* 'huge drier' and *russkij Besitzer* 'Russian owner'. Conversely, if an adjective is infrequent and the frequency with which an adjective and a noun co-occur is high, the trend is towards inserting this adjective-noun co-occurrence into an otherwise Russian sentence, as in the case of *bares Geld* 'cash money' and *alleinerziehende Mutter* 'single mother'.

With regard to the random-effect predictor speaker, we can assert judging by its variance and the corresponding standard deviation that differences among the speakers in the corpus in terms of their preference for one of the two switching patterns are negligible. The highest adjustment to the model's intercept (1.34) is observed with the speaker who contributes most instances of switching in the context of the modified noun phrase to the data set. She exhibits a marked preference for placing switches at the phrase boundary.

In sum, the statistical regression analysis enabled discrimination between the examined usage-based factors in terms of their contribution to explaining the variation in switch placement in the data. Of the four analysed predictors – frequency of the adjective, frequency of the noun and frequency of the adjective-noun co-occurrence and mutual information – frequency of the adjective and co-occurrence frequency were found to account for switch placement in the context of the adjective-modified noun phrase. Although mutual information appeared to have a high predictive power, when included as the only fixed-effect predictor in the model's term, frequency of the adjective performed better in the respective model under the same conditions. Adding MI to the final, minimal adequate

regression model, which was based on co-occurrence frequency and frequency of the adjective, did not result in an improvement of the model quality. The variation under scrutiny can thus be captured by usage-based factors such as the frequency of the adjective, as measured in a German corpus, the frequency with which the noun and the adjective appear together in the same corpus, and the factor speaker, handled as a random effect predictor.

4.6 Conclusions and discussion

Already the early systematic studies of bilingual speech, such as Pfaff (1979a) and Poplack (1980b) report insertion of nominal constituents from one language in contact into the other and emergence of their code-mixed counterparts. As outlined above, the nature and status of inserted nominal constituents are a matter of controversy. Although some previous work (Backus 1996, Boumans 1998) has acknowledged the importance of recurrent combinations of words for the structure of code-mixing, no systematic evidence has been gathered to date to substantiate or refute this claim. Thus, the aim of this chapter has been to examine these two kinds of nominal constituents in detail to show that the choice between them depends on lexical factors, such as collocational ties between words and word frequency, rather than structural non-equivalence, as previously assumed (Myers-Scotton & Jake 1995, Myers-Scotton 2002).

The analysis of Russian-German code-mixing showed that German adjective-noun combinations are frequently inserted in Russian sentences. A vast majority of them represent well-formed German nominal constituents, of which only a fraction are fully-fledged German noun phrases. Crucially, a great number of adjective-noun combinations are inserted in Russian sentences without German determiners, but their internal structure is yet analysable since gender-number agreement is largely maintained. The tendency to omit determiners while preserving the internal structure of the nominal constituents is interpreted as a strategy to enhance structural similarity between the languages involved in code-mixing (cf. Hakimov & Backus n.d.(a); Sebba 2009). However, not all of the inserted adjective-noun combinations represent well-formed German nominal constituents. Since words in these constituents may be subsequent insertions, rather than parts of a multimorphemic insertion, only well-formed German constituents from bilingual sentences were subject to subsequent analysis.

Another bilingual strategy to express an adjective-modified noun phrase is to produce a mixed constituent. In my bilingual corpus, German noun insertions regularly combine with Russian attributive adjectives, but modification of Russian nouns by German attributive adjectives was not attested. Adjective-modified

nominal constituents in the analysed bilingual speech have thus German nouns as heads, and both German and Russian attributive adjectives as modifiers. In other words, a bilingual speaker may switch from Russian to German either inside the adjective-modified nominal constituent, or at the constituent boundary, producing an embedded-language island.

The choice between these two patterns of switch placement was hypothesised to depend on the following factors: frequency of the adjective, frequency of the noun, frequency of their co-occurrence and mutual information, a statistical measure of association between them. The speakers' possible predilections towards one of the patterns was controlled for using the random factor "individual" in the analysis. Corpus linguistic methodology and statistical modelling were used to test the contribution of each of the factors to explaining the variation in the data set.

The findings of the study include two frequency effects. Occurrence frequency of adjectives appears to affect switch placement in the adjective-modified noun phrase in my data. Frequent, more accessible adjectives tend to come from Russian, which is the language of the sentence frame and is thus the more activated language in this case. Attributive adjectives of average or low frequency, which exhibit very specific meanings, are predominantly German. This observation provides evidence for Backus's (1999a,2003) specificity continuum hypothesis introduced in §4.4.1, according to which in bilingual speech, items from another language usually have specific meanings, particularly at early stages of contact, as is the case in the current study. As regards online production, we may conclude that whenever a very specific German adjective of average or low frequency is activated and accessed, it co-activates, or triggers, the head noun in the same language. This may be to the fact that in German, an inflected adjective sets a strong projection with regard to the nature of the following element, which must be a noun (Auer 2007a: 98; cf. Auer 2005). It is important to emphasize that co-activation spreads from the adjective to the noun and not vice versa because not a single German adjective appears to modify a Russian noun in these data, i.e., each inflected German adjective is followed by a German noun, but German nouns freely combine with Russian adjectives. These results shed light on the degrees of activation of the German lexicon/grammar during the processing of the examined constructions. As the adjectives modifying German nouns in otherwise Russian sentences are mainly German and their accessibility in processing, owing to their low and average frequencies of use, is restricted, we may assume that the German lexicon/grammar is highly activated when German nominal constituents, or the so-called Embedded-Language islands, are produced. In other words, the German lexicon/grammar should be more strongly activated

because it delivers very specific adjectives and is also responsible for the co-activation effect. When the role of German is restricted to supplying only nouns involved in mixed nominal constituents, its activation seems to be lower. Hence, the effect of the lexical frequency on the choice between a German monolingual constituent and a mixed Russian-German constituent offers indirect support for Myers-Scotton's (2002: 140) view that the production of embedded-language islands and the production of single embedded-language elements occurring in the matrix language frame require different levels of activation of the embedded language.

That frequency with which nouns and adjectives are used together in German influences switch placement in the examined syntactic context is the other principal finding of this study. The regression analysis documents that lexical words that regularly co-occur repel switches, while word combinations exhibiting loose collocational ties in German and therefore weak associations appear to attract switches. The combination of these two factors accounts for the patterns in switch placement in the data set: If the frequency of the adjective, the most important factor in the statistical model, is low, the chance is high that a German constituent will be produced, regardless of the phrase frequency. If frequency values of both factors are high, a German nominal constituent will be very likely, but a mixed constituent may not be ruled out altogether. In contrast, if only the frequency of the adjective is high but the phrase frequency is low, a mixed constituent is a serious possibility.

Further findings pertain to the tested frequency-based predictors of switch placement in the examined syntactic context. The study confirms Heylen & De Hertog's (2014) claim that frequency of co-occurrence is able to capture multiword units if the syntactic context in which they are analysed is well defined, as in the present study (see also Schneider 2014). That co-occurrence frequency may be a powerful factor influencing online processing is supported by studies of phrase frequency using behavioural data (Arnon & Snider 2010, Janssen & Barber 2012) as well as neurophysiological data (Tremblay & Baayen 2010). Mutual information, which has been observed to overestimate the relevance of rare events, seems to be counterbalanced in my data by frequency of the adjective, the most relevant of the examined frequency-based factors to accounting for switch placement in the adjective-modified noun phrase.

As a syntactic context of variation in switch placement, the adjective-modified noun phrase implies that the involved lower-level constituents are lexical words. Because speakers choose among a virtually unlimited number of lexemes, the effect of co-occurrence frequency on their choices may be restricted. My corpus

indeed includes German adjective-noun combinations in otherwise Russian sentences which could not be attested in the large deWaC corpus. It seems highly plausible that when the choice among possible candidates for selection is more constrained, the effect of co-occurrence frequency will be more robust. In order to test this claim, the next chapter will analyse variation in switch placement in the prepositional phrase, a syntactic context in which the bilingual speaker selects a function word from a rather limited set. Additionally, other usage-based factors such as recency will be investigated as predictors of the examined variation.

5 Code-mixing in the prepositional phrase

This chapter[1] examines patterns of insertional code-mixing in the prepositional phrase. This syntactic context is one the most frequently reported loci for code-mixing: a speaker can switch the language either at the boundary of the prepositional phrase (cf. Bentahila & Davies 1983: 314; Boumans 1998: 271, 315; Clyne 1987: 757; Haust 1995: 169; Pfaff 1979a: 310; Treffers-Daller 1994: 208, 221–224) or within the prepositional phrase, i.e., between the preposition and the noun phrase (Bentahila & Davies 1983: 315; Poplack 1980b: 602; Pfaff 1979a: 310; Stenson 1990: 173, 178). If we restrict our attention to insertional code-mixing, we can subsume these phenomena under the concept of insertion. Either a prepositional phrase or a noun (phrase) is inserted from one contact language into a clause framed by the other contact language. In other words, prepositional phrases involved in code-mixing are either inserted or mixed. These mixing patterns are also regularly found in Russian-German bilingual speech in Germany as documented in my corpus. Despite frequent mention of these patterns in the literature, systematic investigations of variation in this syntactic context are still missing from research, let alone studies uncovering the factors driving this variation. The purpose of this chapter is therefore to describe and account for variation in code-mixing in the prepositional phrase as observed in my Russian-German data. In doing so, I will take a usage-based perspective on the examined patterns and view this variation as an outcome of usage-based factors, which include frequency of (co-)occurrence and repetition of words in discourse.

The structure of the present chapter will be largely identical to the structure of Chapter 4. First, I will briefly outline two accounts of insertion: Carol Myers-Scotton's Matrix Language Framework and Ad Backus' unit hypothesis (they are given a detailed presentation in Chapter 1). The next section will describe inserted and mixed prepositional phrases in my corpus of Russian-German bilingual speech. Section three will be concerned with factors that influence switch placement in the prepositional phrase. In the subsequent section, I will report

[1]This chapter is based on an earlier publication, see Hakimov (2016a).

the results of the statistical analysis testing these factors and their contribution to the scrutinised variation. Finally, I will summarise the results of the study and put them in the broader perspective of usage-based linguistics.

5.1 Previous accounts of mixing in the prepositional phrase

Although code-mixing in the context of the prepositional phrase is widely reported in bilingual speech involving various language pairs, few scholars have attempted to elucidate the factors underlying the choice between placing a switch within the prepositional phrase and switching the language at the phrase boundary. Such variation has been of particular interest to the Matrix Language Framework (MLF) model (Myers-Scotton 1993, 2002). As has been laid out in the previous chapters, the principal tenet of this model is that one of the languages in contact provides the core clause structure, whereas the role of the other language is restricted to the supply of lexical items. The structure of a bilingual clause is hence analysed by determining the division of labour between the involved languages. The language responsible for the core clause structure is labelled as the matrix language (ML), the other, dominated language is the embedded language (EL) (Myers-Scotton 1993: 75–119). As a result of this asymmetry, embedded-language content morphemes such as nouns are inserted into constituents organised by the matrix language to form mixed constituents. Other lines of research refer to this situation as insertional code-mixing (cf. Chapter 1). For example, in (1), the prepositional phrase consists of the Croatian preposition *u* 'to' and the mixed constituent *city-ju* 'city-ACC.SG.F. The mixed constituent emerges because the English noun *city* receives the Croatian inflectional suffix *-ju* of the accusative case, which is assigned by the preposition *u* 'to' to its complement.

(1) Croatian-English (Hlavac 2003: 203)
i iš-l-i smo u *city*-ju
and go-PTCP-PL be.PRS.1PL to -ACC.SG.F
'…and we went to the city…'

Embedded-language noun stems sometimes lack matrix language case markers, as in (2), where the noun *city*, again preceded by the Croatian preposition *u* 'in', does not take the required inflectional suffix of the locative case and remains bare.

(2) Croatian-English (Hlavac 2003: 202)
ja bolje vol-im bi-ti u *city*
I more like-PRS.1SG be-INF in
'…I like more to be in the city…'

In bilingual speech, prepositional phrases consisting of a matrix language preposition and a mixed constituent, as in (1), and those consisting of a matrix language preposition and a bare embedded-language noun, as in (2), alternate with prepositional phrases containing only embedded-language morphemes. Such constituents are covered by the umbrella term "embedded-language islands". Under this model, the English prepositional phrase *down the stairs* in (3) is analysed as an embedded-language island.

(3) Croatian-English (Hlavac 2003: 227)
…i jedan dan ja sam pa-o *down the stairs* i jedan
and one day I be.PRS.1SG fall-PTCP.SG.M and one
je bi-o…
be.PRS.3SG be-PTCP.SG.M
'…and one day I fell down the stairs and one was…'

Myers-Scotton & Jake (1995) explain the appearance of embedded-language islands in code-mixed utterances by mismatches in the grammatical information contained in lemmas, which are defined as entries in the mental lexicon underlying lexical items. Following their view (ibid., 1008–1014), lemmas represent grammatical information at three levels: (*i*) lexical-conceptual structure (semantic/pragmatic features), (*ii*) predicate-argument structure, and (*iii*) morphological realisation patterns. The lack of congruence at one of these levels between the relevant embedded-language lemma and its matrix language equivalent can result in the emergence of an embedded-language island. According to Hlavac (2003: 227), the embedded-language island *down the stairs* in (3) is produced because the content morpheme *down* "for most of its uses in English […] is non-congruent to any Croatian content morpheme equivalent". That is, the embedded-language island emerged because of lacking congruence in the lemmas involved at the level of lexical-conceptual structure. At the same time, Deuchar (2005: 258) contends that semantic/pragmatic differences between lemmas, as in (3), can motivate the appearance of specific switches, but only grammatical congruence determines the possibilities for code-mixing. While in Chapter 4, devoted to the analysis of mixing in the adjective-modified noun phrase, I considered the constituent's internal structure and its word-order pattern as crucial features pertaining to

congruence in the realisation patterns, in the case of prepositional phrases, it is congruence at the level of adposition that is relevant in the first place. Different realisation patterns of adposition indeed seem to regulate the occurrence of embedded-language islands structured as adposition phrases in certain language pairs. For instance:

(4) German-Hungarian (Szabó 2010: 435)
ich ich war dort (.) ich war dort äh *internátus-ba*
I I was there I was there HES boarding.school.SG-ILL
'I was there, I was there in the boarding school.'

(5) Finnish-English (Lehtinen 1966: 226)
ja sitte eh *in the afternoon* isä vai minä meni…
and then HES father or I went
'And then in the afternoon father or I went…'

The Hungarian illative suffix *-ba* in (4) expresses the kind of spatial relations that are coded by the preposition *in* in German. Hence, the Hungarian phrase *internátusba* 'in the boarding school' corresponds to the German prepositional phrase *im Internat* (where *im* is a contracted form, merging the preposition *in* and the determiner *dem*). The phrase *in the afternoon* in (5) is equivalent to the Finnish *iltapäivällä*, which consists of the noun *iltapäivä* 'afternoon' and the adessive suffix *-llä*. We can conclude from these examples that, if one of the languages involved in code-mixing employs prepositions and the other relies on postpositions, a possible outcome of code-mixing is the emergence of embedded-language islands. However, explaining this phenomenon by the incongruence in adposition realisation pattern may fail when the contact languages share the same pattern. Such languages include, for example, Russian and German, which both use prepositions to express spatiotemporal relations. The emergence of embedded-language islands in (4) and (5) may also be attributed to the level of lexical-conceptual structure. Both insertions are conspicuously marked by hesitations, which indicate that the speakers are processing word searches.

An alternative explanation for the occurrence of strings *internátusba* 'in the boarding school' and *in the afternoon* in the bilingual utterances in (4) and (5) rests on the assumption that these multimorphemic forms are holistic units, or chunks, in the corresponding speakers' mental lexicons/grammars. According to the unit hypothesis (Backus 1999a, 2003), embedded-language islands are units inserted into matrix language clausal frames. Under this approach, any string of morphemes or words gains the status of unit once it is entrenched in the

speaker's lexicon. As described in Chapter 1, in order to qualify as units, multimorphemic strings have to either (*i*) demonstrate irregular morphosyntax, (*ii*) express non-compositional semantics, or (*iii*) be of recurrent use (Backus 2003: 90). For the multimorphemic strings in (4) and (5), which exhibit neither morphosyntactic irregularities nor semantic non-compositionality, it is only frequency of use that may account for their status as lexical units. Apparently, the morphemes *in*, *the* and *afternoon*, on the one hand, and *internátus* and *-ba*, on the other, co-occur so frequently in English and Hungarian, respectively, that speakers of these languages represent and retrieve them as units. This idea is in line with Bybee's (2002a) Linear Fusion Hypothesis (cf. Chapter 2). Unit status, determined by usage frequency, is obviously a gradient category since frequency is itself gradient. Although evidence for the unit hypothesis pervades most, if not all, bilingual corpora, reservations to this approach should yet be voiced. First, we cannot exclude the possibility altogether that grammatical incongruence plays no role in the emergence of embedded-language islands. Therefore, prepositional phrases need be analysed in a language pair which employs the same pattern of adposition realisation; Russian and German are good candidates for such a test. Secondly, to consider any embedded-language island a unit would be fatuous because not every multimorphemic embedded-language insertion is a unit (see §1.5, for a discussion of this issue). On this account, embedded-language islands must be submitted to systematic investigation involving both the application of statistical analysis and the presentation of available additional evidence, which could support the unit status of the examined embedded-language islands, independently from code-mixing data. In what follows I thus analyse the prepositional phrases involved in code-mixing in my Russian-German bilingual corpus and supplement this analysis by measuring the examined structures' frequencies in deWaC, the large monolingual German corpus (Baroni & Kilgariff 2006) which I already utilised as a source of distribution information in the study reported in Chapter 4. These data will enable me to tap into co-occurrence frequency as a one of the factors responsible for the scrutinised variation and to thus put the unit hypothesis to a test.

5.2 Prepositional phrases in the corpus of Russian-German bilingual speech

In my Russian-German bilingual corpus, the use of prepositional phrases in bilingual sentences follows one of two principal patterns: most commonly, German nouns occur in Russian prepositional phrases as complements, the other typical

pattern is the so-called long German insertion structured as a fully-fledged prepositional phrase. However, unlike the patterns of adjective-modified noun phrases, described in Chapter 4, the use of prepositional phrases in bilingual sentences, as will be shown below, is more varied.

5.2.1 Patterns of prepositional phrases in bilingual sentences

Prepositional phrases constitute a large part of the switches observed in my bilingual Russian-German corpus: a total of 456 pertinent instances were identified. The investigated syntactic context demonstrates a high degree of variability. I first focus on the most frequent case, namely, prepositional phrases in bilingual sentences with Russian clausal frames. This case falls into two sub-cases and can be represented schematically like this:

(6) $_{R}[\ P\ _{G}[\ N\ _{G}](\text{-INFL})\ _{R}]$
$_{R}[\ _{G}[\ P\ (\text{ART})\ N\ _{G}]\ _{R}]$

where *R* and *G* stand for Russian and German, respectively. While the former pattern involves the insertion of a German noun in a Russian prepositional phrase such that the noun occasionally combines with a Russian inflectional suffix, in the latter pattern a fully-fledged German constituent is embedded in a Russian clausal frame.

The considerable variability which the prepositional phrases exhibit in the corpus is due to (*i*) the code-mixing type, i.e., alternation or insertion (cf. Chapter 1), (*ii*) the choice of the matrix language in the case of insertional mixing, and (*iii*) phrase complexity. 25 prepositional phrases are involved in alternational code-mixing, as in (7).

(7) (LG0503)
mm a esli *zum* *warenkorb* *geh-sch*
INTRJ and if to.ART.DAT.SG.M shopping.basket go.PRS-2SG
'Mm, and if you go to the shopping basket?'

Here, the sequence comprising the Russian connector *a* 'and' and the complementiser *esli* 'if' is juxtaposed with a German string consisting of a prepositional phrase and a finite verb.

The predominant code-mixing type in the data is insertion: with 431 tokens, it constitutes 94.4 per cent of all the instances of prepositional phrases involved in mixing. For example:

(8) (LL-0517)
A: nam že zapreti-l-i kuri-t' v škol-e
DAT1PL PTCL prohibit-PST-PL smoke-INF in school-PREP.SG.F
B: *achso* daže *auf dem schulhof*?
PTCL even on ART.DAT.SG.M school.yard
A: v *schulhof*-e voobšče nel'zja. tol'ko jesli vyxod-iš'
in school.yard-PREP.SG generally forbidden only if go.out-PRS.2SG
am schulhof
at.ART.DAT.SG.M school.yard
A: 'They prohibited us from smoking in school.'
B: 'Ah, even in the school yard?'
A: 'In the school yard it is generally forbidden; only if you go outside the school yard.'

Lines three and four of example (8) illustrate two major patterns of insertional code-mixing in the examined context: either a German prepositional phrase such as *am Schulhof* 'in the school yard' is embedded in the Russian matrix frame, or a German noun, namely, *Schulhof* 'school yard', is used in a Russian prepositional phrase (cf. 6). These examples are in line with the observed tendency to use German nouns and prepositional phrases in Russian discourse. Insertion of Russian nouns and prepositional phrases in German sentences is possible but rare. By and large, insertions of German items in otherwise Russian clauses prevail over Russian insertions in German discourse. An example of a Russian prepositional phrase occurring in a German clause is (9).

(9) (LJ0526)
s kitajc-ami *würde ich leb-en*
with Chinese-INS.PL AUX.COND.1SG NOM1SG live-INF
'I would live with the Chinese.'

In this example, German sets the matrix frame, and accommodates the Russian prepositional phrase *s kitajcami* 'with the Chinese'.

Prepositional phrases adjacent to other insertions constitute a special case, for instance:

(10) (LR0712)
ili ty v *baden-württemberg studier*-u-eš i potom
either NOM2SG in Baden-Württemberg study-ST.PRS-2SG and then

tebe nado v *irgendein ander-es bundesland…*
DAT2SG necessary to some[ACC.SG] other-ACC.SG federal.state[ACC.SG]
'Either you study in Baden-Württemberg and then you have to move to some other federal state….'

The mixed phrase *v baden-württemberg* 'in Baden-Württemberg' in (10) is followed by the morphologically integrated verbal insertion *studierueš* '(you) study'. 18 tokens of this type (approx. 4.0%) could be identified in the corpus. In order to keep the data set homogeneous, these instances were discarded, such that only lone insertions, i.e., those surrounded by the matrix language morphemes, as in (8), entered the data set.

Phrase complexity is a further source of variability when Russian is the matrix language. As expected, simple prepositional phrases take precedence over prepositional phrases with expanded noun phrases. The asymmetry is obvious because the former pattern is twice as frequent in the bilingual corpus as the latter pattern, which is exemplified by the mixed prepositional phrase *v irgendein anderes Bundesland* 'to some other federal state' in (10).

The various patterns reflecting the syntactic variability in the examined context and their counts in the bilingual corpus are summarised in Figure 5.1. The distribution of the prepositional phrases involved in insertional code-mixing reveals that, with Russian being the matrix language, simple prepositional phrases overwhelmingly dominate all other patterns described above ($\chi^2 = 170$, $p < 0.001$). Therefore, the present chapter focuses on switch-placement within this type of structure. The data set contains 86 tokens of German simple prepositional phrases embedded in Russian sentences – $_{R}$[$_{G}$[P (ART) N $_{G}$] $_{R}$] – and 247 German noun complements of Russian prepositions – $_{R}$[P $_{G}$[N $_{G}$](-INFL) $_{R}$].

Some of the involved nouns were subject to further analysis since they were either involved in word-internal mixing or in mixing at a site of non-equivalence, or qualified as established loans. Seven German noun stems occurring in Russian prepositional phrases take not only the corresponding Russian inflectional suffixes but also the Russian diminutive suffix *-ik*, as in (*ot*) *dart-ik-a* '(from) dart-DIM-GEN.SG'[2]. Because such mixed forms consist of German noun stems and Russian derivational affixes, they were regarded as instances of word-internal mixing and were omitted from the data set.

[2]The word *dart-ik*, consisting of the German stem *dart-* and the Russian diminutive suffix *-ik*, presents a blend of the German and Russian terms for 'dart', namely, *Dart* and *drotik*. This blend may have resulted from the speaker's reanalysis of the Russian simplex *drotik* as a suffixed word and her confusion of the terms.

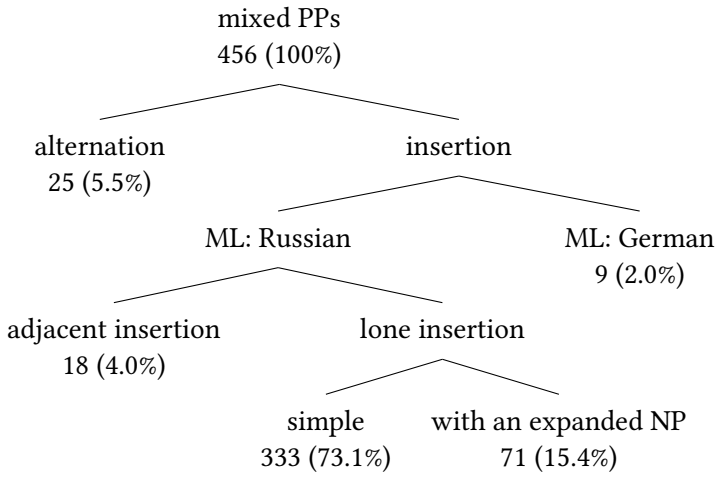

Figure 5.1: Prepositional phrases involved in code-mixing: variability of structures and proportions.

Prepositional phrases resulting from mixing at a site of non-equivalence constitute another equivocal case. Both instances of this case – *na Hälfte* '(up to a) half' and *na Endeffekt* 'finally' – were produced by the same speaker (Rita). The Russian counterpart of the former string is the adverb *napolovinu*, its German equivalent is the prepositional phrase *zur Hälfte*. The Russian correspondents of the latter string (cf. German *im Endeffekt*) include expressions *v (konečnom) itoge* and *v konce koncov*. The unconventional use of the preposition *na* in this context may be attributed to an interference with the adverb *nakonets* 'at last', although its meaning differs slightly from the contextual meaning of the produced form *na Endeffekt*. The latter may be a result of an erroneous retrieval owing to processing difficulties. We may assume that the speaker was aiming at an expression with the stem 'end', either the German *Ende* or the Russian *konec*, but could not retrieve an appropriate preposition. We may also analyse this string as an instance of word-internal mixing if we regard the *na* as a prefix rather than a preposition, just as in *nakonec* (cf. the aforementioned *na Hälfte* versus *napolovinu*). Lack of clarity on whether it would be appropriate to subsume the two instances under the category of word-internal mixing led me to remove these items from the data set.

The last category of noun insertions omitted from the data set includes three nouns which have acquired new meanings in the variety of Russian spoken in Germany. The German noun *Sprache* 'language' is used in the corpus in the form

sprachi, carrying the Russian plural inflection *-i*, to refer to 'language courses'. While entering the vocabulary of Germany's Russian, the meaning of the noun *Heim* 'home' narrowed to the meaning 'home for late repatriates' (cf. German *Spätaussiedlerheim*). Finally, the German adjective *sozial* 'social' began to be used as a noun, its meaning underwent conventionalisation to refer to 'social benefits'; the noun occurs particularly often in the expression *sidet' na soziale* 'to be on supplementary benefit' (cf. Russian *sidet' na* [*social'nom*] *posobii*; the expression from the bilingual corpus was unattested in Russia's Russian). On the basis of the observed semantic changes, it is reasonable to assume that the nouns under scrutiny have become native items in Germany's Russian, i.e., established loans.

In total, 233 instances of German nouns occurring in Russian prepositional phrases entered the final data set.

5.2.2 Frequency distribution of the structures in the data set

Relying on semantic change as an indicator of an item's status as an established borrowing is usually restricted to few special cases. I will therefore utilise the frequency of a German word, or a longer item, in the Russian discourse as a diagnostic of its status as an established loan, as I did in the previous chapter (§4.3.3). As such, the identified German prepositional phrases and nouns appear in otherwise Russian sentences with varying frequencies, but their lion's share occurs in the Russian discourse only once. However, some German items in the corpus, both prepositional phrases and nouns, were used with a higher frequency.

The pertinent prepositional phrases include the following items: the string *zum Beispiel* 'for example' appears four times in the Russian sentences of the corpus; the phrases *am Montag* 'on Monday' and *zum Ausgleich* 'for compensation' are embedded in the Russian discourse three times each, and the strings *im Normalfall* 'normally', *am Dienstag* 'on Tuesday' as well as *am Sonntag* 'on Sunday' occur in the Russian discourse twice each. Overall, the data set contains 86 tokens of embedded German prepositional phrases, which correspond to 79 types.

The German nouns occurring in the examined Russian prepositional phrases, just as the German nouns modified by Russian adjectives examined in the previous chapter, appear in Russian sentences at a higher rate than longer constituents. Yet, only a minor portion of these nouns are embedded in the Russian discourse on a regular basis, with a majority of them occurring in it only once. Determining these nouns' frequencies in otherwise Russian sentences involved the following procedure: Every token of a specific German lexical item (type) was counted

when it appeared in the Russian context as a lone item. For example, the German lexeme *Montag* 'Monday' occurs in the corpus in the following contexts:

(11) (a, b, d: LJ1105; c: LN1107)

a. v *montag* bud-u rabota-t'
in Monday[ACC.SG] AUX.FUT-1SG work-INF
'On Monday I will work.'

b. *montag* bud-u rabota-t'
Monday[ACC.SG] AUX.FUT-1SG work-INF
'Monday I will work.'

c. vot ot *montag* do *freitag* u nas mnogo škol-y
PTCL from Monday till Friday at 1PL.GEN much school-SG.GEN
'From Monday till Friday we have a lot of school-classes.'

d. a my *chemie* že *am* *montag* pisa-l-i
but 1PL.NOM chemistry PTCL at.ART.DAT.SG.M Monday write-PST-PL
'But we wrote chemistry on Monday.'

As the lexeme *Montag* 'Monday' occurs three times in the Russian discourse, namely, in (11a, 11b[3] and 11c) but not in (11d), where it is part of a German prepositional phrase, its frequency in stretches of the Russian discourse amounts to three tokens.

The frequencies with which the examined nouns appear in the Russian discourse in the corpus are given in Table 5.1. The table reveals that hapax legomena, i.e., items occurring in the corpus once, amount to 44.6 per cent of all the analysed German noun lexemes. The German nouns which are inserted in Russian sentences ten or more times are in the minority; their contribution to the totality of the scrutinised nouns is only 3.1 per cent. The majority of German noun lexemes (52.3%) occur in Russian sentences as lone items at rates ranging between one and ten. Therefore, the German nouns examined here parallel the nouns investigated in §4.3.3 in that they also form a heterogeneous class with regard to the frequency of occurrence in the Russian discourse as lone items. In order to achieve some homogeneity among the sporadic and recurrent lexical items under scrutiny (cf. Poplack et al. 1988), I will employ a cut-off threshold scheme, along the lines of reasoning expounded in the previous chapter. German nouns that appear five or more times in the Russian discourse counted as recurrent and were discarded from the data set. The excluded items encompass eight tokens of the aforementioned lexemes, which correspond to 7.9% of the data set.

[3]The noun *Montag* is analysed here as an instance of bare adjunct noun phrase (see Larson 1985, for English, Tajsner 1997, for Polish).

Table 5.1: Frequencies of German noun insertions in Russian sentences as distributed in bilingual corpus.

Word frequency		Number of lexemes	
Absolute, F	Relative	Absolute	%
1	0.00004	70	44.6
2	0.00008	28	17.8
3	0.00012	20	12.7
4	0.00016	14	8.9
5	0.00020	9	5.7
6	0.00024	6	3.8
$6 < F < 10$	–	5	3.2
10	0.00040	1	0.6
$F < 10$	–	4	2.5
Total		157	100.0

In summary, some German noun insertions occur more frequently in Russian sentences than others. Control over the frequency of such nouns in the Russian discourse was achieved by removing frequent lexical items from the data set.

5.3 Factors

In this section, I investigate the question whether the choice between placing a switch within the prepositional phrase or at its boundary may be predicted by usage-based factors such as co-occurrence frequency, word frequency and word repetition. The frequency of the preposition is not considered as a predictor of switch placement because the prepositions used in the identified prepositional phrases are all high-frequency items. The factor "word repetition" is included in conjunction with the impact of repetition priming, or recency in discourse, on choices among functionally equivalent items competing for selection in on-line speech production. I address the research question by adopting the methodology introduced in Chapter (4). A novel step pertains to approaching the factor "co-occurrence frequency". Specifically, I introduce a measure modelling the competition among recurrent word strings structured as prepositional phrases.

5.3.1 Modelling frequency of co-occurrence

Switch placement in the context of a syntactic phrase is hypothesised to be the function of the frequency with which the words constituting the phrase appear together (cf. Bybee 2010: 33). If a preposition and a noun co-occur with a high frequency, they exhibit a strong bond, which is likely to repel a code-switch. Conversely, if the association between the preposition and the noun is loose, i.e., the frequency of co-occurrence between these items is low, a switch is likely to be placed between them. To test these hypotheses, I measured the frequencies with which the nouns extracted from the bilingual corpus co-occur with prepositions in German.

5.3.1.1 Corpus analysis

The strength of association between a given noun and a specific preposition was operationalised as co-occurrence frequency. Modelling associations between words by utilising co-occurrence frequency paralleled the corpus analyses reported in Chapter (4). The nouns used in the prepositional phrases in the data set were investigated in deWaC as to the preposition realisations with which they usually appear together in German. All the possible syntactic formats of the German prepositional phrase were considered: [P N], [P ART N] and [P.ART N]. The analysis procedure is illustrated by the German prepositional phrase *an einem Seil* 'on a rope', occurring in a mixed sentence in (12).

(12) (LL0510)
i čto on-i vs-e *an ein-em* *seil*
and that 3-PL.NOM all-PL.NOM on ART-DAT.SG.M rope
'And that they are all on a rope.'

Combinations of the noun *Seil* with specific prepositions, preposition-article contractions as well as preposition-article combinations were identified in deWaC, and their corresponding frequencies were measured. The results for the item *Seil* are given in Table 5.2. As the utilised corpus was not tagged for the morphological case, it was impossible to distinguish between strings with differing, or syncretic case exponents. For example, the tokens of the strings *an ein Seil* 'onto a rope' and *an einem Seil* 'on a rope' were collapsed automatically to the pattern *an ein Seil*, where *ein* stands for all the case forms of the indefinite article. The individual patterns were further merged to a less specific pattern with the unspecified article – [*an* ART *Seil*] – because the choice of articles largely depends on the contextual information.

Table 5.2: The use of the noun *Seil* 'rope' in the context of the prepositional phrase in deWaC. The abbreviation *d* refers to all the case forms of the definite article, the abbreviation *ein* refers to all the case forms of the indefinite article.

Pattern	Frequency	Pattern	Frequency
am	428	*ins*	51
an ein	340	*durch ein*	44
mit ein	250	*auf ein*	43
auf d	186	*über ein*	30
mit d	115	*ans*	28
an d	107	*in d*	28
mit	84	*aufs*	23
über	65	*um d*	17
vom	62	*zum*	17
ohne	59	*durch d*	15
im	54	*per*	15

By using this procedure, I reordered sets of the prepositions co-occurring with the investigated nouns and calculated their co-occurrence frequencies. The outcome for the example noun is presented in Table 5.3. In a set of co-occurrences of a specific noun with an array of prepositions, such as the one in Table 5.3, the highest frequency value corresponds to the strongest association, as in the combination [*an* (ART) *Seil*], while low-frequency co-occurrences stand for loose associations between the string parts, as in [*per* (ART) *Seil*].

Table 5.3: The prepositions accompanying the noun *Seil* in deWaC.

Preposition	Frequency	Preposition	Frequency
an	903	*durch*	59
mit	449	*ohne*	59
auf	252	*um*	17
in	133	*zu*	17
über	95	*per*	15
von	62		

5.3.1.2 Predicting a chunk

To predict an item in production means to determine the likelihood with which it is produced by the speaker. The information about the distribution of a pattern across its specific instantiations is used to model the relationships between these instantiations in terms of probabilities. If one of the slots of a pattern is kept constant, a particular realisation of the whole pattern can be predicted by utilising the information about the distribution of the specific items in the other slot. The predicted realisation is then an outcome of the competition among the various realisations of the open element. In the present case, prepositions compete with one another in order to become activated together with a specific noun. A co-occurrence distinguished by the highest frequency in a given set has a greater chance of being used in production than a low-frequency sequence. For instance, in the set of the prepositions used with the noun *Seil* 'rope' (see 5.3), the co-occurrence *an – Seil* is more likely to be selected than any other co-occurrence.

The competition in preposition sets is modelled by applying the ratio of odds, based on (Fahrmeir et al. 2007: 119–121). Odds are the ratio of the probability that an event will happen to the probability that an event will not happen. Mathematically, it is formulated as follows:

$$\text{odds}_1 = \frac{F_1}{\sum F_i - F_1}$$

where *F1* stands for the frequency of a specific pattern instantiation and $\sum F_i$ denotes the cumulative frequency of the pattern. The index expresses the relationship between an element in a frequency distribution and the remaining distribution elements. Let me demonstrate this by using the distribution of the pattern [P … *Seil*], whose total frequency amounts to 2061 tokens. The preposition slot can be specified by 11 prepositions (cf. 5.3), and the most frequent co-occurrence is [*an – Seil*] 'on – rope', with a frequency value of 903 tokens. The probability that this realisation is selected in production, as determined by the odds ratio, is 0.779. Odds were computed for all the German strings in the data set, i.e., the nouns preceded by German prepositions. When the preceding prepositions were Russian, odds were calculated for German equivalents of the corresponding Russian prepositions in order that semantic equivalence was maintained. For example, in the case of the mixed string *na miete* 'on rent' in (13), the German preposition *von* was selected because its meaning corresponds to the meaning of the preposition *na* 'on' in the context of the verb *žit'* 'to live' (cf. German *von der Miete leben* 'to live on the rent'); the odds were thus computed for the combination [*von – Miete*], competing with the actually realised string.

(13) (LG05036)
on tol'ko na *miete* i živ-ët
3SG.M only on rent PTCL live.PRS-3SG
'It is only the rent that he lives on.'

The calculated odds ratios were normalised by employing a logarithmic scale, in order to avoid skewing of the distribution (cf. Baayen 2008: 31). An analysis of outliers (cf. Gries 2009: 258) resulted in the omission of three data points with extremely low odds values: the German phrase *neben dem Haus* 'next to the house' and the mixed phrases *naprotiv straße* 'opposite the street' and *krome kuchen-ov* 'except cake-GEN.PL.M'. The excluded data points make up 1.2% of the sample. The relationship between odds values and switch placement is represented in Figure 5.2. The binary variable "switch placement" is on the vertical axis with the values of zero and one, which stand for a switch within a prepositional phrase and a switch at its boundary, respectively; the values of odds are on the horizontal axis. The line depicting the relationship between these variables is a Lowess curve, which represents a function describing the deterministic part of the variation in the data and is generated by locally weighted scatterplot smoothing (Cleveland & Devlin 1988). The Lowess curve shows that the line begin to curve upwards around the logarithmic value of −1.3, which corresponds to the odds value of 0.273. This means that the tendency to switch within a phrase curbs steadily and permanently: with odds reaching the value of 0.273, the phrase boundary gradually becomes a more preferred switch site. That is, the higher the odds for a preposition-noun combination, the higher the likelihood that this combination will be produced in one language, here the embedded language German.

5.3.2 Frequency of the noun

The facilitatory effect of word frequency in language production was already introduced in Chapter 4. Following (MacWhinney 1997: 115), I argued that the associations between frequent words and the specific lexicogrammatical patterns with which they are usually used together may be stronger than such associations involving infrequent words. Under this view frequent words can trigger their lexicogrammatical patterns, including chunks, more easily than rare words. Conversely, lexicogrammatical patterns involving frequent words are more accessible than those involving rare lexemes. While in the previous chapter I analysed and evaluated the frequencies of both the adjectives and the nouns involved in the examined adjective-modified noun phrases, I restrict the analysis of frequency here to nouns because the prepositions under examination are all high-frequency items.

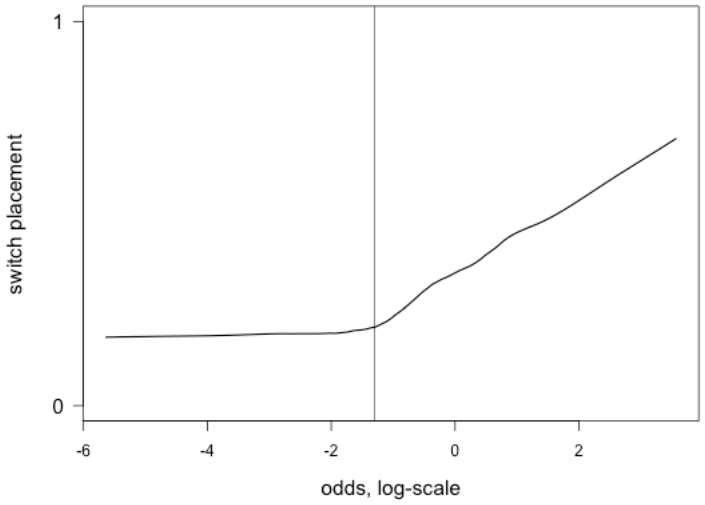

Figure 5.2: The relationship between switch placement and odds (on the logarithmic scale). The values of 0 and 1 on the y-axis stand for switching within and outside the prepositional phrase, respectively.

In the context of the present chapter, frequent German nouns appearing in prepositional phrases, unlike rare nouns, are assumed to have a more pronounced ability to trigger, or co-activate, prepositions typically occurring with them. To test this hypothesis, frequencies of the nouns from the examined prepositional phrases were obtained from the deWaC corpus. 5.4 provides some of the prepositional phrases under investigation and the corresponding noun frequencies. The first five prepositional phrases include low-frequency nouns and exhibit phrase-internal switches, whereas the last five prepositional phrases include high-frequency nouns and contain only embedded-language, i.e., German, elements.

The obtained frequency values were logarithmically transformed (cf. Baayen 2008: 31). The datum *Erotikshop* 'sex shop' was considered an outlier, due to its extremely low frequency in the corpus. The relationship between noun frequency and switch placement is plotted in Figure 5.3. The variable "switch placement" is on the vertical axis, and its values zero and one stand for a phrase-internal switch and a switch at the phrase boundary, respectively; noun frequency is on the horizontal axis. The Lowess curve demonstrates the effect of word frequency on switch placement: an increase in the noun frequency positively correlates with the tendency to switch the language at the phrase boundary. This is in line with the aforementioned hypothesis that the lexicogrammatical patterns involving high-frequency words are more accessible in production. Whether the

Table 5.4: Prepositional phrases switched inside and at the phrase boundary and the frequencies of the involved nouns, as measured in deWaC.

PPs with switches placed within and outside the PP	F_N
v *erotikshop* 'to the sexshop'	19
na *sporttag* 'on a sports day'	65
s *gummisohlen* 'with rubber soles'	80
za *einzimmerwohnung* 'for a studio appartment'	126
v *prüfungsstress* 'under stress from exams'	249
nach hause '(to) home'	230408
neben dem haus 'near the house'	230408
in der kirche 'in church'	281589
in die stadt 'to the city'	480767
am ende 'in the end'	579078

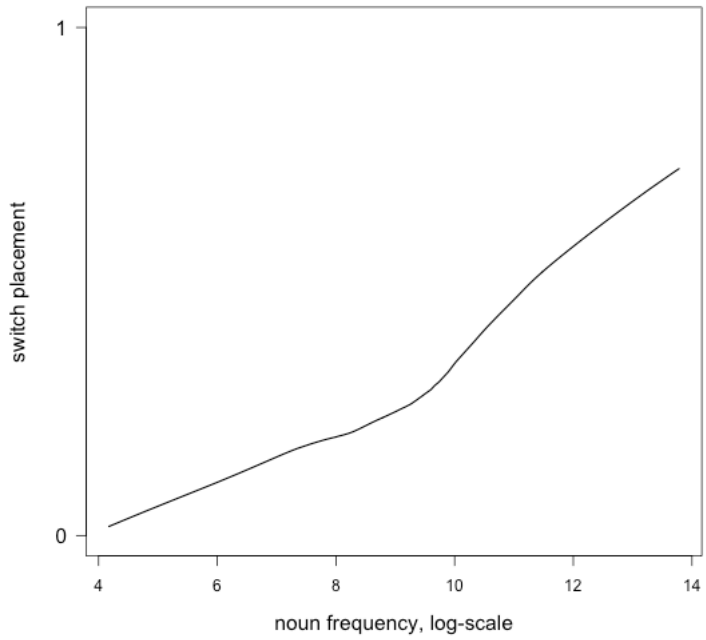

Figure 5.3: The relationship between switch placement in the context of the prepositional phrase and the frequency of the noun (on the logarithmic scale).

frequency of the noun significantly contributes to predicting switch placement is tested in a regression model below.

5.3.3 Word repetition

The implicit memory effect of priming has been shown to affect monolingual language production in various domains of language (see §2.3 for further details). In bilinguals, experimental studies have focused on cross-language priming effects in lexical access (Dijkstra et al. 1998, Kroll & Stewart 1994, van Hell & de Groot 1998) and the production of syntactic constructions (Loebell & Bock 2003, Salamoura & Williams 2006, Schoonbaert et al. 2007). Only recently has experimental research approached priming effects in bilingual speech. For example, (Kootstra et al. 2010) report that participants in their experiments tended to switch languages at the same position as in the prime sentence. This result, as Kootstra et al. report in their (2012) study, is driven by both the presence of cognates in the prime sentence and word repetition. In the conducted experiments, the subjects repeated a code-mixed sentence and then described a picture by using another code-mixed sentence. The authors show that lexical repetition between the prime sentence and the target picture, or the presence of a cognate in the prime and the target are capable of priming code-switches in sentences. Analyses of naturally occurring code-mixing have traditionally neglected priming effects, an exception to this trend is the work by (Travis & Torres Cacoullos 2016). The authors have found that in the New Mexico Spanish-English Bilingual corpus, the distribution of syntactic structures in a specific syntactic context, namely, the expression of the Spanish first-person singular subject pronoun, largely depends on both language-internal and cross-language priming effects. We can conclude from these studies that repetition of words and structures in discourse should be given due consideration in analyses of naturally occurring bilingual speech.

In the context of the prepositional phrase, an occurrence of a preposition in discourse can lead to a repeated selection of this preposition in subsequent discourse, provided that semantic compatibility is maintained. That is, once a preposition is produced, it is highly likely to be selected again in the same language, for instance:

(14) (LG05036)
priš-l-o-s' kogda v *lahr* zaexa-l-i
be.necessary-PST-SG.N-REFL when to Lahr[ACC.SG.M] come-PST-PL
perv-ym del-om v *krankenhaus* exa-t'
first-INSTR.SG.N thing-INSTR.SG.N to hospital[ACC.SG.M] drive-INF
'First thing when (they) came to Lahr, they had to drive to hospital.'

The mixed prepositional phrase in (14) consists of the Russian preposition *v* 'to' and the German noun *Krankenhaus* 'hospital'. The speaker may alternatively have selected the German preposition *in* 'to' to produce the phrase *ins Krankenhaus* 'to hospital' (in this phrase the preposition-article combination *in das* is usually realised as a contracted form). Nevertheless, the speaker selects the Russian preposition *v* 'in/to', which she produced as part of the string *v Lahr* 'to Lahr' in the previous clause. In other words, the first occurrence of the preposition *v* appears to prime the use of the same lexical item with the German noun *Krankenhaus* 'hospital' in later discourse.

Nouns involved in the examined prepositional phrases also seem impervious to the pervasive effect of repetition priming, for instance:

(15) (LVa0510)
tam *WELLE* na nix polete-l-a ili
there wave[NOM.SG] onto 3PL.ACC go.over-PST-F.SG or
poli-l-a-s' nu i esli posmotr-et' čo tam *auf*
spout-PST-F.SG-REFL PTCL and if look.at-INF what there on
der *WELLE* vidno
ART.DAT.SG.F wave visible
'A wave went over them or spouted, and if you look at what is visible on the wave…'

The noun from the inserted German prepositional phrase *auf der Welle* 'on the wave' in (15) appears in the prior discourse in the same language. Instead of repeating this lexical item in German in the subsequent prepositional phrase, the speaker may have produced the Russian equivalent *volna* as well. Yet, the opposite is the case: the speaker uses the German noun *Welle* persistently. The activation of this lexeme has presumably co-activated the German preposition *auf* 'on' associated with it. Furthermore, in the light of the analysis of example (14), according to which the choice of a specific preposition depends on its selection in the prior discourse, we could expect the Russian equivalent of this preposition, namely, *na* 'on', to occur rather than its German realisation, for the Russian preposition appears in the phrase *na nih* 'on(to) them' in the preceding clause. The result would have been a mixed phrase *na Welle* 'on the wave' (cf. Russian *na volne*). A brief examination of possible motivations, or factors, underlying the choices among alternatives available to bilinguals makes evident that we cannot tell with certainty which of the considered motivations plays a decisive role in determining the speaker's choice to use a concrete alternative, unless we examine these motivations systematically and subject them to statistical tests.

The data were coded for the presence, or absence, of the examined prepositions and nouns in the window of eight seconds in the prior discourse (i.e., 5 ± 3 seconds, cf. Szmrecsanyi 2006: 189). The relationship between switch placement and word repetition in the same language is given in Figure 5.4. The upper panel concerns the repetition of prepositions and the lower panel refers to the repetition of nouns. As we can see in the upper panel, the proportion of prepositions repeated in mixed phrases is skewed; the asymmetry indicates the tendency to switching language within the prepositional phrase, given a prior occurrence of a specific preposition in the same language. The lower panel reveals relatively similar proportions of nouns that appear prior to the investigated German insertions. This observation demonstrates that noun repetition does not affect switch placement in the prepositional phrase in my bilingual corpus. Nevertheless, we cannot exclude an interaction between the repetition of a noun and the repetition of a preposition, especially as in the case of phrase repetition. For example:

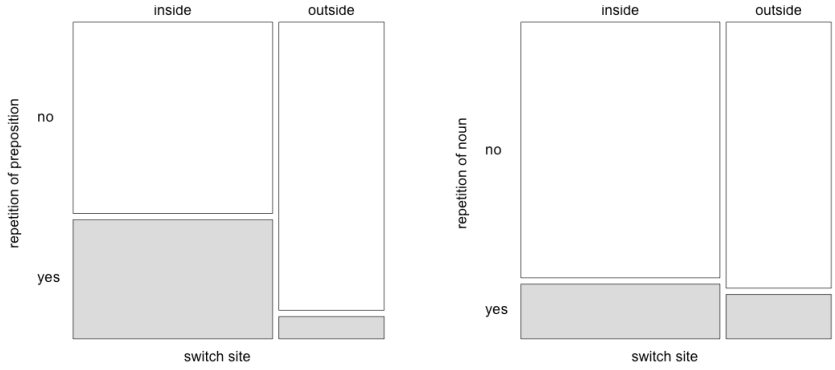

Figure 5.4: The relationship between switch placement and word repetition in the same language. The upper panel represents prior occurrences of target prepositions, the lower panel demonstrates prior occurrences of target nouns.

(16) (HO1007)
ja im govor-ju *wenn ich* *was* *von* *gott*
1SG.NOM 3PL.DAT tell-PRS.1SG if 1SG.NOM something from god
will *geh-e* *ich* *IN die* *kirche* a oni
want.PRS.1SG go.PRS-1SG 1SG.NOM to ART.ACC.SG church but 3PL.NOM
mne nača-l-i *IN der* *kirche* kak budto s *gott*
1SG.DAT begin-PST-PL in ART.DAT.SG church as if with god[?SG]

čë-to oni ne tak dela-jut čë *gott*
something.ACC 3PL.NOM NEG so do-PRS3PL what.ACC god[NOM.SG]
ime-et v vid-u
have-PRS3SG in view-LOC.SG
'I tell them, if I want something from God, I go to church, but they began like, in church people do not treat God as he intends them to.'

Here, the phrase *in die Kirche* 'to church' appears in a German clause for the first time and is then echoed in a mixed clause, albeit with the article in a different case. Although this kind of repetition is relatively rare, an interaction between the variables "repetition of preposition" and "repetition of noun" would be able to account for this example. In most cases, however, occurrences of prepositions in the prior discourse affect the choice of the preposition in the target phrase. In other words, prepositions seem to be stronger primes than nouns.

To summarise, I have shown thus far that four factors are predictive of switch placement in the context of the prepositional phrase: (*i*) the odds based on the frequency with which prepositions and nouns appear together, (*ii*) the frequency of the noun, (*iii*) the presence, or absence, of prepositions in the same language in the prior discourse and (*iv*) the presence, or absence, of nouns in the prior discourse, albeit only when used with the repeated preposition. As detailed above, instances such as (15) cannot be attributed to a single factor. This circumstance necessitates examining the pertinent factors as potential determinants of the scrutinised variation by means of a statistical test, which would simultaneously take account of the factors as well as their interactions.

5.4 Statistical prediction of switch placement

The aim of this section is to investigate the factors described above as they compete and interact with one another while conditioning switch placement in the prepositional phrase. Addressing the issues of competition and interaction entails a number of questions: How predictive is noun frequency of switch placement, all things being equal? Which factor is the most important predictor of the observed variation? Do individual preferences for switch placement override the regularity of the identified linguistic tendencies? Like §4.5, this chapter investigates these issues by using the generalised linear mixed model. Probabilities of binary outcomes – a switch within a prepositional phrase and a switch at the boundary of a prepositional phrase – will be determined on the basis of the predictor variables, i.e., the factors analysed above. Significant interactions

between the dependent variable "switch placement" and the predictor variables will provide objective evidence of the relevance of each of the predictors.

5.4.1 Model fitting

In order to obtain a minimal adequate regression model, the common procedure (Baayen 2013, Szmrecsanyi 2013) was employed, which is as follows. The maximal model contained the four factors considered above as main effects: the odds, based on the frequency with which the examined nouns are used with prepositions; the frequency of the involved noun; the prior occurrences and non-occurrences of the target noun; and the prior occurrences and non-occurrences of the target preposition. The maximal model also included interactions between these factors. The speakers' individual differences in mixing behaviour – a propensity to insert German nouns into Russian prepositional phrases or, rather, a predilection for maintaining the integrity of the phrase – may alter the tendencies determined solely by the fixed factors. These individual differences are considered in the random variable "speaker". A random effect for the variable "item" could not be included owing to its high variability, namely, 197 various lexemes appear in the 244 prepositional phrases under analysis. The model was thus run without the by-item random variable. The model simplification procedure consisted in the omission of the factors and interaction terms that added no significant explanatory power to the model. According to Baayen (2008: 281), the calculation of the *C* index of concordance is the basis for estimating the explanatory power of main effects and interaction terms. The process of model reduction resulted in the exclusion of the following interaction terms from the model: noun frequency × odds and prior noun × noun frequency. Subsequently, the inclusion of the by-subject random effect was justified by the estimation of the *C* index of a generalised linear model without random effects. The model that included the random factor appeared to perform significantly better than the model without a random effect (which is in line with established practice in similar cases, cf. Bresnan et al. 2007; Tagliamonte & Baayen 2012). The final, minimal adequate model is presented in Table 5.5.

5.4.2 Model evaluation and model discussion

The minimal adequate model reported in Table 5.5 is of high quality. The model correctly predicts 84% of all instances of switch placement in the data set, while the categorical prediction of switch placement by always guessing a switch placed within the prepositional phrase will be correct in 69% of all cases. The *C* index

Table 5.5: Predicting switch placement in the context of the prepositional phrase: minimal adequate generalised linear mixed model. Predicted odds ratios are for switches placed at the phrase boundary. Significance codes: *significant at $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Factor	Odds	Est.	Pr(> |z|)	
(Intercept)	0.012	−4.427	0	***
Prior preposition	0.030	−3.496	0	***
Prior noun	0.440	−0.820	0.178	
Noun frequency	1.568	0.449	0	***
Log odds	1.359	0.307	0.001	**
Prior preposition × prior noun	24.753	3.209	0.007	**
Random effect:				
Speaker				
(intercept, $N = 19$, variance $= 0.915$, $\sigma = 0.956$)				
Summary statistics:				
N		244		
% correct predictions (% baseline)		84 (69)		
C index of concordance		0.871		
Somer's Dxy		0.742		

of concordance between the predicted probability and the observed binary outcomes is 0.871, which indicates that the model has high predictive power. The performance indicator Somers' D_{xy}, a rank correlation coefficient between predicted probabilities and observed binary response, is 0.725, which again signals the model's high predictive capacity. Regarding the random effect for the variable "speaker", we can conclude from its estimated variance and standard deviation that the variation among speakers, although minor, still contributes to the distribution of the examined data.

The main effects in the model provide positive evidence for the assumptions formulated above. The signs of the regression coefficients (Estimates) in Table 5.5 reveal the direction of the adjustment to the intercept. Hence, a prior occurrence of the target preposition and that of the target noun condition the placing of the switch between the preposition and the noun, whereas both odds, based on co-occurrence frequency, and noun frequency favour switching at the phrase boundary. Consider the odds ratios reported in Table 5.5. Among the fixed-effect factors, repetition of the preposition has the largest effect size, which means that

if the target preposition appears in the prior discourse, the likelihood of switching the language at the phrase boundary decreases by 97%. That is, there is a high probability that the previously occurring preposition will be selected in the target prepositional phrase. The interaction between a prior occurrence of the target preposition and that of the target noun in the discourse is another strong effect: If both the preposition and the noun appear in the preceding discourse, the odds for switch placement at the phrase boundary increase by a factor of approximately 25. This interaction term accounts for the case illustrated by example (15). But when the occurrences of the noun and the preposition in the prior discourse do not interact, both of them favour switch placement within the prepositional phrase. Although the predictor "prior occurrence of noun" does not attain statistical significance, it interacts with the factor "prior occurrence of preposition", enhancing the model's predicative capacity. This is the main reason for retaining this predictor in the minimal adequate model. Moreover, this interaction reaches a high level of statistical significance.

After considering the effect sizes of the factors contributing to the minimal adequate model, it is necessary to discuss these factors' overall importance. This parameter is visualised in Figure 5.5 by plotting the decrease in the Akaike Information Criterion (AIC) of the model if a factor is removed from the minimal model. According to Szmrecsanyi (2013), more sizable decreases in the AIC criterion of a factor stand for its increased overall importance. In predicting switch placement in the context of 244 prepositional phrases from the data set, the most important factor is "prior occurrence of the target preposition". Noun frequency is the second most relevant predictor, followed by "(log) odds", which is based on the measured co-occurrence frequency. The model's Akaike Information Criterion sinks drastically when the factor "noun frequency" is removed from the model. The extent of this decrease appears to substantially exceed the extent of the decrease of factor "(log) odds". The interaction between the prior occurrence of the noun and that of the preposition is ranked last. The prior occurrence of the noun is not among the plotted factors owing to its low level of importance for the model, when included without an interaction term.

A somewhat surprising result of the minimal adequate model is that the predictor "noun frequency" is more important in accounting for variation in the data than the factor "(log) odds", which is a measure of competition among co-occurrences of nouns and prepositions. A question arises as to why "(log) odds" is a less important predictor than "noun frequency". Although "(log) odds" is the only factor that takes preposition-noun co-occurrences into consideration, we cannot rule out the possibility that it performs in the model inadequately. A possible reason for this under-performance is the severe competition among

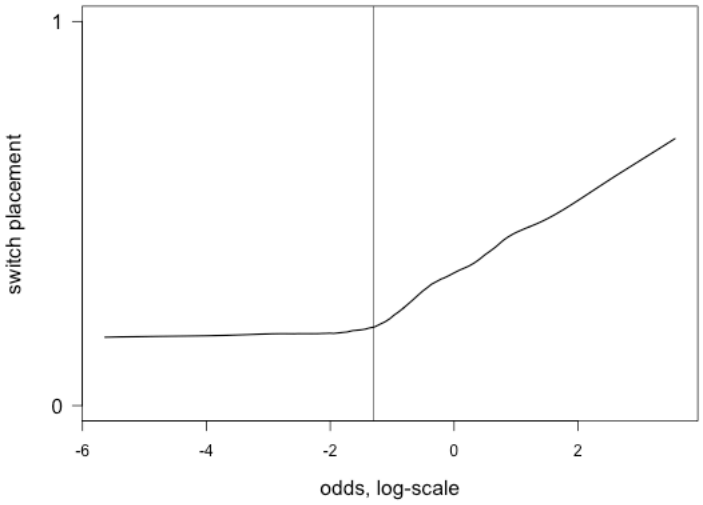

Figure 5.5: Importance of factors in model: decrease in Akaike Information Criterion (AIC) if factor removed. (The table representation is based on Szmrecsanyi 2013.)

chunks involving high-frequency nouns. As these nouns combine with several prepositions at high rates, the joint frequency of such combinations are likely to prevail over the frequency of a particular preposition-noun combination in focus. In other words, the higher the frequency of the noun, the larger the number of its collocates, usually referred to as "family size", and the more difficult it is for a specific combination to come out on top of the distribution and become selected in production. Table 5.6 provides an illustration of these intricacies.

Table 5.6 presents a set of prepositions co-occurring with the high-frequency noun *Ende* 'end'. As this noun combines with an array of prepositions at high rates, we may describe the competition among the thirteen prepositions as stiff. The preposition *an* 'at' accounts for the lion's share in the overall distribution of the [P – *Ende*] pattern. In fact, it is this preposition which the speaker selects. However, the summated frequencies of the noun's remaining companions outbalance the frequency of the competition winner *an* 'at'. As a result, the top collocate's odds value is lower than 1. It is noteworthy that the other high-frequency nouns in the data set are also affected. We can thus contend that instead of exerting a direct influence on switch placement, noun frequency may act as a leverage for combinations of prepositions and high-frequency nouns.

To summarise, in this section I have demonstrated that frequency-based factors, such as "noun frequency" and "odds", based on co-occurrence frequency, as

Table 5.6: Prepositions co-occurring with the German noun *Ende* 'end' and their respective co-occurrence frequencies.

Co-occurring prepositions	Frequency of co-occurrence
an 'at'	157220
zu 'to'	67684
bis 'by'	22152
nach 'after'	18279
gegen 'towards'	14240
seit 'since'	13063
von 'from'	7214
vor 'before'	6993
ohne 'without'	5242
mit 'with'	5153
für 'for'	3824
ab 'from'	3398
auf 'to'	2408

well as the processing-related factor "word repetition (priming)" effectively predict switch placement in the context of the prepositional phrase. Speaker variation in switching the language at the phrase boundary or within the phrase was found to be marginal.

5.5 Conclusions and discussion

This chapter has explored variation in switch placement in the prepositional phrase in Russian-German code-mixing as an outcome of factors related to intricacies of language use in terms of the speakers' experience with language as such, i.e., in the global sense, and their sensitivity to the use of linguistic elements in the immediately preceding discourse, i.e., at the local level. Specifically, it has focused on the role of factors such as word frequency, competition among multiword strings varying in usage frequency, and recency of a word in the discourse. Additionally, I have shown that a combination of corpus-linguistic, computational and statistical methods proves to be useful for analyses of variation in code-mixing patterns.

The principal findings of the study include two frequency effects: First, frequencies with which prepositions and nouns co-occur influence switch place-

ment in the context of the prepositional phrase. If a noun and a specific preposition appear together more frequently than all the other combinations of this noun with prepositions, a switch between this noun and the given preposition is unlikely. Conversely, if the frequency of co-occurrence of a noun and a specific preposition is low, the chance for a switch to be placed within the prepositional phrase is high. The fact that high-frequency preposition-noun combinations are impervious to phrase-internal switching is explained in terms of their holistic representation in the speakers' mental lexicon. Whenever two or more items co-occur on a regular basis in one of the languages involved in code-mixing, speakers store and retrieve these multiword sequences from memory as units, and their production becomes automatised (cf. Chapter 2). This finding accounts for embedded-language islands as word strings welded together in language usage. Similar to the results reported in Chapter 4, it corroborates Backus' (2003) unit hypothesis. The competition among multiword strings structured as prepositional phrases was modelled by (log) odds.

Another finding concerns a strong correlation between the frequency of the nouns involved in the examined prepositional phrases and switch placement. As reported above, German high-frequency nouns are accompanied by German prepositions in mixed sentences more often than by Russian prepositions. That is, the speakers switch the language to German at the phrase boundary in this case. I have argued that this tendency is due to the fact that high-frequency nouns usually form several recurrent multiword strings with various prepositions. The stiff competition among these strings is likely to weaken the predictive capacity of the competition measure odds. As a result, the frequency of the noun appears to be a more important predictor of the variation in the data than the odds. However, as stated above, the influence of noun frequency on switch placement may be indirect. Interestingly, in the context of the adjective-modified noun phase (cf. Chapter 4), noun frequency appeared to exert no significant influence on the variation in the mixing patters. Another interpretation of the noun frequency effect pertains to the accessibility of high-frequency nouns and their lexicogrammatical patterns, including multiword chunks, in production. Being extremely accessible, these nouns may easily co-activate the contexts in which they typically occur. Arnon & Snider (2010) have presented evidence of a similar effect in language comprehension.

The statistical analysis conducted reveals that the most important predictor of switch placement in the examined context is a prior occurrence of the preposition in the prior discourse. The chances for producing the preposition in one of the involved languages are higher if this preposition also occurs in the same language

during the previous eight seconds of discourse, provided that semantic compatibility is maintained. The occurrence of the preposition in the prior discourse, which is usually framed by the matrix language Russian, has, in most cases, the use of the Russian preposition in the target phrase, or placing the switch within the prepositional phrase. This recency effect is so strong that it overrules the effects of frequency discussed above. This tendency can be attributed to the following factors: First, the investigated prepositions are distinguished by high usage frequencies, and are extremely polysemous (the data set includes no secondary prepositions). Apparently, their polysemous nature and frequent use are interdependent. Second, like pronouns, which are the focus of analysis in (Travis & Torres Cacoullos 2016), prepositions are function words and seem to go unnoticed in language production. That is, function words appear to be stronger primes than content words. Third, code-mixing occasionally co-occurs with self-repair, in which case part of the lexical material remains preserved in the target structure, as in the example below.

(17) (LR0316)
stra- straße(n)bahn priezža-l v dvadcat' četyre minut-y v(.)
tram[NOM.SG] arrive-PST.M.SG at twenty four minute-GEN.SG at
also v *einundzwanzig* dvadcat' četyre
PTCL at twenty-one twenty four
'The tram arrived at twenty-four minutes, well, at nine twenty-four.'

The preposition *v* 'at' appears before the mixed phrase *v einundzwanzig dvadcat' četyre* 'at twenty-one twenty-four' in an instance of self-repair, but also earlier, at the beginning of the utterance.

Future work should therefore include follow-up research designed to examine patterns of code-mixing with regard to self-repair as well as repetition priming. An important question for future studies is to elucidate the role of noun frequency for switching preferences in the prepositional phrase and to develop and test other models of competition among functionally and structurally similar word strings.

The findings presented in this chapter parallel the results reported in Chapter 4 in that in the prepositional phrase as well as in the adjective-modified noun phrase, the choice between producing a mixed constituent and inserting an embedded-language island, appears to depend on the frequency of the linguistic material and its accessibility in the memory, referred to as recency in discourse and repetition priming. This chapter has also shown that in addition to usage frequency, the choice is influenced by the competition among the prepositions

occurring with a specific noun at varying rates. In situations in which a clear winner is identifiable, the probability of selecting the string with this preposition is extremely high.

The difference between the distribution of the mixing patterns in the prepositional phrase and in the adjective-modified noun phrase pertains to the nature of the elements involved. While the latter syntactic context is determined by open-class words, the syntactic context in the present chapter involves combinations of open-class and closed-class words. As mentioned in Chapter 4, the selection of nouns is an unconstrained process (Boumans 1998), whereas the selection of adjectives is less unconstrained (e.g., German attributive adjectives never modify Russian nouns). As to the selection of prepositions, the options are much more restricted. Of the analysed nouns, the majority regularly occurs with a handful of prepositions. That is, the choice between a Russian, and a German preposition is conditioned by the limited size of the available inventories. Against this background, it would be interesting to explore the selection of elements in a context permitting choices among even fewer alternatives. Regarding the competition among the available options, we may assume that it will be much stronger owing to a reduced number of competing candidates. It well may be that in this situation, frequency will play even a more crucial role in conditioning the mixing patterns. One context which allows us to investigate variation in mixing patterns involving a small inventory of elements is plural marking. In this context, embedded-language nouns in sentences framed by the matrix language either receive matrix language plural markers, or retain their embedded-language plural markers. It is to this study that I turn in the next chapter.

6 Plural marking of German noun insertions in bilingual sentences

The study presented in this chapter[1] differs from the analyses reported in the previous chapters in that it explores mixing patterns in the expression of plural, i.e., at the level of the morphological word, whereas the studies in the previous chapters investigated code-mixing beyond the level of the morphological word. The focus on morphological processes has ramifications for the types of mixing patterns and the number of elements, or forms, competing for selection in bilingual speech production.

In Russian-German bilingual sentences in which Russian is the matrix language, embedded-language, i.e., German, nouns may either receive plural marking from Russian, or retain their German markers, for instance:

(1) (LV120224-11)
tam očen' xoroš-ie *geschenk*-i. tam možno
there very good-NOM.PL present-NOM.PL.M there possible
punkt-y mnogo *punkt*-ov sobra-t' na *baby*-k-ax.
point-NOM.SG.M many point-NOM.SG.M collect-INF on baby-DIM-PREP.PL
na privivk-ax... na *u* *untersuchung-en*
on jab-PREP.SG.F on check-up check-up(F)-PL
'They have very good presents. You can earn many bonus points for babies, for jabs, for check-ups.'

This example contains several German nouns embedded in the Russian discourse. For now, let us focus on the nouns *Geschenk* 'present' and *Untersuchung* 'check-up'. The former noun receives the Russian inflectional suffix *-i* cumulatively expressing several categories including the plural, whereas the latter noun is inserted together with its German plural marker *-en*.[2] Again a distinction is

[1]This chapter is based on an earlier publication, see (Hakimov 2016b).

[2]The analysis of the noun *Punkt* '(bonus) point' is less straightforward. I am inclined to analyse it as a German insertion because the Russian word *punkt* does not refer to a bonus point, which is commonly expressed as *ball*. The speaker, who grew up in Germany, is likely to be

made between mixed constituents (e.g., *geschenki*) and embedded-language islands (e.g., *untersuchungen*). Crucially, the speaker might have handled these words differently, producing the embedded-language island *geschenke* and the mixed constituent *untersuchungax*. It is the choice between these two patterns of plural marking, which is the object of study in this chapter. Unlike the embedded-language islands analysed in the previous chapters, the islands examined here lack syntactic independence and are labelled as internal embedded-language islands (Myers-Scotton 2002: 149–150)[3]. Embedded-language plurals have often been discussed in the literature (see Backus 1996, 1999a, 2003, Boumans 1998, Muhamedowa 2006, Myers-Scotton 1993, 2002), but the proposed explanations for their emergence are still a matter of controversy. The question I address in this chapter is whether variation in patterns of plural marking of code-mixed nouns can be effectively predicted by a combination of structural and usage-based factors.

The first hypothesis of this study is that the frequency distribution of singular and plural forms in the embedded language, German, will influence patterns of plural marking of German nouns in Russian sentences. Words that are often used in their plural forms in one language are inserted as plurals into another language (Backus 1996, 1999a, 2003). Highly entrenched German plurals will thus retain German plural markers in Russian sentences. Another consequence of the frequency distribution of singular and plural forms in the embedded language is that nouns frequent in the singular in German will receive their plural marking from Russian, the matrix language. As a matrix language, Russian restricts the possibilities for accommodation of German noun insertions on the morpho-phonological level. The Russian nominal system favours stems featuring consonants in the final position. Therefore, the second hypothesis is that German noun stems exhibiting vowels stem-finally will retain the German plural marker. That

unfamiliar with the Russian expression since the system of bonus points is a relatively new phenomenon in Russia. Yet, an orthodox analysis would have to conclude that this item is not a German insertion. The next noun is the German acronym *U*, utilised by a specific German health insurance company for *Untersuchung* 'check-up'. The speaker handles it as a Russian indeclinable noun, which is in line with the Russian grammar. Finally, the German noun *Baby* receives both an inflectional and a derivational, namely, diminutive suffix (for more details, see §6.5.1).

[3]The problem with considering embedded-language plurals as syntactically independent was raised by Boumans (1998: 36–37). I agree that this view is difficult to maintain for language constellations in which at least one of the languages relies on affixes fusing the number and the morphological case, such as Russian. Yet, I will keep this term in order to highlight the difference between embedded-language islands representing co-occurrences of words and those representing morpheme co-occurrences.

is, the phonemic shape of German nouns is a factor determining the choice of language for marking the plural on inserted stems. Another motivation behind the observed variation is a mismatch between the nominal systems of German and Russian: the Russian system exhibits a lesser degree of syncretism and contains more morphological cases than German. The final hypothesis is therefore that an insertion will receive a Russian plural affix if the slot of the insertion requires a non-core case. This is of particular interest in a situation of contact between Russian and German, two fusional languages, as previous research on variation in plural marking of noun insertions has thus far only involved pairs of agglutinative and analytical languages.

This chapter is organised as follows: Section 1 presents a survey of patterns of plural marking on code-mixed nouns by drawing on instances observed in corpora of bilingual speech involving various language pairs. Section 2 reports the state of the art of code-mixing research concerned with explaining the distribution of these patterns in mixed sentences. Section 3 then offers an overview of the systems of plural marking in Russian and German. Section 4 is concerned with the analysis of the factors determining the studied variation: these include the frequencies of the singular and plural forms of the inserted lexical items as distributed in the embedded language, phonemic shape of the German noun insertions, including the segment featured stem-finally, and the morphosyntactic context in which the nouns appear, i.e., the morphological case required by the slot in which they are to be inserted. Each of these factors is a tendency rather than an absolute rule, therefore they are evaluated statistically in Section 5, which describes the statistical model testing the interplay of factors when predicting the use of the examined patterns. Section 6 summarises and discusses the results.

6.1 Typology of marking plural on code-mixed nouns

Code-mixed nouns can receive plural marking in three different ways: the first possibility results from the morphological processes of the matrix language, as in (2a); the second possibility to mark the plural on code-mixed nouns is to retain the plural marker used with the word in the embedded language, i.e, the language of the inserted stem, as in (2b); the final option is double plural marking, which refers to the case when the noun receives two plural markers: one from the embedded language and the other from the matrix language, as in (2c).

(2) a. $_{A}[\ _{B}[\ \text{N}\ _{B}]\text{-PL}\ _{A}]$
b. $_{A}[\ _{B}[\ \text{N-PL}\ _{B}]\ _{A}]$
c. $_{A}[\ _{B}[\ \text{N-PL}\ _{B}]\text{-PL}_{A}]$

6.1.1 Type 1: A morphological process of the matrix language

Marking the plural by the morphological process of the matrix language is especially frequent in situations in which the matrix language is an agglutinative language. In (3), the Dutch noun *activiteit* 'activity', embedded into a Turkish clause, is used with the Turkish plural suffix *-ler-*, and in (4), the English noun *rule* acquires the Finnish stem formant *-i* and the plural marker *-t*.

(3) Turkish-Dutch (Backus 1996: 150)
activiteit-ler-i yapacağız dedik
activity-PL-ACC do:FUT;1PL said:1PL
'we said we were going to activities'

(4) Finnish-English (Halmari 1997: 60)
joo missä kummassa ne *rule*-i-t on?
yeah where ever 3PL -SF-PL are
'Yeah, where on earth are those rules?'

The high frequency of this kind of insertion in language pairs with an agglutinative language as the matrix language, and its rarity in pairs of languages with fully fusional Arabic (Boumans 1998: 180; Nortier 1990: 189) and moderately fusional French and Dutch as the matrix language (Treffers-Daller 1994) led Muysken to hypothesise that fusional languages are resistant to this pattern (2000: 77). However, from as early as Hasselmo (1972: 265–266) there is documented evidence in the literature against this claim (see Budzhak-Jones 1998: 174, for Ukrainian-English; Hlavac 2003: 73, for Croatian-English; Stenson 1990: 180, for Irish-English; Szabó 2010: 352, for German-Hungarian). In line with this evidence are the following examples of embedded-language nouns in Russian sentences:

(5) Russian-English (Benson 1960: 173)
ves' *place* by-l zaparkova-n *car*-ami
whole place[NOM.SG] be-PST[SG.M] park-PART[SG.M] -INSTR.PL
'The whole place was parked full of cars.'

(6) Russian-Hebrew (Naiditch 2008: 48)
kogda u vas nač-n-ut-sja *bagrut*-y?
when with 2GEN.PL begin-PERF-3PL-REFL exam-NOM.PL
'When will your exams begin?'

The nouns *car* in (5) and *bagrut* 'exam' in (6) take Russian inflectional affixes expressing the plural and a morphological case, which is also a typical pattern in the Russian-German code-mixing data analysed below.

6.1.2 Type 2: A morphological process of the embedded language

As mentioned above, another possibility to express the plural on noun insertions in bilingual sentences is to resort to a process employed by the embedded language so that an internal embedded-language island occurs. In (7), the noun *window* is inserted together with its English plural marker *-s*; similarly in (8), the plural form *dvarim* 'things' of the Hebrew noun *davar* is embedded in the Russian clause.

(7) Russian-English (Benson 1960: 173)
ona poš-l-a *clean*-ova-t' *window-s*
3SG.F go-PST-SG.FS -SF-INF -PL
'She went to clean windows.'

(8) Russian-Hebrew (Naiditch 2008: 48)
est' neskol'ko *dvar-im* kotor-ye my dolžn-y zna-t'
be.PRS several PL\thing-PL which-ACC.PL 1PL obliged-1PL know-INF
'There are several things we have to know.'

Both plurals in (7) and in (8) lack case marking, which is required in Russian, and thus count as internal embedded-language islands. The Russian sentence in (9) contains the Estonian form *vaimusid* 'ghosts', which is inflected for the partitive case. This is obviously possible owing to a functional similarity in this specific context between the Estonian partitive and the Russian genitive, required by the verb *bojus'* 'am afraid'.

(9) Russian-Estonian (Mürkhein 1970: 117, quoted from Verschik 2004: 437)
ja boj-u-s' *vaim-u-sid*
1SG fear-1SG-REFL ghost-SF-PARTITIVE.PL
'I am afraid of ghosts.'

Some agglutinative languages, such as Tamil and Kazakh (Muhamedowa 2006: 67), have been reported to combine embedded-language plural markers with matrix language case markers, as in (10). This is indicative of the gradience of the morphosyntactic integration of noun insertions.

(10) Tamil-English (Sankoff et al. 1990: 81)
only kalaimagaLLataan *movie-s*-e patti peece ille
only proper.name.LOC.only movie-PL-ACC about talk NEG
'Only in KalaimagaL there is no talk about the movies.'

6.1.3 Type 3: Double plural marking

Double plural marking is probably the most frequently described case of double morphology documented in the literature and has therefore attracted the most scholarly interest (Backus 1992: 96, 1996: 151, 1999a: 98–99; Boumans 1998: 90–91; Muhamedowa 2006: 152–156; Myers-Scotton 1993: 132–135, Myers-Scotton & Jake 1995). For example:

(11) Kazakh-Russian (Muhamedowa 2006: 152)
sonday ülken *zvanij-a*-lar-ï bar
such big title-PL.NOM-PL-POS3SG EXIST.3SG
'He has many titles.'

Here the Russian plural noun *zvanija* 'titles' additionally receives the Kazakh plural marker *-lar-*. Another example comes from Moluccan Malay-Dutch code-mixing in the Netherlands and showcases the use of reduplication as a productive process of plural marking in Moluccan Malay (Voigt 1994: 50–56):

(12) Moluccan Malay-Dutch (Voigt 1994: 52)
kalau minum terus ada di punya *pitje-s-pitje-s*-nya
if drink continue exist DET seed-PL-seed-PL-DET
'If I continue drinking there'll be seeds.'

6.2 Previous explanations

Prior studies have predominantly focused on double plural marking in mixed nouns, and double morphology in general, since the latter was considered an exceptional case in the early version of Myers-Scotton's 1993 influential Matrix Language Frame (MLF) model. The proponents of the MLF model attributed double morphology to erroneous access in production (Myers-Scotton 1993: 132–136; Myers-Scotton & Jake 1995: 1000). Although Myers-Scotton (1993: 134) suggests possible scenarios for double morphology, it is not clear what factors determine its occurrence in bilingual corpora.

A structural explanation of the phenomenon was suggested by Boumans (1998: 91), who considers a mismatch in the morphological marking processes to be a crucial factor determining the emergence of double morphology. According to Boumans, "[t]he likelihood of double marking appears to increase when each language marks the same feature in a different manner, for instance, by means of prefixes and suffixes" (ibid.). This approach sheds light on a number of cases of double morphology found in bilingual corpora, including, for instance, the Moluccan

Malay-Dutch example (12). The morphological processes of plural marking mismatch in this case because Malay uses reduplication whilst Dutch employs suffixes. The dissimilarity leads to the emergence of a form with the plural marked twice. Boumans' proposal has been widely echoed in the literature. Indeed, it not only holds for double plurals, but also for other cases of double morphology (see Myers-Scotton 2002: 91, for double marking of determiners and infinitives; Muysken 2000: 104 and Muhamedowa 2006: 152–156, for doubled adpositions; Szabó 2010: 346–352, for doubled adpositions and determiners). Myers-Scotton (2002: 150) extends this principle to the case of plural marking on inserted nouns by the embedded language. It can, for example, aptly account for the emergence of the internal embedded-language island in (8). Its emergence may be attributed to differences in the morphological processes involved in plural marking: while Hebrew relies on both stem change and suffixation, Russian draws only on inflectional suffixes.

This factor, however, has a restricted explanatory power for it cannot account for the use of double morphology in two cases. First, it is silent on the appearance of embedded-language islands in bilingual speech involving languages which use similar morphological processes for marking a particular feature, such as Kazakh and Russian (cf. example 11). Second, even when the processes for marking plural do diverge, embedded-language markers may alternate with matrix language markers, as in (13):

(13) English-Hebrew (Olshtain & Blum-Kulka 1989: 70)

Mother:	What do I have to do when I go to your *gan* (nursery.school)?
Son (15):	Wait, they are on strike, the *gan-im* (nursery.school-PL)?
Mother:	No, the *ozrot* (PL\assistant) are on strike. In order not to close down the *gan* the mothers are taking *tor-ot* (turn-PL) to help the *ganenet*-s (nursery.teacher-PL). The assistants to the *ganenet*-s (nursery.teacher-PL). They're young women who have taken a year's course to be an *ozeret* (assistant) for *gan-im* (nursery.school-PL).
Son (15):	So why are they going on strike and not the *gananot* (PL\nursery.teacher)?

This example demonstrates high variability of plural marking within a short piece of conversation. Hebrew morphology is retained in most of the plural forms: *ganim* 'nursery schools', *ozrot* 'assistants', *torot* 'turns' and *gananot* 'nursery teachers'. Of these plurals, the forms *ozrot* and *gananot* can be explained by a mismatch in the morphological processes because they express the plural by

using stem change, and not suffixation as English nouns. Nevertheless, on another occurrence the noun *ganenet* 'nursery teacher' receives the English plural marker *-s*.

In discussing the same phenomenon, Backus (1996: 151) posits an alternative proposal. Within a Cognitive Linguistics framework, he provides a usage-based explanation of double marking. Emphasising the role of entrenchment of certain plural forms, he observes that some nouns occur in their plural form more often than in their singular form. The high degree of entrenchment of these plural forms leads to a facilitation of their activation in production, and is attributed to a mismatch in the frequency distribution of the plural and singular forms of nouns. In his later work, Backus (1999a: 98) elaborates this idea further : "[… embedded-language plurals] are established lexical units for the speakers who use them. This means they are chunks […]: they are lexical units that consist of more than one morpheme." This assumption is based on the idea that the strength of cognitive representations at various linguistic levels is a function of usage frequency, and it is thus in line with usage-based linguistic theory (Bybee 1985, 2006, 2010). A rich memory for language (Langacker 1987, 2000, Tomasello 2003, Bybee 2010) allows for both plural and singular forms of same nouns to exist independently in the lexicon. (More details, including the findings on the processing of singulars and plurals, are reported in §2.2.3.) Although Backus's reasoning is highly plausible, it cannot be viewed as a single principle determining the emergence of embedded-language plurals. If this were the case, any embedded-language noun that appeared with an embedded-language plural marker would be granted the status of a lexical unit. When modelling the probability of activating a plural form in production, we need to consider the competition between the representations of plural and singular forms of the inserted lexemes. Baayen et al. (1997: 97) describe this competition in terms of distribution dominance and distinguish between singular-dominant and plural-dominant nouns: "Either the singulars were much more frequent than their plurals (singular-dominant) or the plurals had a much higher surface frequency than their singulars (plural-dominant pairs)."

As to other factors, phonotactic restrictions and morphophonological regularities have been largely neglected in the literature as possible motivations for the use of plural marking patterns. However, as will be shown below, they are among important requirements imposed by the matrix language on items that are to be inserted in the structure it frames.

In sum, the appearance of embedded-language plurals as internal embedded-language islands, or as parts of double-marked constituents can be attributed either to a mismatch in the processes of plural marking, or singular-plural distribution dominance, i.e., an asymmetry in the token frequency of a lexical item's

singular and plural. Some other factors such as morphophonology have not yet been considered pertinent thus far. A limitation that pervades all the approaches to double morphology discussed is their sole reliance on individual factors and therefore a tendency to overlook the multi-factorial nature of the phenomenon. In order to account for the variation in the data, we need to analyse the possible driving forces behind it systematically and in their interaction.

6.3 Plural marking in Russian and German

Russian-German code-mixing provides a good testing ground for the interaction of factors that contribute to variable patterns of plural marking because they fusional languages. Both inflect nouns for number, gender and case, but demonstrate varying degrees of syncretism and systematicity in the encoding of these grammatical categories.

6.3.1 Russian

Russian noun declension fuses case, number, and gender marking. The system is stem-based and synthetic, and can be best described by stipulating declensional classes. However, there is no consensus in the literature regarding the number of the declensional classes and the determinants for distinguishing them (cf. Corbett 1991, 2003). In this paper I adopt Zaliznjak's 2002 [1967], 2009 [1977] approach to the Russian nominal inflection. The following declensional classes are relevant for the presentation of the data below: the "masculine", the "feminine", and the zero-marked class. Based on the gender and phonological type of the stem, three principal classes are traditionally differentiated, each being prototypical for the respective gender: the masculine, the feminine, and the neuter. The three classes are presented in 6.1. Of these classes, the "masculine" and "feminine" classes are the most productive ones (Zaliznjak 2002 [1967]: 218; Timberlake 2004: 148). It is thus not surprising that they play a crucial role in the integration of the other-language noun stems in Russian sentences in bilingual speech. This tendency has been reported for Russian-Hebrew code-mixing (Naiditch 2008) and is observed in my Russian-German data. Crucially, the distinctions between the declensional classes in the plural, when compared to those in the singular, are minimal: each morphological case has one inflection per class except for the genitive and the nominative, whose forms coincide with those of the accusative.

The last relevant class relies on zero marking and is restricted to a group of loan words ending in vowels, except unstressed *-a* preceded by a consonant

Table 6.1: Inflections of the Russian nominal declension (adapted from Zaliznjak 2009 [1977]: 26).

		singular			plural		
		m	n	f	m	n	f
nominative		∅	-o	-a	-y (-i)	-a	-y (-i)
genitive		-a		-y (-i)	-ov (-ej)	∅	
dative		-u		-e	-am		
accusative	inanimate	∅	-o	-u	-y (-i)	-a	-y (-i)
accusative	animate	-a	-o	-u	-ov (-ej)	∅	
instrumental		-om		-oj	-ami		
prepositional			-e		-ax		

(Timberlake 2004: 148–149). This class includes nouns such as *xudi* 'hoodie', *boa* 'feather boa', *kafe* 'cafè', *kenguru* 'kangaroo', *kino* 'cinema'. None of the nominal grammatical categories is marked overtly on the nouns of this class; compare *ljubimoe kafe* 'a favorite cafè' versus *raznye kafe* 'different cafès'.

Russian employs the genitive singular to express plurality when nouns follow the cardinal numerals two, three, and four, for example: *dv-a dom-a* 'two-M house-GEN.SG.M' and *dv-e knig-i* 'two-F book-GEN.SG.F'. If we assume conceptual plurality, the instances of German noun insertions *dv-a mensch-a* 'two-M person-GEN.SG.M' and *dv-e flasch-i* 'two-F bottle-GEN.SG.F' should be considered as a case of plural marking.

6.3.2 German

Like Russian, German noun inflection fuses case, number, and gender marking. Furthermore, the nominal system is also traditionally analysed in terms of declensional classes (Eisenberg 2006: 158; *Duden 04*: 229). Because the plural demonstrates a high degree of syncretism, some scholars handle the patterns of plural marking and the patterns of the nominal inflection in the singular separately, given that overt case marking is the exception in the paradigm (Flämig 1991, Helbig & Buscha 2001, Hentschel & Weydt 2003). In the plural, the only case opposition, characteristic of most patterns of plural marking, is between the dative, which is marked by the suffix *-(e)n*, and the non-dative. Interestingly, German noun insertions in my data set never carry the exponent of the dative in the plural. In other words, these German plurals do not exhibit German case morphology.

German employs four mechanisms of plural marking (Flämig 1991: 480): (i) zero marking (*Schüler* 'pupil[SG]') – *Schüler* 'pupil[PL]'), (ii) the use of umlaut, or vowel alternation (*Garten* 'garden[SG]' – *Gärten* 'PL\garden'), (iii) suffixation (*Arm* 'arm[SG]' – *Arm-e* 'arm-PL'), (iv) a combination of umlaut and a plural suffix (*Buch* 'book[SG]' – *Büch-er* 'PL\ book-PL'). The patterns of German plural marking, drawing on these mechanisms, are given in Table 6.2.

Table 6.2: Patterns of German plural marking (adapted from Flämig 1991: 480).

Pattern no.	Example of singular	Pattern		Example of plural
		Suffix	Umlaut	
1	*Lehrer* 'teacher'	-∅	–	*Lehrer*
2	*Kloster* 'cloister'	-∅	+	*Klöster*
3	*Tag* 'day'	-e	–	*Tage*
4	*Kopf* 'head'	-e	+	*Köpfe*
5	*Kind* 'child'	-er	–	*Kinder*
6	*Mann* 'man'	-er	+	*Männer*
7	*Name(n)* 'name'	-n	–	*Namen*
8	*Mensch(en)* 'person'	-(e)n	–	*Menschen*
9	*Auto* 'car'	-s	–	*Autos*

6.4 Patterns of plural marking on code-mixed German nouns in the bilingual corpus

A total of 153 German noun insertions in Russian sentences were identified in my corpus of Russian-German bilingual speech as marked for the plural. The German nominalised adjectives *Mehrsprachige* 'multilinguals', *Russlanddeutsche* 'Russian Germans' and *Verwandte* 'relatives' were not counted as nouns because they decline like adjectives and do not have stable plural forms.

Russian plural markers are found on 72 insertions. Together with the plural number, these forms also express all possible morphological cases. In (14), for instance, the German noun *Augenarzt* 'ophthalmologist' is used with the Russian inflection of the genitive plural.

(14) (LA050310)

malo *augenarzt*-ov
few ophthalmologist-GEN.PL
'There are few ophthalmologists.'

Of the German noun insertions found in the corpus, 69 appear with German plural markers, as in (15).

(15) (LG050311)
čë na nej za *klamotte-n* ode-t-y
what on 3PREP.SG for rag-PL clothe-PART-PL
'What rags is she wearing?'

Here, the colloquial German noun *Klamotten* 'rags' (referring to 'clothes') is inserted into a Russian clause frame in its plural form.

Assignment of German noun insertions to one of the two patterns – $_{R}[\ _{G}[$ N-PL $_{G}]\ _{R}]$ and $_{R}[\ _{G}[$ N $_{G}]$-PL $_{R}]$ – is uncomplicated for most instances. However, attributing certain insertions to one of the patterns on the basis of their shape is not always straightforward. Such is the case of German nouns ending in /r/. This phoneme has two typical phonetic realisations in the corpus: a near-open central vowel [ɐ] and an alveolar vibrant consonant [r]. It is necessary to note that the consonantal /r/ in German is phonologically real. For example, the phoneme /r/ is realised as a vowel in the German suffix *-er* when the phoneme occurs in the word-final position, as in *Lehrer* ['leːʀɐ] 'teacher(M)' and *jüng-er* ['jʏŋɐ] 'young-COMP'. However, when another suffix is added to {-er} such that the phoneme occurs in an intervocalic position, /r/ is realised as a consonant[4], as in *Lehrer-in* ['leːʀərɪn] 'teacher-F' and *jüng-er-e* ['jʏŋərə] 'young-COMP-SG'. This variability has direct consequences for the morphological integration of the noun insertions with /r/ in the stem-final position.

In order to become integrated into the Russian declensional system, a noun stem usually has to feature a consonant in the final position (see §6.5.1 for further details). Hence, integration of the German nouns featuring /r/ stem-finally is unambiguous when /r/ is realised as a consonant. This is the case in the following eight tokens: *ausländе*[r]-*ov* 'foreigner-GEN.PL', *baue*[r]-*á* 'peasant-NOM.PL', *hamste*[r]-*y* 'hamster-NOM.PL', *hauptsemina*[r]-*y* 'advanced.seminar-ACC.PL', *opfe*[r]-*y* 'loser-NOM.PL', *penne*[r]-*y* 'tramp-NOM.PL', *pflaste*[r]-*ax* 'plaster-PREP.PL', *studiengebüh*[r]-*y* 'tuition.fee-ACC.PL'. In parallel to these forms, the data contain

[4]In German, at least four consonantal realisations of the phoneme /r/ are distinguished: [r], [ɾ], [ʀ] and [ʁ]. These variants occur in free variation, which can be attributed to different regional varieties of German (Kohler 1995: 165–166; Ramers & Vater 1992: 37–38, 110–112).

instances of German noun insertions with the vowel [ɐ] in the stem-final position, these include *anfänger* 'beginner', *aschenbecher* 'ashtray' (two tokens), *dinosaurier* 'dinosaur' (two tokens), *finger*, *inliner* 'rollerblade', *kleiderständer* 'coat-stand', *mitarbeiter* 'employee', *obstbecher* 'fruit cup', *zigeuner* 'gipsy', *zuschauer* 'spectator'. As the singular and plural forms of these nouns coincide in German, the plural is marked only syntactically, i.e., on the noun phrase but not on the nominal stem (the dative being an exception): *ein artiger Schüler* 'a good pupil' versus *viele artige Schüler* 'many good pupils' (see pattern 1 in 6.2). Russian also allows for this strategy, as it has a small group of stems with vowels in the final position whose plural and case forms are marked by zero. Therefore, German noun insertions with the stem-final /r/ realised as the vowel [ɐ] express the plural syntactically in Russian sentences as well, for example:

(16) (LB071401)
ty ne naš-l-a tak-ie *kleiderständ*[ɐ]
2SG NEG find-PST-SG.F such-ACC.PL coat.stand[PL]
'Have you found such coat stands?'

Here, the plural is marked on the adjective *takie*, and the noun *kleiderständer* is analysed as a Russian indeclinable.

The corpus contains one instance of syntactic marking of the plural on a stem featuring a vowel in the final position. Overt morphological marking of the plural is absent on the German noun *LKW* [ɛlkaˈveː], [ˈɛlkaveː] 'lorry' (17) although in German it has separate overt forms for the singular and the plural, i.e., *LKW* versus *LKWs*. I analyse this word as zero-marked in line with the Russian zero declensional class because the sentence is otherwise ungrammatical.

(17) (LR07141)
nu kogda èt-i [ɛlkaˈveː] grëban-ye proezža-jut
PTCL when this-NOM.PL lorry[PL] jiggered-NOM.PL pass.by.PRS-3PL
'But when those jiggered lorries pass by.'

This example supports the analysis that German nouns with a vowel in the stem-final position, when occurring in plural contexts, are treated as zero-marked in Russian discourse. Considering the fact that marking the plural by zero and expressing it only syntactically is the more economic strategy, the question arises why any of the stems ending in /r/ are marked for plural overtly at all. Note that there are eight tokens of nouns with the coronal trill that receive overt Russian plural markers and ten instances of stems with the vocalic realisation of /r/ at

the end. The tendency to mark the plural may be explained by the idea that plurality is expressed because it satisfies speakers' intentions (Myers-Scotton 2002: 150). In order to be more explicit, speakers may prefer an overt marker because the default case in Russian is to mark plural overtly (apart from the small class of zero-marked nouns, cf. Zaliznjak 2002 [1967]: 218). Additionally, due consideration must be given to inter-speaker variation in the pronunciation of /r/, i.e., whether or not the speakers tend to vocalise the coronal trill in all possible German contexts. Unfortunately, these issues cannot be investigated in depth in the current chapter due to the scarcity of the data. Nonetheless, in the following analysis, the nominal stems with the final vocalised /r/ will be considered as possible candidates for overt plural marking. As discussed above and shown in §6.5.1, they can either mark the plural by zero or take an overt Russian plural marker.

The examined data contain no instances of double plural marking. The distribution of the patterns of plural marking on code-mixed German nouns in the sample is given in Table 6.3. Below I will analyse the factors determining this variation. I will start with the influence of morphophonology on the patterns of plural marking by narrowing the focus on the bilingual speakers' strategies to integrate German nouns with stem-final vowels into Russian morphophonology. Subsequently, I will broaden my perspective to include all the pertinent insertions in the data set, regardless of their sound shapes.

Table 6.3: Distribution of patterns of plural marking with code-mixed German nouns in the sample.

Plural marking	Tokens	%
morphological (overt)		
Russian: $_{R}[\ _{G}[\ N\ _{G}]\text{-}PL\ _{R}]$	73	47.7
German: $_{R}[\ _{G}[\ N\text{-}PL\ _{G}]\ _{R}]$	69	45.1
syntactic (covert): $_{R}[\ A\text{-}PL\ _{G}[\ N\ _{G}]\ _{R}]$	11	7.2
Total	153	

6.5 Determinants of overt plural marking on German code-mixed nouns

The analysis of variation of plural marking on German noun insertions in Russian sentences builds on three important observations. First, Russian morpho-

phonology restricts the possibilities of using a German stem with Russian inflectional suffixes. Second, the frequency distribution of a lexeme's singular and plural forms in German can predict which of the two mixing patterns is likely in bilingual production. Third, a partial overlap between the declensional systems of German and Russian appears to be another factor determining the selection of the mixing patterns with German noun insertions in plural contexts.

6.5.1 Morphophonological restrictions on overt Russian plural markers with German nominal stems

As shown in the previous section by the example of German nouns with the vowels [ɐ] and [eː] in the stem-final position, both the morphophonological restrictions of the matrix language and the sound shape of a lexical item to be inserted determine whether the item undergoes morphological integration into the matrix language or whether it features the embedded-language morphology. In the following I will demonstrate that the Russian declensional system restricts the use of Russian plural inflections with German nouns, depending on their phonemic shape.

German nouns' base forms, i.e., the forms of the nominative singular, end in either a vowel or a consonant. A major portion of German nouns receiving Russian plural inflections has stems with consonants in the final position, such as *Geschenk-i* 'present-NOM.PL', *Netz-y* 'net-ACC.PL', and *Beispiel-ej* 'example-GEN.PL'. The data set contains 57 tokens of this type. Noun stems with vowels in the final position are much less frequent: there are only 16 instances of such stems in the sample. They fall into two groups depending on whether the stem-final vowel is stressed or unstressed. Of the 16 instances, 14 stems exhibit unaccented vowels in the final position. The most frequent noun of this type, the lexeme *Sprache* 'language', takes Russian plural inflections four times. In contrast, there are only two tokens of a German noun with an accented stem-final vowel. Both tokens are the instances of the lexeme *LKW* 'lorry', though the two occurrences differ.

Let us consider the first group of nouns, which have unstressed vowels in the stem-final position. As the initial phonemes of the Russian nominal inflections are vowels, these stems' unstressed final vowels are deleted in order to avoid hiatus: -CV + -V(C) > -C~~V~~V(C). The forms resulting from this process are listed below:

(18) *Flasche* + i > (tri) *Flasch*-i '(three) bottle-GEN.SG' (LR0125)
Grippe + -y > *Gripp*-y 'flu-NOM.PL' (LAR022411)
Konto + -y > *Kont*-y 'account-NO.PL' (LA05036)

Kunde + -ov > *Kund*-ov 'customer-GEN.PL' (LA022415)
Kunde + -am > *Kund*-am 'customer-DAT.PL' (LA05038)
Sache + -i > *Sach*-i 'thing-NOM.PL' (LB110526)
Sprache + -i > *Sprach*-i 'language.(course)-ACC.PL'[5] (LB110526, LV022410)
Sprache + -ax > *Sprach*-ax 'language.(course)-PREP.PL' (LV022410)
Sprache + -ami > *Sprach*-ami 'language.(course)-INSTR.PL' (LV022410)
Türke + -i > *Türk*-i 'Turk)-NOM.PL' (LR0316)
Zwetschge + -i > *Zwetsch*[k]-i 'plum-NOM.PL' (LR0714)
Zwetschge + ∅ > *Zwetschek* 'GEN.PL\plum[GEN.PL]' (LR0714)

The vowel sound at the end of the stems in (18) is almost always a schwa; in *Konto*, however, it is a peripheral vowel. From this we can assume that stem-final vowels are deleted regardless of their quality. With regard to the form *Zwetschek* 'GEN.PL\plum[GEN.PL]', its occurrence is by virtue of similarity with the genitive plural forms of Russian feminine and neuter nouns such as *toček* 'GEN.PL.F\point[GEN.PL.F]' and *sloveček* 'GEN.PL.N\small.word[GEN.PL.N]'. The base forms of these stems end in *-čk-* and they alternate with forms involving the unstressed epenthetic vowel [ɪ] in the genitive plural, i.e., *-čk-* [tʃk] ~ *-ček* [tʃɪk] (cf. the aforementioned nouns' base forms *točk-a* 'point-NOM.SG.F' and *slovečk-o* 'small.word-NOM.SG.N').

The sample contains one exception to stem-final deletion, namely, the morphologically integrated form *babyki* 'babies'. In producing this form, the speaker not only selects the Russian plural inflection *-i* but also attaches the consonant [k] to the German noun *Baby*. There are two possible explanations for this case: first, if the consonant is not inserted, the inflection is not perceivable acoustically, i.e., ['be:bɪi] or ['be:bɪɪ]; second, the combination *-ik-* may count as a Russian diminutive suffix (e.g., *dom* 'house' – *dom-ik* 'little house'), considering its semantic congruence with the German *Baby* 'very young child'. Note that a different speaker uses the same lexical item with another very productive diminutive suffix *-ičk-* (cf. Švedova 2005 [1980][a]: 209–210) to produce a singular form *babyčka* (HO1007). Here, just as in the aforementioned instance, the final *-i* undergoes a reanalysis and becomes part of the Russian suffix, i.e., *Baby* + *-čk-a* > *bab-ičk-a* 'baby-DIM-NOM.SG'.

Regarding inserted German items with accented vowels in the stem-final position, speakers insert a suffix element instead of deleting the stressed vowel. The noun *LKW* 'lorry' is the only noun affected:

[5] As mentioned in §5.2.1, Russian Germans use the German word *Sprache* 'language' in the plural to refer to language courses. Because of the obvious semantic change this plural noun may be analysed as an established loan.

(19) (LJ07141)
[ɛlkaˈveː] + -am (DAT.PL) > [ɛlkaˈveːʃkəm] (lorry-DIM-DAT.PL)
[ɛlkaˈveː] + -∅ (GEN.PL) > [ɛlkaˈveːʃɪk] (GEN.PL\lorry-DIM[GEN.PL])

Reanalysis is involved here again; the stressed vowel [e] is treated as part of the suffix *-ešk-*, an allomorph of the diminutive suffix *-k-* (Švedova 2005 [1980][a]: 2010). Interestingly, a similar strategy is observed in vernacular Russian: loan words, which are subsumed by the zero declensional class in standard Russian (cf. the use of the word *LKW* in 17), receive (diminutive) suffixes in order to become inflected overtly. For instance, nouns of standard Russian such as *kafé* 'cafè', *pjuré* 'purèe' and *sidí* 'CD' are inflected as *kaféška*, *pjuréška* and *sidíška* in colloquial Russian.

In sum, stems with final consonants are susceptible to integration into the Russian declensional system and easily take Russian inflections. The integration of stems with unaccented final vowels is achieved through the deletion of these vowels or the insertion of a suffix element such that the stems feature a consonant in the final position. Stems with accented final vowels are problematic because they can only be integrated into the declensional system by means of consonantal suffixal elements (see Table 6.4). Although this process is attested only with one noun in the sample, integration into the declensional system with a possibility of pluralisation by inflections appears only achievable when the stem ends in a consonant. In other words, direct use of overt Russian inflections with German nouns depends on these nouns' phonemic shape, and it is constrained with this type of stems. This restriction may explain why nouns such as *Schlittschuh* [ˈʃlɪtʃuː] 'skate' (LJ1221), *Presswehe* [ˈpʁɛsˌveːə] 'pushing contraction' (LV022408), and *CD* [tseːˈdeː] (LD0405) are used in the corpus solely with German plural markers (cf. **Schlittschuh-i*[6], **Pressweh(e)-i*, **CD-i*).

In this section, I have shown that analysis of morphophonological regularities of a matrix language may be fruitfully explored to account for the variation of plural marking on noun insertions in language mixing. However, the restriction formulated above is limited only to the three mentioned instances (*Schlittschuhe* 'skates', *Presswehen* 'pushing contractions', and *CDs*) because only few German stems end in an accented vowel. Moreover, the occurrence of these plurals in Russian sentences could also be attributed to other factors, such as their frequent

[6]Interestingly, the form *šui* [ˈʃuˑɪ] is not impossible in Russian: it corresponds to the genitive singular of the toponym *Šuja* (a city in Central Russia). Nevertheless, the underlying form of this noun is /ˈʃuj-/, not /ˈʃu-/ (which is evidenced by the derived anthroponym *Šujskij*). Likewise, the underlying form of the word *idè-i* 'ideas' is /idej-/ (see Itkin 2007: 246, for this stem alternation).

Table 6.4: Morphophonological processes allowing the pluralization of German noun stems ending on vowels.

Stem-final vowel	Morphophonological process	Tokens	Types
unaccented	deletion	12	8
	insertion of a suffix element	1	1
accented	insertion of a suffix element	2	1

use in German. For example, the word *Presswehe* 'pushing contraction' is a plural-dominant noun (*Duden online* 2013). As the distribution dominance of a noun's singular and plural forms in monolingual usage may be a more pervasive factor than the morphophonological restriction outlined above, the following section will scrutinise the inserted nouns' frequencies in the singular and in the plural. Another factor influencing the choices between German and Russian plural markers pertains to the differences and similarities between these languages' case systems.

6.5.2 Factors determining the language for plural marking: coding and modelling

As mentioned above, when analysing the variation in plural marking on German noun insertions in Russian sentences, it is not always possible to attribute the choice between German and Russian overt markers to a single factor. Rather, the use of one of the patterns should be seen as an outcome of several factors, interacting in bilingual production online. These factors include: (1) a mismatch in the frequencies of the plural and singular forms, (2) the stem-final segment of the base form (nominative singular), a consonant or a vowel, and (3) the morphological case required by the slot in which the noun is inserted. Each of these could be analysed to account for some part of the data, yet it is impossible to say how relevant a factor is in terms of the overall variation observed because their effects differ in strengths. Therefore, I perform statistical modelling in order to first disentangle the impact of each in determining overt plural marking and then to predict the outcome of the competition between them. To do this, I utilise the generalised linear mixed model.

6.5.2.1 Frequencies of the noun's singular and plural forms

To date, contact linguistics literature has not explored the role of a mismatch in the frequency distribution of a noun's singular and plural forms as a factor determining the use of the embedded-language plural marking with code-mixed nouns. Backus mentions this effect in his monograph (1996), but does not explore its potential as a determinant of plural marking on inserted nouns empirically. Against the background of the present study, I hypothesise that German nouns inserted in otherwise Russian sentences retain the embedded-language, i.e., German, plural marking if they are used in Germany more frequently as plurals than singulars. Conversely, if the frequency of German noun's plural form is similar to, or lower than, the frequency of the noun's singular form, it is more likely to receive a Russian plural marker as a result of the pressure the matrix language exerts to morphologically integrate embedded-language stems.

To test these assumptions, an analysis of the frequencies of the inserted German nouns' singular and plural forms was conducted in the deWaC corpus, which was utilised in the case studies reported in the previous chapters. Nouns with separate singular and plural forms present the clearest and prototypical case, they mark the plural by either adding a suffix or applying umlaut. For example, the singular form of the noun *Situation* occurs 228 thousand times in the corpus, whilst its plural form, marked by the inflection *-en,* occurs 48 thousand times. Somewhat less straightforward was the task of determining the frequencies of the nouns with identical forms in both numbers. Conducting separate counts of such nouns in the singular and in the plural was impossible because the corpus was not tagged for the morphological number. Automated counts of singular versus plural uses could not be performed for two groups of nouns: one is distinguished by the inflection *-(e)n*, which appears in both numbers in all morphological cases except for the nominative (and sometimes genitive) singular (patterns 7 and 8 in Table 6.2); the other group includes masculine nouns devoid of overt plural marking except in the dative plural, these stems end in *-l*, *-n* and *-r* (pattern 1 in Table 6.2). In order to obtain the number values of these ambiguous forms, all sentences in which they appear in deWaC were extracted and subjected to automatic parsing by the mate-tools pipeline (Björkelund et al. 2010). Among the lexemes whose frequencies could not be measured in deWaC automatically are *Obstbecher* 'fruit cups' and *Ballerinas*. The word *Ballerinas* refers to ballet-dancers and shoes, and its singular and plural may be used with differing frequencies depending on the meaning (while the plural may be more recurrent in the meaning 'shoes', the singular may be more frequent with reference to ballet-dancers). However, the two meanings could not be discriminated automatically in the corpus analysis. As to

Obstbecher, it was not attested in the corpus at all. In total, 151 items remained for further analysis.

The competition between the examined nouns' plural and singular forms was modeled by employing the odds (see Fahrmeir et al. 2007: 119 and Chapter 5.3.1.2). Odds is the ratio of the likelihood that an event will happen to the likelihood that an event will not happen. For the competition between plurals and singulars, this ratio was calculated in the following way:

$$\text{odds} = \frac{F_{PL}}{F_{SG}}$$

where F_{PL} is the frequency of a noun's plural and F_{PL} is the frequency of a noun's singular. The odds express the relation between the strengths of representation of a noun's plural and singular forms. In other words, plurals are characterised as having a stronger, or weaker, memory trace than their corresponding singulars. When the odds equal 1, the representations of both forms are regarded as equally strong. If the odds are larger than 1, the plural form has a stronger memory trace than the singular, and is thus more likely to become activated as a unit in production. Congruently, the value of the odds below 1 indicates a stronger representation of the singular form. Researchers studying morphological processing of plurals have referred to this mismatch as frequency dominance (Baayen et al. 1997). They distinguish between plural-dominant and singular-dominant lexemes. The frequency of plural-dominant words is higher in the plural than in the singular, whereas singular-dominant lexemes have a higher usage frequency in the singular. Dominance is similar to the odds in that it also relies on the frequency of the plural relative to the frequency of the singular.

If we assume that the odds model the competition between the representations of plurals and singulars, we can predict that in bilingual production, when the embedded-language and the matrix language plural marking compete, the embedded-language plurals will be produced if they have odds greater than 1.

Table 6.5 presents examples of the analysed lexical items' singulars and plurals, their respective frequencies and the odds of the plural given the singular. The items were selected according to the pattern of plural marking they use and their corresponding odds values, which are either higher or lower than 1. The first two items showcase nouns whose plurals rely on suffixation, the following two mark the plural by simultaneously applying suffixation and umlaut, and the last four have identical forms in both numbers. Considering the odds ratios reported here, we can assume that the plural forms *Studiengebühren* 'tuition fees' and *Bundesländer* 'federal states' are represented in the mental lexicon/grammar more strongly than their singular counterparts. The words *Situation* and

Parkplatz 'car park' demonstrate a reverse distribution: their singulars are more common and obviously more strongly entrenched in the language users' minds. The nouns *Türke* 'Turk' and *Kunde* 'customer' do not have separate singular and plural forms, they take the inflection *-n* not only in the plural but also in the non-nominative cases in the singular. We may thus hypothesise that the exemplar clusters corresponding to their inflected forms, namely, *Türken* and *Kunden*, are stronger than those linked to their base forms. However, I cannot address this issue here since there are only nine items belonging to this pattern in the bilingual corpus. Future work focusing on the cognitive representation and processing of homophonous forms in bilinguals and monolinguals is clearly needed. In order to maintain consistency, counts of these nouns in deWaC were carried out according to the general procedure: case distinctions were ignored, and a lexeme's plurals were counted separately from its singulars. That is, the base, or nominative singular, form and the inflected singular form, marking the genitive, accusative and dative singular, were taken together. The same procedure was applied to the other group of nouns distinguished by identical singular and plural forms, e.g., masculine nouns featuring stem-final /r/. These forms are exemplified by two items at the bottom of Table 6.5, i.e., *Pflaster* 'plaster' and *Ausländer* 'foreigner'.

Table 6.5: Some of the examined lexical items' singulars and plurals with their respective occurrence frequencies obtained from DeWAC and the odds of the plural given the singular. Overt plural markers are in bold.

Singular	F_{SG}	Plural	F_{PL}	Odds
Situation 'situation'	228,379	*Situation**en***	48,579	0.213
Studiengebühr 'tuition fee'	895	*Studiengebühr**en***	23,474	26.228
Parkplatz 'car park'	13,839	*Parkplätz**e***	8,230	0.595
Bundesland 'federal state'	16,857	*Bundesländ**er***	69,728	4.136
Türke(n) 'Turk'	14,432	*Türke**n***	3,811	0.264
Kunde(n) 'customer'	77,580	*Kunde**n***	132,604	1.709
Pflaster 'plaster'	3,491	*Pflaster*	1,533	0.439
Ausländer 'foreigner'	12,128	*Ausländer*	42,238	3.483

After the plural-singular ratios were calculated for the German nouns marked for the plural, the logarithm of their values was taken in order to avoid skewing in the distribution (Baayen 2008: 31). These values are given in Figure 6.1. Figure 6.1a depicts the ordered values of the odds, whilst Figure 6.1b shows its

quartiles. As can be seen in both, the data points are distributed more sparsely around the extreme values than around the median. The values on both ends of the scale present outliers, which correspond to the items *Grundkenntnis* 'basic knowledge' (log(odds) = 4.66) and *Grippe* 'flu' (log(odds) = −4.86). While the former is extremely infrequent as a singular, the latter is exceedingly rare as a plural. In order to enhance normality, the outliers were removed from the data set. The number of the discarded data points amounts to 1.3% of the sample. The distribution of the odds values without outliers is represented in Figure 6.1c, and its quarterlies, in Figure 6.1d. A comparison of Figure 6.1a with Figure 6.1c indicates that the distribution in Figure 6.1c is more centered around zero, and thus closer to normality.

Figure 6.2 shows the relationship between the plural-singular ratio on the logarithmic scale and the language of the overt plural marker. The binary variable overt plural marker is on the vertical axis and has the values of one and zero, which represent the overt plural marker language, Russian, as the matrix language, and German, as the embedded language; the values of the odds are on the horizontal axis. The line depicting the relationship between the two variables is a LOWESS (locally weighted scatterplot smoothing) curve (Cleveland & Devlin 1988). As indicated in Figure 6.2, the curve is virtually symmetrical around the logarithmic value of −0.3, which corresponds to the odds value of 0.740. In the plot we can also see that the line has two points of inflection, separating the central interval of the curve between the logarithmic values of −1.3 and 0.6. The central interval, also symmetrical around −0.3, stands for a gradual, transitional area. Interestingly, a comparison of the inflection points 0.277 and 1.822 at opposite ends of the spectrum also shows a symmetrical distribution of the lexical items in the data set: there are as many singular-dominant nouns as there are plural-dominant ones. As is evident from the shape of the curve, embedded-language nouns with embedded-language plural markers correspond to plural dominant nouns, whereas embedded-language nouns featuring matrix language overt plural markers correspond to singular-dominant nouns. As such, the higher an embedded lexical item's plural-singular ratio, or the odds, the more likely this item will appear with embedded-language plural marking in bilingual speech. Thus, embedded-language plurals appear to be accessed and selected in online speech production as units.

6.5.2.2 Stem-final segment of the insertion

In this section I analyse the stem-final segment of an inserted lexical item as a predictor of the language of its plural marker. I identify stem-final segments of

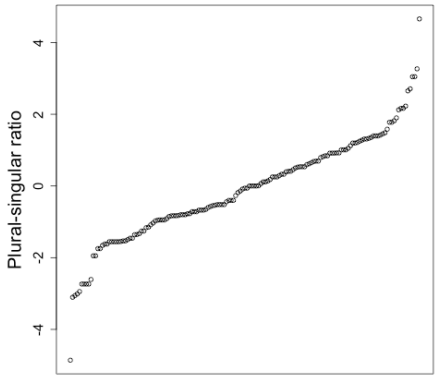

Figure 6.1a

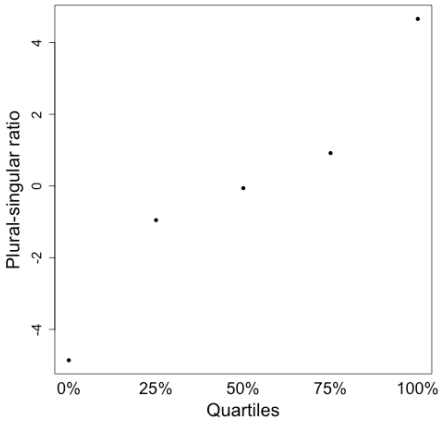

Figure 6.1b

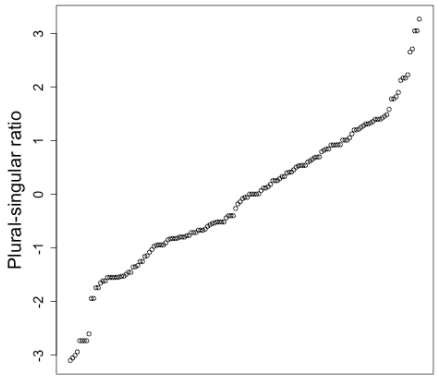

Figure 6.1c

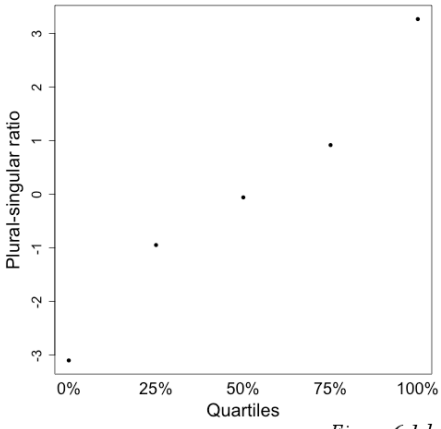

Figure 6.1d

Figure 6.1: The ordered values of the plural-singular odds on the logarithmic scale (a) and its quartiles (b); before removing the outliers (c) and after removing outliers (d).

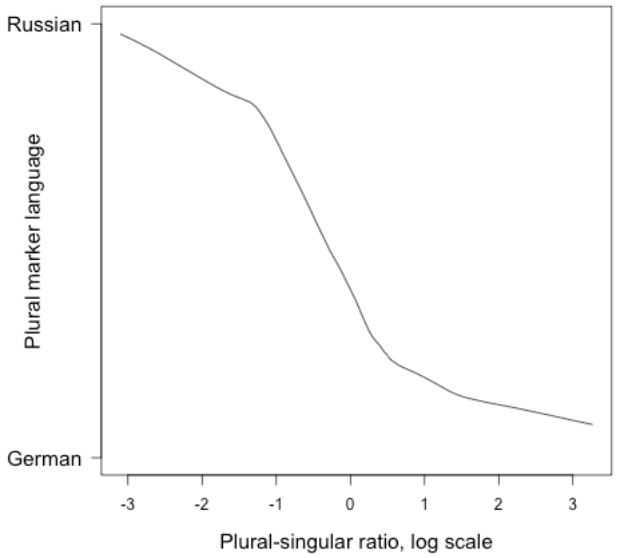

Figure 6.2: The relationship between the plural-singular ratio (on the logarithmic scale) and the language of the overt plural marker. Russian is the matrix language (ML), German is the embedded language (EL).

the nouns' base forms, i.e., the forms in the nominative singular. For example, the stem-final segment of the word *Türke* 'Turk' is a vowel. As shown in 6.5.1, several strategies are employed to enable the addition of Russian inflectional suffixes, marking the plural, to German stems ending in vowels. As mentioned above, these strategies aim at hiatus resolution at the morpheme boundary. One strategy is to delete the vowel, if it is unaccented, regardless of its quality, as in (19), another option is to insert a consonant, or a consonant cluster, which, as a rule, correspond to parts of Russian derivational suffixes. Crucially, the need to resolve the morphophonological mismatch may also result in the use of German plural markers on German stems ending in vowels. This is particularly characteristic of stems whose final vowels are accented. In contrast to the morphophonological constraint formulated above, the assumption here is that even German nominal stems with unaccented vowels in the final position may tend to occur with German plural markers in otherwise Russian sentences. In other words, if the stem-final segment of the inserted lexical item is a vowel, the chance that this item will be used with a German plural marker is high, and vice versa, if the final segment of an inserted stem is a consonant, the preference for a Russian inflection in plural contexts will be strong.

The data were coded for the presence, or absence, of a vowel in the stem-final position of inserted German nouns. In the case of the stem-final /r/, its consonantal and vocal realisations were treated separately and were respectively coded. Figure 6.3 displays the relationship between the sound at the stem end and the language of the plural marker. According to Figure 6.3, the proportion of embedded-language plural markers is skewed depending on whether stems feature final vowels. This is in line with the hypothesis that noun insertions with vowels in the stem-final position favour German plural markers, while Russian plural markers are more common on stems ending in consonants. Whether the segment in the stem-final position of a German insertion has explanatory power regarding the variation in the data will be further examined and discussed in §6.6.

6.5.2.3 The morphological case of the slot

The mismatch in the nominal systems of German and Russian could be considered an additional motivation to use Russian inflectional suffixes when inserting German lexical material into Russian sentences. The languages are nonequivalent in two respects: First, the Russian morphological cases outnumber the German cases six to four. Second, case in German is principally not marked morphologically on the stem, but rather syntactically on the determiner (see §6.3.2). However, my corpus contains no instances of insertion of fully-fledged noun phrases, nor is morphological marking by the dative suffix *-(e)n* attested on inserted German plurals. Hence German plurals remain unmarked for the morphological case. In contrast, the degree of syncretism among Russian plural inflections is sizeably lower. In order to mark the case overtly and thus in line with the general pattern of the Russian nominal system, speakers may tend to utilise Russian plural markers, especially if the case projected on the slot is a non-core case, i.e., neither the nominative nor the accusative.

The nominative and the accusative case have a special status in the Russian declensional system. In the data set, the nominative plural is generally expressed by the inflection *-i* and its phonological allomorph *-y*, with the exception of the inflection of the masculine plural *-á*, occurring in the form *bauer-á* 'peasants' in analogy with such Russian nouns as *professor-á* and *traktor-á*. The same inflections are used for the accusative plural of inanimate nouns. Note that no animate nouns are attested in the data set in the form of the accusative plural (which require inflections identical with the genitive plural). In other words, apart from one instance of the inflection *-a*, the nominative plural and the accusative plural appear both to be marked solely by the inflection *-i* and its allomorph *-y*. As such,

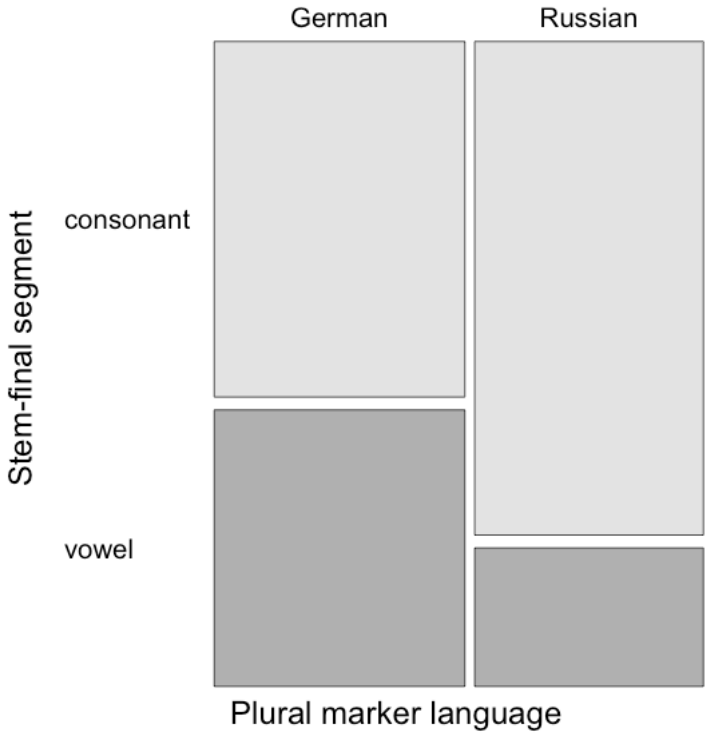

Figure 6.3: The relationship between the choice of the language for overt plural marking on German nominal insertions and the final sound of the stem. Russian is the matrix language (ML), German is the embedded language (EL).

the inflection *-i (-y)* is regarded as a prototypical plural marker of Russian nouns; the inflection *-y* also marks plural on predicative short adjectives, as in *umn-y* 'clever-PL', and the inflection *-i* marks the plural on verbs in the past tense, as in *by-l-i* 'be-PAST-PL', thus reinforcing their status as prototypical plural markers. Additional evidence for this assumption comes from language acquisition. As the most frequent plural inflection in the declensional system of Russian nouns, *-i (y)* is the first plural inflection to be acquired by Russian learning children and is the one most often generalised in the process of acquisition (Gagarina & Voeikova 2009). As previously outlined, the status of this formative is similar to the status of German plural inflections in that they are also distinguished by syncretism. We may assume that similarity between German plural markers and Russian inflections expressing the plural in the core cases facilitates insertion of German

plural forms. In other words, if the case projected on the slot is nominative or accusative, German plurals can be accommodated easily. The opposite seems to hold as well: German noun stems will take Russian plural inflections if the slots in which they are inserted require non-core cases.

These hypotheses determined the coding of the data with respect to the case. The German noun insertions examined were analysed for the morphological case projected on the slot in which they were inserted. Following the argumentation above, the items fell into two groups: one group included the items in nominative and accusative slots, and the other group was made up of the items in the non-core cases, namely, the genitive, the dative, the instrumental and the prepositional. Thus, the predictor "case of the slot" was handled as a binary variable. The relationship between the case projected on German noun insertions and the language of the plural marker is displayed in 6.4. The case of the slot is on the vertical axis, the language of the plural marker is on the horizontal axis. The data in 6.4 reveal that overall, the nominative and the accusative are more frequent than all the other cases. Additionally, the proportions of the embedded-language plural markers and the matrix language plural markers are asymmetrical in terms of the case required by the slot. The large proportion of German plurals in the nominative and accusative slots indicates that these slots, as expected, favour the insertion of German plurals. I interpret this fact by similarity in the status between the Russian inflectional suffix(es) of the nominative/accusative plural and German plural markers. Their status is similar because like German plural markers, which express the number, the Russian plural suffix *-i* (*-y*) also expresses the number in the first place, blurring the distinction between the nominative and the accusative. Finally, as anticipated, inflections of the matrix language are preferred when a non-core case is required. Given these considerations, the case projected on the slot will be included in the multi-factorial statistical analysis below as a factor determining the language of the plural marker.

6.6 Statistical model

In this section, I will investigate the factors presented thus far with regard to the extent to which they compete with, or assist, one another in determining the plural marking on German nouns inserted into a Russian matrix structure. I will also examine and assess each of the competing factors' explanatory power. To approach these issues, I will adopt the procedure employed in the foregoing chapters, which included the fitting and evaluation of a generalised linear mixed model (see Baayen 2008: 278–284). In this model, significant interactions

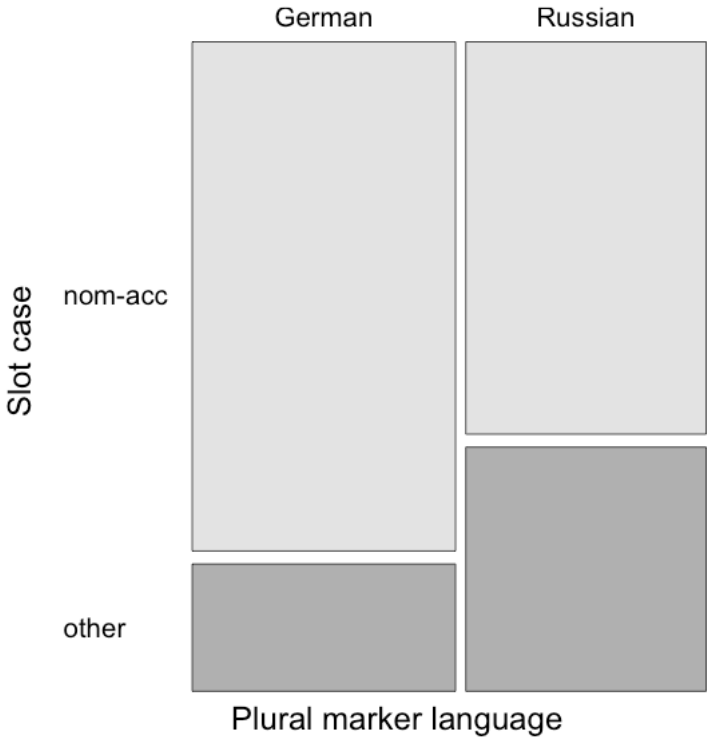

Figure 6.4: The relationship between the choice of the language for overt plural marking on German (EL) nominal insertions and the case projected by the slot. Russian is the matrix language (ML).

between the predictor variables outlined above and the binary dependent variable "the language of the plural marker on code-mixed German nouns" will provide tangible evidence for the relevance of the predictor variables.

6.6.1 Model fitting

In order to obtain a minimal adequate regression model, the common procedure was employed (Baayen 2008, Szmrecsanyi 2013), which included fitting the maximal model with the three factors outlined above as main effects: the odds, based on the distribution frequencies of the inserted noun's singulars and plurals in the embedded language; this item's stem-final segment, i.e., a vowel or a consonant; and the morphological case of the slot in which the item is inserted. Additionally, the maximal model included the interactions between these factors. The

speakers' individual differences in marking plural on German nominal insertions, i.e., the tendency to either retain the German plural marker or to add Russian plural inflections, was measured by means of the variable "speaker". The variable "speaker" was handled as a by-subject random effect. Unfortunately, adding "item" as a random effect was impossible due to the high variation in this variable: 151 German noun insertions correspond to 110 different lexical items. The model was thus run without the by-item variable. The model simplification consisted in the exclusion of factors and interactions without any significant contribution to the explanatory power of the model. Following Baayen (2008: 281), the estimation of explanatory power of the interaction terms and main effects is based on the calculation of the *C* index of concordance. In line with the reduction procedure, all interaction terms were excluded from the model (plural-singular ratio × stem-final segment, plural-singular odds ratio × case of the slot, stem-final segment × case of the slot). The final, minimal adequate model is given in Table 6.6.

Table 6.6: Predicting the language of the overt plural marker: minimal adequate generalized linear mixed model. Predicted odds ratios are for Russian (or matrix language) overt plural markers. Significance codes: *significant at $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Factor	Odds	Est.	Pr(> |z|)	
(Intercept)	2.147	0.764	0.096	
Plural-singular ratio	0.443	−0.815	0.000	***
Stem-final segment ('vowel')	0.339	−1.082	0.014	*
Slot case ('NOM or ACC')	0.383	−0.961	0.031	*
Random effect:				
Speaker				
(intercept, $N = 12$, variance $= 0.439$, $\sigma = 0.662$)				
Summary statistics:				
N		151		
% correct predictions (% baseline)		79 (81)		
C index of concordance		0.838		
Somer's Dxy		0.676		

6.6.2 Model evaluation and model discussion

After the model is fit to the data, the fit can be evaluated. The minimal adequate model reported in Table 6.6 is of high quality. The model correctly classifies 79%

of the data overall. Regarding the categorical prediction of the plural marker language, i.e., when always guessing one variant, the model correctly predicts 77% of German plural markers and 81% of Russian plural inflections. Furthermore, the fit is estimated by the measure of predictive accuracy which relates the observed realisations of the plural marker to the predicted outcome of the model (Bresnan et al. 2007). The outcome of the model is the choice of the German or Russian plural marker, denoted by 0 and 1 respectively. The accuracy measure counts any probability > 0.5 as correct for the Russian plural marker. The measure of predictive accuracy is provided in Table 6.7. The model correctly predicts the use of German plural markers with the lexical items analysed in 85% of the cases, but it has more difficulty predicting the use of Russian plural markers, performing with only 72% of correct predictions. The *C* index of concordance between the predicted probability and the observed binary outcome is 0.838, which indicates that the model has real predictive power. Performance indicator Somers' Dxy, a rank correlation coefficient between predicted probabilities and observed binary response, is 0.676, which also attests some predictive capacity of the model. As to the random effect "speaker", the variance and standard deviation of the by-subject effect reveal minimal variation. The speakers in the sample favour neither language for marking the plural on German insertions at the individual level.

Table 6.7: Model accuracy. Classification table for the minimal adequate model. (The table representation is based on Bresnan et al. 2007).

		Predicted		% correct
		0	1	
Observed	0	67	12	85
	1	20	52	72
Overall				79

The main effects in the model offer persuasive evidence in favour of the hypotheses formulated above. The signs of the regression coefficients (Estimate) in Table 6.6 reveal the direction of the adjustment to the intercept. Given this, we can conclude that the factors a noun's plural dominance, or high plural-singular ratio, vowel as the final segment of the stem, and the nominative or accusative case of the slot disfavour the use of Russian plural markers with German code-mixed nouns. Conversely, any of the non-core cases projected on the slot, a consonant as the final segment of the stem, and a noun's singular dominance, or

low plural-singular ratio, demonstrate a preference for Russian plural markers. Consider the odds listed in Table 6.6. Both the stem-final segment and the case of the slot have comparable effect sizes, that is, the odds for using a Russian plural marker decrease by approximately 66% if the stem features a vowel in the final position, and by 62% if the case of the slot is either nominative, or accusative. The strongest effect is exerted by the plural-singular ratio, or dominance: the odds for Russian overt plural markers fall by 56% at every one-unit increase in the plural-singular ratio on the logarithmic scale. This stands for the increase of this ratio by 2.7 on the linear scale. In other words, if the plural-singular ratio increases by 2.7, the odds for using a Russian plural marker falls by 56%. The overall importance of the factors is given in Figure 6.5 by plotting the decreases in the Akaike Information Criterion of the model when a factor is removed from the minimal model. As mentioned in the previous chapters, more sizable decreases in the AIC criterion of a factor stand for its greater overall importance. Thereby, when predicting the language of the overt plural marker used with 151 German noun insertions in Russian sentences, the most important factor is the plural-singular ratio. The second most important predictor is the phonemic shape of the stem, namely, the presence or absence of a vowel in the stem-final position. The case projected by the slot on the inserted lexical item is ranked last. Crucially, in slots projecting the nominative or the accusative case (coded as one level of the binary variable "case of the slot") the proportions of German and Russian plural markers on noun insertions are equally large, and the asymmetry exhibited is due to the number of instances of the other cases projected. In this case, German noun insertions do not take German plural markers as frequently as Russian ones.

This analysis shows that the frequency-based plural-singular ratio, a continuous variable, and structural factors such as the case of the slot and the the inserted lexical item's stem-final segment reliably predict the language of the overt plural marker on German lexical items inserted in Russian matrix frames, the most important predictor of this variation being the plural-singular ratio. The variation in the speakers' preferences to use either Russian or German plural markers was taken into consideration and found to be negligible.

6.7 Conclusions and discussion

This study has addressed the question of whether in a situation of marking plural on code-mixed lone nouns, these nouns retain the plural morphology of their language, i.e., the embedded language, or receive plural markers from the matrix

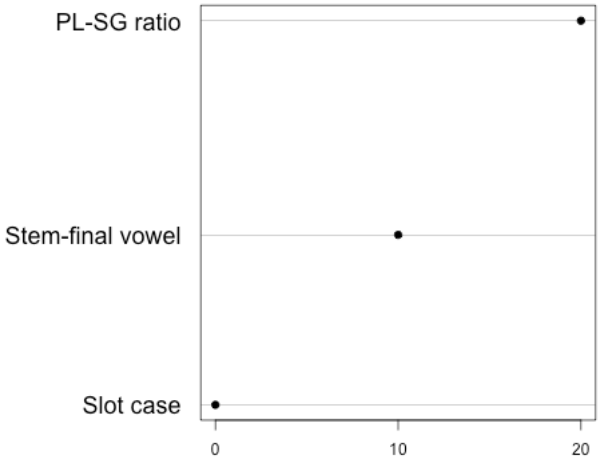

Figure 6.5: Importance of factors in model: decrease in Akaike Information Criteri-on (AIC) if factor removed. (The table representation is based on Szmrecsanyi 2013).

language: $_{R}$[$_{G}$[N $_{G}$]-PL $_{R}$] or $_{R}$[$_{G}$[N-PL $_{G}$] $_{R}$]. Prior research has largely focused on explaining the use of either double plural morphology or embedded-language plural morphology with code-mixed nouns, and attributed the use of these patterns to structural factors (Boumans 1998: 91; Myers-Scotton 2002: 91, 150), erroneous access in production (Myers-Scotton 1993: 132–136; Myers-Scotton & Jake 1995: 1000) or high frequency of some plurals in the source language (Backus 1996: 151, 1999a: 97–99, 2003: 93–100). However, these assumptions have neither been systematically examined, nor tested on monolingual material. In this study I have analysed the extent to which the choice of the language of plural marking on code-mixed nouns is determined by the frequency distribution of the inserted items' singulars and plurals in the embedded language and the structural requirements imposed by the matrix language. The research questions are addressed through corpus analyses and statistical modelling.

This investigation reveals three main findings. The first concerns a frequency effect: the frequency with which a noun's plural occurs in the embedded language, i.e., German, appears to determine the language of the plural marker on code-mixed nouns. In a situation of competition between a lexeme's singular and plural forms during online production, a plural-dominant item tends to be selected as a whole and inserted into a matrix clause together with its plural marker. If the inserted lexical item is a singular-dominant noun, only the stem,

which corresponds to the base form in these data, will be activated in production and will receive the plural marker from the matrix language, namely, Russian. This provides compelling evidence for the effect of frequency assumed earlier in (Backus 1996, 1999a, 2003).

The entrenchment of high-frequency plurals observed in the Russian-German code-mixing data concurs with similar findings from previous studies in language processing, first-language acquisition and typology. Experimental research into processing of singulars and plurals has produced substantial evidence that plural-dominant words are accessed faster than singular-dominant words (New et al. 2004, Sereno & Jongman 1997, Baayen et al. 1997). This frequency effect is interpreted as evidence for the fact that a plural-dominant word is accessed directly and has an allocated autonomous representation in the mental lexicon. In the speech of children acquiring Russian, the first plural forms are nouns that are usually used as plurals, for example, *glaz-a* 'eye-NOM.PL', *jagod-y* 'berry-NOM.PL' and *grib-y* 'mushroom-NOM.PL' (Gagarina & Voeikova 2009: 198). Interestingly, children have been found to use these forms with reference to singular entities (Ceitlin 2000: 91), which reinforces the idea that plural-dominant nouns are learned and retrieved as holistic units. Furthermore, in a corpus-based cross-linguistic study of number marking, Haspelmath & Karjus (2017) propose semantic groups of lexemes which are more frequent as plurals than singulars, these include paired body-parts, paired items, small animals, fruit/vegetables, people, and ethnic groups. It is indeed striking that many of the insertions with the embedded-language plural markers in both my data and other bilingual corpora fall into one of the suggested categories. The frequency effect observed in my data might also be approached from a semantic perspective.

The second finding of this study concerns the relevance of the phonemic shape of the inserted word to the likelihood of receiving an inflectional suffix from the matrix language. I have shown that a German lexical item with an accented vowel in the stem-final position cannot take Russian inflectional suffixes directly since the Russian declensional system depends on stems with consonants in the final position. If the final segment of an inserted German stem is an accented vowel, either a compromise strategy such as the use of epenthetic consonants is employed, or German plural forms are produced. Therefore, the Russian declensional system restricts the use of Russian inflectional suffixes with German stems depending on their phonemic shapes. In the subsequent statistical model, I reconsidered this absolute constraint as a probabilistic factor because the across-the-board analysis revealed that the presence of a vowel in the stem-final position, whether accented or not, results in favouring the use of German plural markers. This effect is grounded in processes of similarity identification between Rus-

sian inflected words and German insertions at the level of morphophonology. Instances of accommodation of German nouns when bilingual speakers reanalyse some of the stems' material as part of Russian derivational suffixes demonstrate that when similarity between German and Russian forms is lacking, speakers construct similarity online (cf. Hakimov & Backus n.d.[a]), in order to produce forms satisfying the requirements of the matrix language and being thus acceptable in the bilingual community.

The third result relates to the other structural factor: the mismatch in the systems of case marking in the plural between German and Russian. When the matrix structure projects a non-core case on the slot in which a German lexical item is inserted, the tendency is to employ Russian inflections, characterised by a fusion of the number and the case. This finding can be interpreted as a manifestation of the pressure exerted by the matrix language to produce well-formed Russian constituents. Nonetheless, many German plural nouns occur in slots requiring the nominative or accusative case. This situation is explained by the fact that the functions of German plural inflections and the Russian inflection of the nominative and accusative case *-i (y)* coincide: owing to form syncretism, both express the plural number rather than the case. This kind of equivalence results in the ease of inserting German plurals in such slots. We may also assume that the declensional systems of the two languages are perceived as incompatible in the other morphological cases because in slots projecting these cases, bilingual speakers tend to use Russian (=matrix language) plural formatives, enabling unambiguous case-marking.

The findings of this study have far-reaching consequences for models of code-mixing because they clarify the emergence of the so-called internal embedded-language islands. Firstly, the embedded-language plural markers are regarded as syntactically inactive (Myers-Scotton 2002: 92) or parts of chunks (Backus 1999a: 98). If we combine the concept of syntactic inactivity with the idea of holistic storage, we can assert that embedded-language plural markers are syntactically inactive inasmuch as they are part of the representations of plural forms. These plurals are so strongly entrenched in the mental lexicon/grammar that they are retrieved as wholes in bilingual production. If a plural is weakly entrenched, a matrix language plural marker is produced. Secondly, by handling both categorical variables – the phonetic and the morphosyntactic context – as probabilistic tendencies rather than absolute constraints, the study has provided greater understanding of non-trivial idiosyncrasies observed in bilingual language production. Thirdly, the findings demonstrate that in language mixing at the level of morphology, bilingual speakers keep track of the regularities and distribution properties of both the embedded language and the matrix language. While the

matrix language determines the morphosyntactic and phonetic context and thus influences the choice between the competing mixing patterns, the distribution properties of the embedded-language material appears to play even a greater role in predicting the examined variation. Similarity identification, which manifests itself at the level of morphophonology and the level of morphosyntax, again involves activation of the representations attributed to both languages. The results reported here are encouraging, and should be validated in further studies of other bilingual corpora.

The study presented in this chapter differs from the studies reported in the previous chapters in that it showcases how frequency-based and structural factors can be usefully investigated together as factors influencing choices between variant mixing patterns in bilingual speech. I have interpreted the effects of contextual structural factors within the framework of usage-based contact linguistics, arguing that processes of similarity detection, and even construction, are at the core of the choices which bilingual speakers make in online speech production.

7 Summary and outlook

Research into the structure of bilingual speech has traditionally focused on the distribution of elements of language A in the discourse framed by language B. Little attention has been drawn to the distributional properties of the inserted elements as determined by the usage thereof in language A, i.e., in the monolingual mode. This book has demonstrated that these elements' distributional properties in language A in terms of their co-occurrences as well as the competition among these co-occurrences decisively influence the structure of bilingual speech. Multimorphemic forms and multiword sequences which are frequent in that language tend to appear in bilingual speech preserving their integrity. A usage-based explanation of this effect refers to the fact that frequently used multimorphemic forms and multiword strings leave strong memory traces in the language user's mind. During language comprehension and production, these representations are more easily activated than the representations corresponding to those structures' component parts. Hence, holistic storage of particular linguistic structures leads to the production thereof en bloc. The analyses in this book present evidence that pieces of language regularly used together in the embedded language, which are also referred to as chunks, appear to repel switches, and conversely, elements with no, or few, frequent companions in the embedded language easily combine with the material of the matrix language in bilingual speech. Furthermore, I have shown that prediction of structural patterns of mixing, or of the choices bilingual speakers make, usually involves other factors. These include not only usage-based, or distributional, factors such as frequency and recency, but also structural factors such as overlaps and mismatches between the involved languages' structural patterns. Because under a usage-based model of language, categorisation (which is responsible for the emergence of abstract linguistic structure) is grounded in the process of similarity identification, not only frequency effects, but also similarity effects may be usefully interpreted in one framework. Thus, even the aforementioned similarities and differences in structural patterns are interpreted in terms of the speakers' linguistic experience: bilingual speakers compare, whether consciously or unconsciously, familiar, or well-entrenched, forms of one language with forms available in the other language, and nonce forms, which may lay the ground for linguistic innovations.

The studies presented in this book investigated variation in three frequently reported loci of code-mixing: the adjective-modified noun phrase, the prepositional phrase and plural marking on nouns. Each of these contexts involves the occurrence of embedded-language islands, i.e., multimorphemic or multiword strings of the embedded language in the discourse framed by the matrix language. While co-occurrence frequency, the basis of chunking, was found to facilitate the use of embedded-language islands in each of the examined structural contexts, the frequency of the words constituting the islands also determined the likelihood of their appearance in bilingual sentences, namely, the frequency of the noun in the prepositional phrase and the frequency of the adjective in the adjective-modified noun phrase.

In adjective-modified noun phrases involved in code-mixing, the frequency of the adjective was observed to negatively correlate with the tendency to produce a German embedded-language island. Whilst frequent adjectives, which express very general meanings, usually came from the more activated, matrix language, i.e. Russian, low-frequency adjectives were predominantly German. This situation was interpreted in terms of Backus's specificity continuum, according to which embedded-language lexical items with specific meanings are frequently involved in code-mixing, and the strong syntactic projection of inflected German adjectives (Auer 2005, 2007a), which are followed exclusively by German nouns in the examined data. Hence, the embedded-language islands structured as adjective-modified noun phrases corresponded either to chunks, or were triggered by German low-frequency adjectives, distinguished by specific meanings.

In the prepositional phrase, the bias to produce embedded-language islands was observed to positively correlate not only with the probabilistic factor odds, based on the frequency of co-occurrence between the preposition and the noun, but also the frequency of the noun. I interpreted the effect of the noun frequency as evidence for the tendency of high-frequency nouns to trigger their preposition companions, which are part of highly recurrent multiword sequences. Prior discursive context appeared to exert the strongest influence on the choice between German and Russian prepositions with German noun insertions. That is, if a preposition occurred in the immediately preceding discourse, this preposition was likely to appear in the examined prepositional phrase in the same language again. This priming effect was interpreted as evidence for the high accessibility of function words to bilingual speakers after their previous retrieval a moment ago during the same interaction.

As in the previous case studies, plural marking on German noun insertions in Russian sentences exhibited two patterns: the speakers either produced German plurals, i.e., internal embedded-language island, or combined German stems

with Russian inflectional suffixes, fusing number and case. The choice between a mixed constituent and an embedded-language island was found to result from the following factors: the plural-singular distribution ratio, the phonemic shape of the inserted stem and the morphological case of the slot in which the noun is inserted. According to the minimal adequate regression model fit to the data, the most important predictor of this variation was the plural-singular ratio: plural-dominant nouns, i.e. nouns that are more frequently used as plurals than as singulars, tended to retain their German plural marking in mixed sentences, whereas singular-dominant nouns received plural marking from Russian. The stem's phonemic shape was the factor of second importance: nouns with vowels in the stem-final position retained their German plural marking more often than nouns with consonants in the stem-final position. This result was regarded as a morpho-phonological restriction that the matrix language exerts on inserted noun stems. Finally, the non-equivalence of the German and Russian nominal systems, with varying degrees of syncretism, affected the preference for one of the examined patterns: German nouns in slots projecting the genitive, the dative, the instrumental, or the prepositional case received Russian inflectional suffixes more frequently than German nouns in slots requiring the nominative or the accusative. In effect, bilingual speakers could insert German plurals in slots projecting the core cases more amply than in slots requiring the non-core cases. This finding was considered to result from (i) a mismatch between German plural paradigms, virtually neutralising case distinctions, and Russian plural forms in the non-core cases, each being distinguished by a unique formative, and (ii) a similarity between the German plural paradigms, made up, with one exception, of a single form, and the Russian plural formatives in the core cases, distinguished by nominative-accusative syncretism. Hence, this similarity is of the relational type (Gentner & Markman 1997).

When we compare the various predictors identified through the application of statistical modelling, co-occurrence frequency was found to exert an effect on the variation in every case study. It was the most important predictor in the variation in plural marking of German noun insertions. In this case study as well as in the analysis of switch placement in the prepositional phrase, co-occurrence frequency was modelled in relational terms with regard to the other items entering the co-occurrence distributions. While in the study of plural-marking, the competition was only between two forms, the plural and the singular (i.e., the base), the competition between prepositions accompanying a specific noun involved more than ten items. The analysis of mixing in the adjective-modified noun phrase dispensed with a relational measure of the frequency with which a specific noun appeared with a particular adjective and utilised the observed fre-

quency of co-occurrence because some nouns are used with a virtually unlimited number of adjectives. A comparison of the impacts of co-occurrence frequency on the variation in mixing patterns in the various contexts yields a conclusion that the fewer competitors in a distribution, the more robust the effect of co-occurrence frequency.

As concerns the inter-speaker variation in the utilised data, it was found to be negligible in every case study. This result lent support to my assumption that the group of speakers represented in the corpus was homogeneous and their personal preferences in code-mixing did not significantly influence the distribution of the scrutinised patterns.

Essentially, a usage-based approach to the analysis of code-mixing, as proposed in this book, proved not only empirically robust but extremely fruitful in providing psychologically plausible explanations for code-mixing patterns, such as embedded-language islands. The reported results provide tangible evidence that patterns in code-mixing depend on (i) gradient facts of usage, (ii) similarities between the involved languages' structural patterns, and (iii) the immediate linguistic and discursive context. Additionally, my findings can be considered as corroborative evidence of the usage-based claim that language users represent multiword and multimorphemic amalgams and heavily rely on them in language production.

Although the results are encouraging, a more accurate and robust comparison of bilingual and monolingual data would be possible when large corpora of spoken language are available. In view of the applied methodology, in future it would be interesting to model recency of linguistic structures in discourse as a gradient rather than a discrete factor. An important question for future studies is to determine a more precise measure of chunk competition. For example, instead of odds it might be worth employing more complex information theoretical measures such as entropy.

In view of the findings reported in this book, future work in contact linguistics should focus on the followings directions: First, it will be necessary to test the relevance of the explanations reported for the emergence of embedded language islands to the surface structure of the matrix language. Most interesting would be an in-depth analysis of the slots which accommodate embedded-language material and the sequences preceding these slots in terms of their interruptibility and cohesion. Particularly exciting would be a study into the relationship between code-mixing, on the one hand, and syntactic constructions and lexical cohesion, on the other hand. It would also be beneficial to apply the approach laid out in the present work to analyses of alternational code-mixing and to examine

whether the results are valid for situations of congruent lexicalisation, including the case of dialect-standard mixing. Recent work by Goria (n.d.) provides positive evidence for the role of lexically specific chunks and constructions in alternational mixing as well. What seems to be even more promising is to extend the findings of this research to the languages in contact that belong to different typological types because a central issue in the future research agenda will be to investigate the structure of bilingual speech as resulting from (i) perceived and constructed similarity, (ii) patterns of usage and (iii) processing biases (see the papers in Hakimov & Backus n.d.[b]). One direction could explore the relationship between similarities in the morphological systems of the contact languages and the outcomes of language contact. In this regard, the issue of formal and relational similarity could be brought into a sharper focus.

References

Abramowicz, Łukasz. 2007. Sociolinguistics meets exemplar theory: Frequency and recency effects in (ing). *University of Pennsylvania Working Papers in Linguistics* 13(2). 27–37.

Adamou, Evangelia. 2016. *A corpus-driven approach to language contact: Endangered languages in a comparative perspective* (Language contact and bilingualism 12). Boston; Berlin: De Gruyter.

Adamou, Evangelia & Xingjia Rachel Shen. 2019. There are no language switching costs when codeswitching is frequent. *International Journal of Bilingualism* 23(1). 53–70.

Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity* (Studies in language and linguistics). London; New York: Longman.

Alegre, Maria & Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40(1). 41–61.

Altenberg, Bengt. 1990. Speech as linear composition. In Graham Caie, Kirsten Haastrup, Arnt Lykke Jakobsen, Joergen Erik Nielsen, Joergen Sevaldsen, Henrik Specht & Arne Zettersten (eds.), *Proceedings from the Fourth Nordic Conference for English Studies*, 133–134. Copenhagen: University of Copenhagen.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In Anthony Paul Cowie (ed.), *Phraseology: Theory, analysis, and applications* (Oxford Studies in Lexicography and Lexicology), 101–122. Oxford; New York: Clarendon Press.

Alvarez-Cáccamo, Celso. 1998. From "switching code" to "code-switching": Towards a reconceptualisation of communicative codes. In Peter Auer (ed.), *Code-switching in conversation: Language, interaction and identity*, 29–50. London; New York: Routledge.

Amuzu, Evershed Kwasi. 2010. *Composite codeswitching in West Africa: The case of Ewe-English codeswitching*. Saarbrücken: LAP Lambert Academic.

Amuzu, Evershed Kwasi. 2013. Bilingual serial verb constructions: A comparative study of Ewe-English and Ewe-French codeswitching. *Lingua* 137. 19–37.

Anderson, John R. 1982. Acquisition of cognitive skill. *Psychological Review* 89. 369–406.

Arnon, Inbal. 2011. Units of learning in language acquisition. In Eve V. Clark & Inbal Arnon (eds.), *Experience, variation and generalization: Learning a first language* (Trends in Language Acquisition Research 7), 167–179. Amsterdam: John Benjamins.

Arnon, Inbal & Eve V. Clark. 2011. Why *brush your teeth* is better than *teeth* — children's word production is facilitated in familiar sentence-frames. *Language Learning and Development* 7(2). 107–129.

Arnon, Inbal & Uriel Cohen Priva. 2013. More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech* 56(3). 349–371.

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1). 67–82.

Auer, Peter. 1984. *Bilingual conversation*. Amsterdam; Philadelphia, PA: John Benjamins.

Auer, Peter. 1988. A conversation analytic approach to code-switching and transfer. In Monica Heller (ed.), *Codeswitching anthropological and sociolinguistic perspectives*, 187–213. Berlin; New York: Mouton de Gruyter.

Auer, Peter (ed.). 1998. *Code-switching in conversation: Language, interaction and identity*. London: Routledge.

Auer, Peter. 1999. From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism* 3(4). 309–332.

Auer, Peter. 2005. Projection in interaction and projection in grammar. *Text – Interdisciplinary Journal for the Study of Discourse* 25(1). 7–36.

Auer, Peter. 2007a. Syntax als prozess. In Heiko Hausendorf (ed.), *Gespräch als prozess. linguistische aspekte der zeitlichkeit verbaler interaktion*, 95–124. Tübingen: Gunter Narr.

Auer, Peter. 2007b. Why are increments such elusive objects? An afterthought. *Pragmatics* 17(4). 647–658.

Auer, Peter. 2011. Code-switching/mixing. In Ruth Wodak, Barbara Johnstone & Paul Kerswill (eds.), *The SAGE handbook of sociolinguistics*, 460–478. Thousand Oaks, CA: Sage Publications.

Auer, Peter. 2014. Language mixing and language fusion: When bilingual talk becomes monolingual. In Juliane Besters-Dilger, Cynthia Dermarkar, Stefan Pfänder & Achim Rabus (eds.), *Congruence in contact-induced language change, language families, typological resemblance, and perceived similarity* (linguae & litterae 27), 294–334. Berlin, Boston, PA: De Gruyter.

Auer, Peter. 2015. Hermann Paul's Principles: Translations and reflections. In Peter Auer & Robert W. Murray (eds.), *Hermann Paul's Principles: Translations and reflections* (linguae & litterae 51), 177–208. Berlin; Boston, PA: De Gruyter.

Auer, Peter & Raihan Muhamedova. 2005. "Embedded language" and "matrix language" in insertional language mixing: Some problematic cases. *Journal of Italian Linguistics* 17(1). 35–54.

Auer, Peter & Robert W. Murray (eds.). 2015. *Hermann Paul's Principles: Translations and reflections* (linguae & litterae 51). Berlin; Boston, PA: De Gruyter.

Baayen, Harald. 1992. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991* (Yearbook of Morphology), 109–149. Dordrecht: Kluwer Academic.

Baayen, Harald & Robert Schreuder. 1999. War and peace: Morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and Language* 68(1–2). 27–32.

Baayen, Harald R. 2008. *Analyzing linguistic data.* Cambridge, England; New York: Cambridge University Press.

Baayen, Harald R. 2013. Multivariate statistics. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 337–372. Cambridge, England; New York: Cambridge University Press.

Baayen, R. Harald, Ton Dijkstra & Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37(1). 94–117.

Backus, Ad. 1992. *Patterns of language mixing. A study in Turkish-Dutch bilingualism* (Turcologica 11). Wiesbaden: Otto Harrassowitz Verlag.

Backus, Ad. 1996. *Two in one: Bilingual speech of Turkish immigrants in the Netherlands* (Studies in multilingualism 1). Tilburg, Netherlands: Tilburg University Press.

Backus, Ad. 1999a. Evidence for lexical chunks in insertional codeswitching. In Bernt Brendemoen, Elizabeth Lanza & Else Ryen (eds.), *Language encounters across time and space: Studies in language contact*, 93–109. Oslo: Novus.

Backus, Ad. 1999b. The intergenerational codeswitching continuum in an immigrant community. In Guus Extra & Ludo Th. Verhoeven (eds.), *Bilingualism and migration* (Studies on language acquisition 14), 261–279. Berlin; New York: Mouton de Gruyter.

Backus, Ad. 2001. The role of semantic specificity in insertional codeswitching: Evidence from Dutch-Turkish. In Rodolfo Jacobson (ed.), *Codeswitching worldwide II* (Trends in linguistics 126), 125–154. Berlin; New York: Mouton de Gruyter.

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

Backus, Ad. 2006. Turkish as an immigrant language. In Tej K. Bhatia & William C. Ritchie (eds.), *The handbook of bilingualism* (Blackwell handbooks in linguistics), 689–724. Malden, MA: Blackwell Publishers.

Backus, Ad. 2013. A usage-based approach to borrowability. In Eline Zenner & Gitte Kristiansen (eds.), *New perspectives on lexical borrowing: Onomasiological, methodological and phraseological innovations* (Language Contact and Bilingualism [LCB] 7), 19–39. Berlin, Boston: De Gruyter.

Backus, Ad. 2015. A usage-based approach to code-switching: The need for reconciling structure and function. In Gerald Stell & Kofi Yakpo (eds.), *Code-switching between structural and sociolinguistic perspectives* (linguae & litterae 43), 19–38. Berlin; Boston, PA: De Gruyter.

BAMF, Bundesamt für Migration und Flüchtlinge. 2020. *Migrationsbericht der Bundesregierung. Migrationsbericht 2018*. Berlin: Bundesamt für Migration und Flüchtlinge FIII – Migration und Integration: Dauerbeobachtung und Berichtsreihen Referat 22B – Statistik.

Bannard, Colin & Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science* 19(3). 241–248.

Bannard, Colin & Michael Ramscar. 2007. Reading time evidence for storage of frequent multiword sequences. In *Proceedings of Architectures and Mechanism of Language Processing Conference (AMLaP-2007)*. Turku, Finland.

Baroni, Marco & Adam Kilgariff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). Trento, Italy.* 87–90. Morristown, NJ: Association for Computational Linguistics.

Bates, Elizabeth & Brian MacWhinney. 1989. Functionalism and the competition model. In Brian MacWhinney & Elizabeth Bates (eds.), *The Crosslinguistic study of sentence processing*, 3–73. Cambridge, England; New York: Cambridge University Press.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge, England; New York: Cambridge University Press.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113(2). 1001–24.

Benson, Morton. 1960. American-Russian Speech. *American Speech* 35(3). 163–174.

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

Bentahila, Abdelâli & Eyrlis E. Davies. 1998. Codeswitching: An unequal partnership? In Rodolfo Jacobson (ed.), *Codeswitching worldwide*, 25–49. Berlin; New York: Mouton de Gruyter.

Berend, Nina. 1998. *Sprachliche Anpassung: Eine soziolinguistisch-dialektologische Untersuchung zum Russlanddeutschen* (Studien zur deutschen Sprache 14). Tübingen: G. Narr.

Berend, Nina & Claudia Maria Riehl. 2008. Russland. In Ludwig M. Eichinger, Albrecht Plewnia & Claudia Maria Riehl (eds.), *Handbuch der deutschen Sprachminderheiten in Mittel- und Osteuropa*, 17–58. Tübingen: G. Narr.

Bergmann, Pia. 2012. Articulatory reduction and assimilation in n#g sequences in complex words in German. In Philip Hoole, Lasse Bombien, Marianne Pouplier, Christine Mooshammer & Barbara Kuhnert (eds.), *Consonant clusters and structural complexity* (Interface explorations 26), 311–341. Berlin; Boston, PA: De Gruyter.

Bhatt, Rakesh Mohan. 1997. Code-switching, constraints, and optimal grammars. *Lingua* 102(4). 223–251.

Biber, Douglas & Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In Hilde Hasselgård & Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 181–190. Amsterdam; Atlanta, GA: Rodopi.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371–405.

Biber, Douglas, Stig Johanson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, England; New York: Longman.

Bien, Heidrun, R. Harald Baayen & Willem J. M. Levelt. 2011. Frequency effects in the production of Dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes* 26(4-6). 683–715.

Birdsong, David. 2009. Interpreting age effects in second language acquisition. In Judith F Kroll & A. M. B. de Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, 109–127. Oxford; New York: Oxford University Press.

Björkelund, Anders, Bernd Bohnet, Love Hafdell & Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, 33–36. Beijin.

Blankenhorn, Renate. 2003. *Pragmatische Spezifika der Kommunikation von Russlanddeutschen in Sibirien: Entlehnung von Diskursmarkern und Modifikatoren sowie Code-switching* (Berliner slawistische Arbeiten 20). Frankfurt am Main; New York: Peter Lang.

Bley-Vroman, Robert. 2002. Frequency in production, comprehension, and acquisition. *Studies in Second Language Acquisition* 24(02). 209–213.

Blom, Jan-Petter & John Joseph Gumperz. 1972. Social meaning in linguistic structure: Code-switching in Norway. In John Joseph Gumperz & Dell H. Hymes (eds.), *Directions in sociolinguistics: The ethnography of communication*, 407–434. New York: Holt, Rinehart & Winston.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind* (Topics in English Linguistics 83). Berlin; Boston, PA: De Gruyter.

Boas, Hans C. 2003. *A constructional approach to resultatives* (Stanford Monographs in Linguistics). Stanford, CA: CSLI Publications.

Bock, J. Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3). 355–387.

Bod, Rens. 2000. The storage vs. computation of three-word sentences. In *Proceedings of architectures and mechanisms in language processing 2000 (AMLap-2000)*. Leiden, The Netherlands.

Boeschoten, Hendrik & Peter Broeder. 1999. Zum Interferenzbegriff in seiner Anwendung auf die Zweisprachigkeit türkischer Immigranten. In Lars Johanson & Jochen Rehbein (eds.), *Türkisch und Deutsch im Vergleich* (Turcologica 39), 1–22. Wiesbaden: Otto Harrassowitz Verlag.

Bokamba, Eyamba G. 1989. Are there syntactic constraints on code-mixing? *World Englishes* 8(3). 277–292.

Boumans, Louis. 1998. *The syntax of codeswitching: Analysing Moroccan Arabic/-Dutch conversation* (Studies in multilingualism 12). Tilburg: Tilburg University Press.

Bourdieu, Pierre. 1991. *Language and symbolic power*. Trans. by John B. Thompson & Gino Raymond. Cambridge, England; Malden, MA: Polity Press.

Bowerman, Melissa. 1973. *Early syntactic development: A cross-linguistic study with special reference to Finnish*. Cambridge, England; New York: Cambridge University Press.

Boyd, Jeremy K. & Adele E. Goldberg. 2011. Learning what not to say: The role of statistical preeemption and categorization in "a"-adjective production. *Language* 81(1). 1–29.

Bradley, Dianne C. 1981. Lexical representation of derivational relation. In Mark Aronoff & Mary-Louise Kean (eds.), *Juncture*, 37–55. Saratoga, CA: Anma Libri.

Braine, Martin D. S. 1976. *Children's first word combinations* (Monographs of the Society for Research in Child Development 41, Serial No. 164). 1–104.

Brehmer, Bernhard. 2007. Sprechen Sie Qwelja? Formen und Folgen russisch-deutscher Zweisprachigkeit in Deutschland. In Tanja Anstatt (ed.), *Mehrsprachigkeit bei Kindern und Erwachsenen: Erwerb, Formen, Förderung*, 163–185. Tübingen: Attempto-Verlag.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald R. Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Edita-KNAW-Royal Netherlands Academy of Arts & Sciences.

Budzhak-Jones, Svitlana. 1998. Against word-internal codeswitching: Evidence from Ukrainian-English bilingualism. *International Journal of Bilingualism* 2(2). 161–182.

Bullock, Barbara E, Jacqueline Serigos & Almeida Jacqueline Toribio. N.d. Exploring a loan translation and its consequences in an oral bilingual corpus. *Journal of Language Contact* 13. 612–635.

Bullock, Barbara E. & Almeida Jacqueline Toribio (eds.). 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge, England; New York: Cambridge University Press.

Burani, Cristina & Alfonso Caramazza. 1987. Representation and processing of derived words. *Language and Cognitive Processes* 2(3-4). 217–227.

Bybee, Joan & Clay Beckner. 2015. Language use, cognitive processes and linguistic change. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 503–518. New York: Routledge.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam; Philadelphia, PA: John Benjamins.

Bybee, Joan L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 65–85. Stanford, CA: CSLI Publications, Center for the Study of Language & Information.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge, England; New York: Cambridge University Press.

Bybee, Joan L. 2002a. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram F. Malle (eds.), *The evolution of language out of pre-language*, 107–132. Amsterdam; Philadelphia, PA: John Benjamins.

Bybee, Joan L. 2002b. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14. 261–290.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford; New York: Oxford University Press.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge, England: Cambridge University Press.

Bybee, Joan L. 2013. Usage-based theory and exemplar representations. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 49–69. Oxford; New York: Oxford University Press.

Bybee, Joan L. & Clay Beckner. 2009. Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 827–855. Oxford; New York: Oxford University Press.

Bybee, Joan L. & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2). 323–55.

Bybee, Joan L. & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in American English. *Linguistics* 37(4). 575–596.

Cantone, Katja Francesca & Jeff MacSwan. 2009. Adjectives and word order: A focus on Italian-German codeswitching. In Ludmila Isurin, Donald Winford & Kees De Bot (eds.), *Multidisciplinary approaches to code switching*, vol. 41 (Studies in bilingualism), 243–277. Philadelphia, PA: John Benjamins Pub. Company.

Caramazza, Alfonso, Alessandro Laudanna & Cristina Romani. 1988. Lexical access and inflectional morphology. *Cognition* 28(3). 297–332.

Ceitlin, Stella N. 2000. *Jazyk i rebënok: Lingvistika detskoj reči [Language and child: Linguistics of child speech]*. Moscow: VLADOS.

Chang, Brian Hok Shing. 2009. Code-switching with typologically distinct languages. In Barbara E Bullock & Almeida Jacqueline Toribio (eds.), *The Cambridge handbook of linguistic code-switching*, 182–198. Cambridge, England; New York: Cambridge University Press.

Christiansen, Morten H. & Nick Chater. 2016. *Creating language: integrating evolution, acquisition, and processing*. Cambridge, MA: The MIT Press.

Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Clark, Eve V. 2009. *First language acquisition*. Cambridge, England; New York: Cambridge University Press.

Clark, Ruth. 1974. Performing without competence. *Journal of Child Language* 1(01). 1–10.

Cleveland, William S. & Susan J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83(403). 596–610.

Clyne, Michael. 1987. Constraints on code switching: How universal are they? *Linguistics* 25(4). 739–764.

Clyne, Michael G. 2003. *Dynamics of language contact: English and immigrant languages* (Cambridge approaches to language contact). Cambridge, England; New York: Cambridge University Press.

Colé, Pascal, Cécile Beauvillain & Juan Segui. 1989. On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language* 28(1). 1–13.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2003. Types of typology, illustrated from gender systems. In Frans Plank (ed.), *Eurotyp: Noun phrase structure in the languages of Europe: 7*, 289–334. Berlin; New York: Mouton de Gruyter.

Corrigan, Roberta, Edith A. Moravcsik, Hamid Ouali & Kathleen M. Wheatley (eds.). 2009. *Formulaic language* (Typological studies in language 82-83). Amsterdam; Philadelphia, PA: John Benjamins.

Croft, William. 2000. *Explaining language change: An evolutionary approach* (Longman linguistics library). Harlow, England; New York: Longman.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford; New York: Oxford University Press.

Dąbrowska, Ewa. 2004. Rules or schemas? Evidence from Polish. *Language and Cognitive Processes* 19(2). 225–271.

Dąbrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3). 437–474.

Dąbrowska, Ewa. 1997. The LAD goes to school: A cautionary tale for nativists. *Linguistics* 35(4). 735–766.

Davies, Mark. 2004–. *BYU-BNC*. Based on the British National Corpus from Oxford University Press. http://corpus.byu.edu/bnc/ (9 October, 2013).

Davies, Mark. 2008–. *The Corpus of Contemporary American English*. http://corpus.byu.edu/coca/ (9 October, 2013).

de Tinguy, Anne. 2003. Ethnic migration of the 1990s from and to the Successor States of the Former Soviet Union: 'Repatriation' or privileged migration? In Rainer Münz & Rainer Ohliger (eds.), *Diasporas and ethnic migrants: Germany, Israel, and post-Soviet successor states in comparative perspective*, 112–127. London; Portland, OR: Frank Cass.

Deuchar, Margaret. 2005. Congruence and Welsh-English code-switching. *Bilingualism: Language and Cognition* 8(03). 255–269.

Diessel, Holger. 2011. Review article of *Language, usage and cognition* by Joan Bybee. *Language* 87. 830–844.

Diessel, Holger. 2016. Frequency and lexical specificity. A critical review. In Heike Behrens & Stefan Pfänder (eds.), *Experience counts: Frequency effects in language* (linguae & litterae 54), 209–237. Berlin; Boston, PA: de Gruyter.

Dietz, Barbara & Peter Hilkes. 1993. *Rußlanddeutsche: Unbekannte im Osten. Geschichte, Situation, Zukunftsperspektiven.* 2., durchges. Aufl. (Geschichte und Staat 292). München: Olzog.

Dietz, Barbara & Heike Roll. 1998. *Jugendliche Aussiedler — Porträt einer Zuwanderergeneration.* Frankfurt am Main, New York: Campus Verlag.

Dijkstra, Ton, J. B. Walter & Jonathan Grainger. 1998. Simulating cross-language competition with the bilingual interactive activation model. *Psychologica Belgica* 38(3-4). 177–196.

Dressler, Wolfgang U. 2003. Degrees of grammatical productivity in inflectional morphology. *Rivista di Linguistica* 15(01). 31–62.

Du Bois, John W. 2014. Motivating competitions. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 263–281. Oxford, New York: Oxford University Press.

Duden online. 2013. Bibliographisches Institut GmbH. http://www.duden.de/node/764273/revisions/1149030/view (16 May, 2013).

Dürscheid, Christa. 2002. „Polemik satt und Wahlkampf pur" – Das postnominale Adjektiv im Deutschen. *Zeitschrift für Sprachwissenschaft* 21(1). 57–81.

Eckert, Penelope. 2000. *Linguistic variation as social practice: The linguistic construction of identity in Belten High* (Language in society 27). Malden, MA: Blackwell Publishers.

Ehret, Katharina, Christoph Wolk & Benedikt Szmrecsanyi. 2014. Quirky quadratures: On rhythm and weight as constraints on genitive variation in an unconventional data set. *English Language and Linguistics* 18(02). 263–303.

Eisenberg, Peter. 1999. *Grundriss der deutschen Grammatik.* Vol. 2: Der Satz. Stuttgart; Weimar: Metzler.

Eisenberg, Peter. 2006. *Gundriss der deutschen Grammatik.* Vol. 1: Das Wort. Stuttgart: J.B. Metzler.

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking and point of order. *Studies in Second Language Acquisition* 18. 91–126.

Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008a. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3). 375–396.

Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008b. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3). 375–396.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse* 20(1). 29–62.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart: University of Stuttgart. (Dissertation).

Fabricius-Hansen, Cathrine, Peter Gallmann, Peter Eisenberg, Reinhard Fiehler & Jörg Peters. 2009. *Duden 04. Die Grammatik: Unentbehrlich für richtiges Deutsch*. Dudenredaktion (ed.). 8., überarbeitete Auflage. Mannheim; Wien; Zürich: Dudenverlag.

Fahrmeir, Ludwig, Rita Künestler, Iris Pigeot & Gerhard Tutz. 2007. *Statistik: Der Weg zur Datenanalyse*. 6. Aufl. Berlin; Heidelberg: Springer.

Fano, Robert M. 1961. *Transmission of information: A statistical theory of communications*. New York: Wiley & MIT Press.

Field, Fredric W. 2002. *Linguistic borrowing in bilingual contexts* (Studies in language companion series Vol. 62). Amsterdam; Philadelphia: John Benjamins.

Flämig, Walter. 1991. *Grammatik des Deutschen: Einführung in Struktur- und wirkungszusammenhänge*. Berlin: Akademie Verlag.

Forbach, Gary B., Robert F. Stanners & Larry Hochhaus. 1974. Repetition and practice effects in a lexical decision task. *Memory & Cognition* 2(2). 337–339.

Frauenfelder, Uli H. & Robert Schreuder. 1992. Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991* (Yearbook of Morphology), 165–183. Dordrecht: Kluwer Academic.

Gagarina, Natalia & Maria D. Voeikova. 2009. Acquisition of case and number in Russian. In Ursula Stephany & Maria D. Voeikova (eds.), *Development of nominal inflection in first language acquisition: A cross-linguistic perspective* (Studies on language acquisition 30), 179–217. Berlin; New York: Mouton de Gruyter.

Gal, Susan. 1979. *Language shift: Social determinants of linguistic change in bilingual Austria*. New York: Academic Press.

Gardner-Chloros, Penelope. 1991. *Language selection and switching in strasbourg* (Oxford studies in language contact). Oxford; New York: Clarendon Press; Oxford University Press.

Gardner-Chloros, Penelope. 2009. *Code-switching*. Cambridge, England; New York: Cambridge University Press.

Gasimov, Zaur (ed.). 2012a. *Kampf um Wort und Schrift: Russifizierung in Osteuropa im 19.-20. Jahrhundert* (Veröffentlichungen des Instituts für Europäische Geschichte Mainz 90). Göttingen: Vandenhoeck & Ruprecht.

Gasimov, Zaur. 2012b. Zum Phänomen der Russifizierung. Einige Überlegungen. In Zaur Gasimov (ed.), *Kampf um Wort und Schrift: Russifizierung in Osteuropa im 19.-20. Jahrhundert* (Veröffentlichungen des Instituts für Europäische Geschichte Mainz 90), 9–26. Göttingen: Vandenhoeck & Ruprecht.

Gentner, Dedre & Arthur B. Markman. 1997. Structure mapping in analogy and similarity. *American Psychologist* 52. 45–56.

Giles, Howard. 1980. Accommodation theory: Some new directions. *York Papers in Linguistics* 9. 105–136.

Giraudo, Hélène & Jonathan Grainger. 2003. A supralexical model for French derivational morphology. In Egbert M. H. Assink & Dominiek Sandra (eds.), *Reading complex words: Cross-language studies*, 139–157. New York: Kluwer Academic.

Goldbach, Alexandra. 2005. *Deutsch-russischer Sprachkontakt: Deutsche Transferenzen und Code-switching in der Rede Russischsprachiger in Berlin* (Berliner slawistische Arbeiten 26). Frankfurt am Main; New York: Peter Lang.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure* (Cognitive theory of language and culture). Chicago: University of Chicago Press.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Science* 7. 219–224.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford; New York: Oxford University Press.

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3). 289–316.

Goria, Eugenio. 2018. *Inglese e spagnolo a Gibilterra. Le dinamiche del discorso bilingue*. Bologna: Caissa Italia.

Goria, Eugenio. 2021. The road to fusion: the evolution of bilingual speech across three generations of speakers in gibraltar. *International Journal of Bilingualism* 25(2). 384–400.

Goria, Eugenio. N.d. Complex items and units in extra-sentential code switching: Spanish and English in Gibraltar. *Journal of Language Contact* 13. 540–574.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony Paul Cowie (ed.), *Phraseology: Theory, analysis, and applications* (Oxford Studies in Lexicography and Lexicology), 145–160. Oxford; New York: Clarendon Press.

Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier & Daniel Jurafsky. 1999. Effects of collocational strength and contextual predictability in lexical production. *Proceedings of the Chicago Linguistic Society* 35. 151–166.

Grenoble, Lenore A. 2003. *Language policy in the Soviet Union* (Language policy 3). Dordrecht; Boston, PA: Kluwer Academic Publishers.

Gries, Stefan. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.

Gries, Stefan Th. 2010. Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora. In *Proceedings of corpus linguistics*, 1–14. Liverpool: University of Liverpool.

Gries, Stefan Th. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520–548.

Gries, Stefan Thomas. 2009. *Statistics for linguistics with R: A practical introduction*. Berlin: Mouton de Gruyter.

Grillborzer, Christine & Roland Meyer. 2008-2009. *ReBiSlav: Das Regensburger Korpus slavisch-deutscher Bilingualer*. https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/netzwerke/regensburger-korpora/index.html (13 October, 2020).

Grosjean, François. 1985. The bilingual as a competent but specific speaker-hearer. *Journal of Multilingual and Multicultural Development* 6(6). 467–477.

Gullberg, Mariane, Peter Indefrey & Pieter Muysken. 2009. Research techniques for the study of code-switching. In Barbara E Bullock & Almeida Jacqueline Toribio (eds.), *The Cambridge handbook of linguistic code-switching*, 21–39. Cambridge, England; New York: Cambridge University Press.

Haiman, John. 2014. Six competing motives for repetition. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing motivations in grammar and usage*, 246–260. Oxford; New York: Oxford University Press.

Hakimov, Nikolay. 2016a. Effects of frequency and word repetition on switch-placement. In Monika Reif & Justyna A. Robinson (eds.), *Cognitive perspectives on bilingualism*, vol. 17 (Trends in applied linguistics), 91–125. Boston, PA; Berlin: de Gruyter.

Hakimov, Nikolay. 2016b. Explaining variation in plural marking of German noun insertions in Russian sentences. In Heike Behrens & Stefan Pfänder (eds.), *Experience counts: Frequency effects in language* (linguae & litterae 54), 21–60. Berlin; Boston, PA: de Gruyter.

Hakimov, Nikolay. 2017. Ein gebrauchsbasierter Ansatz zur Analyse von Code-Mixing. *Zeitschrift für Dialektologie und Linguistik* 84(2). 308–335.

Hakimov, Nikolay. N.d. Lexical frequency and frequency of co-occurrence predict the use of embedded-language islands in bilingual speech: Adjective-modified nominal constituents in Russian-German code-mixing. *Journal of Language Contact* 13. 501–539.

Hakimov, Nikolay & Ad Backus. N.d.(a). Usage-based contact linguistics: Effects of frequency and similarity in language contact. *Journal of Language Contact* 13. 459–481.

Hakimov, Nikolay & Ad Backus (eds.). N.d.(b). *Usage-based contact linguistics: Effects of frequency and similarity in language contact. Special issue of the Journal of Language Contact.*

Halliday, M. A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Halmari, Helena. 1997. *Government and codeswitching: Explaining American Finnish* (Studies in bilingualism 12). Amsterdam; Philadelphia: Benjamins.

Haspelmath, Martin & Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics* 55(6). 1213–1235.

Haspelmath, Martin & Andrea D. Sims. 2010. *Understanding morphology*. 2nd ed (Understanding language series). London; New York: Routledge.

Hasselmo, Nils. 1972. Code-switching as ordered selection. In Evelyn Scherabon Firchow, Kaaren Grimstad, Nils Hasselmo & Wayne A. O'Neill (eds.), *Studies for Einar Haugen*, 261–280. The Hague, Paris: Mouton.

Haugen, Einar Ingvald. 1953a. *The Norwegian language in America: A study in bilingual behavior*. Vol. 1, The Bilingual Community. Philadelphia, PA: University of Pennsylvania Press.

Haugen, Einar Ingvald. 1953b. *The Norwegian language in America: A study in bilingual behavior*. Vol. 2, The American Dialects of Norwegian. Philadelphia, PA: University of Pennsylvania Press.

Haugen, Einar Ingvald. 1974 [1956]. *Bilingualism in the americas: a bibliography and research guide*. Vol. 26 (Publication of the American Dialect Society). University, AL: University of Alabama Press.

Haust, Delia. 1995. *Codeswitching in gambia: Eine soziolinguistische Untersuchung von Mandinka, Wolof und Englisch in Kontakt.* Köln: Rüdiger Köppe Verlag.

Hay, Jennifer. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6). 1041–1070.

Helbig, Gerhard & Joachim Buscha. 2001. *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Berlin; New York: Langenscheidt.

Heleniak, Timothy. 2003. The end of an empire: Migration and the changing nationality composition of the Soviet successor states. In Rainer Münz & Rainer Ohliger (eds.), *Diasporas and ethnic migrants: Germany, Israel, and post-Soviet successor states in comparative perspective*, 131–154. London; Portland, OR: Frank Cass.

Hentschel, Elke & Harald Weydt. 2003. *Handbuch der deutschen Grammatik*. Berlin: Walter de Gruyter.

Herzfeld, Anita. 1980. Creoles and standard languages: contact and conflict. In Peter H. Nelde (ed.), *Sprachkontakt und Sprachkonflikt* (Zeitschrift für Dialektologie und Linguistik – Beihefte (ZDL-B) 32), 83–90. Stuttgart: Franz Steiner Verlag.

Herzfeld, Anita. 1983. The creoles of Costa Rica and Panama. In John A. Holm & Geneviève Escure (eds.), *Central American English*, 131–156. Heidelberg: Groos.

Heylen, Kris & Dirk De Hertog. 2014. Automatic term extraction. In Hendrik J. Kockaert & Frieda Steurs (eds.), *Handbook of terminology*, 199–219. Amsterdam; Philadelphia, PA: John Benjamins.

Hlavac, Jim. 2003. *Second-generation speech: Lexicon, code-switching and morpho-syntax of Croatian-English bilinguals*. Bern; Oxford: Peter Lang.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London; New York: Routledge.

Hoi Ying Chen, Katherine. 2015. Styling bilinguals: analyzing structurally distinctive code-switching styles in Hong Kong. In Gerald Stell & Kofi Yakpo (eds.), *Code-switching between structural and sociolinguistic perspectives* (linguae & litterae 43), 163–183. Berlin, Boston, PA: De Gruyter.

Itkin, Il'ja. 2007. *Russkaja morfonologija [Russian morphophonology]*. Moscow: Gnozis.

Ivanova, Victoria & Irina Tivyaeva. 2015. Teaching foreign languages in Soviet and present-day Russia. *Zbornik Instituta za pedagoška istraživanja* 47(2). 305–324.

Jäger, Gerhard & Anette Rosenbach. 2008. Priming and unidirectional language change. *Theoretical Linguistics* 34(2). 85–113.

Jakobson, Roman, C. Gunnar M. Fant & Morris Halle. 1976 [1952]. *Preliminaries to speech analysis: The distinctive features and their correlates*. 11th ed. Cambridge, MA: The MIT Press.

Janssen, Niels & Horacio A. Barber. 2012. Phrase frequency effects in language production. *PLoS ONE* 7(3). e33202.

Janssen, Niels, Yanchao Bi & Alfonso Caramazza. 2008. A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes* 23(7-8). 1191–1223.

Jeschiniak, Niels & Willem J. M. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20. 824–843.

Johanson, Lars. 1993. Code-copying in Immigrant Turkish. In Guus Extra & Ludo Th Verhoeven (eds.), *Immigrant languages in Europe*, 197–221. Clevedon, England; Bristol, PA; Adelaide: Multilingual Matters.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & John W. Mullennix (eds.), *Talker variability in speech processing*, 145–166. San Diego: Academic Press.

Jolsvai, Hajnal, Stewart McCauley & Morten H. Christiansen. 2013. Meaning overrides frequency in idiomatic and compositional multiword chunks. In Markus Knauff, Michael Pauen, Natalie Sebanz & Ipke Wachsmuth (eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 692–697. Austin, TX: Cognitive Science Society.

Joshi, Aravind. 1985. Processing of sentences with intrasentential code-switching. In David R. Dowty, Lauri Karttunen & Arnold M. Zwicky (eds.), *Natural language parsing*, 190–205. Cambridge, England: Cambridge University Press.

Jurafsky, Dan. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 39–95. Cambridge, MA: MIT Press.

Jurafsky, Daniel, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological studies in language 45), 229–254. Amsterdam; Philadelphia, PA: John Benjamins.

Kapatsinski, Vsevolod. 2005. Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Berkeley Linguistic Society* 30. 481–492.

Kapatsinski, Vsevolod. 2010. Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech* 53(1). 71–105.

Kapatsinski, Vsevolod & Joshua Radicke. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. In Roberta Corrigan, Edith A. Moravcsik, Kathleen M. Wheatley & Hamid Ouali (eds.), *Formulaic language* (Typological studies in language 82-83), 499–520. Philadelphia, PA: John Benjamins.

Kassin, Saul, Steven Fein & Rose Markus Hazel. 2012. *Social psychology*. 9th edn. Belmont, CA: Cengage Learning.

Keune, Karen, Mirjam Ernestus, Roeland van Hout & R. Harald Baayen. 2005. Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory* 1(2). 183–223.

Kohler, Klaus J. 1995. *Einführung in die Phonetik des Deutschen*. 2., neubearb. Aufl (Grundlagen der Germanistik 20). Berlin: E. Schmidt.

Kootstra, Gerrit Jan, Janet G. Van Hell & Ton Dijkstra. 2012. Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition* 15(4). 1–23.

Kootstra, Gerrit Jan, Janet G. van Hell & Ton Dijkstra. 2010. Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language* 63(2). 210–231.

Kroll, Judith F. & Erika Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33(2). 149–174.

Kruszewski, Mikołaj. 1995. *Writings in general linguistics: On Sound Alternation (1881) and Outline of Linguistic Science (1883).* E. F. Konrad Koerner (ed.). Amsterdam; Philadelphia: John Benjamins.

Labov, William. 1972. *Sociolinguistic patterns* (Conduct and communication series). Philadelphia: University of Pennsylvania Press.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites.* Vol. 1. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar: Descriptive application.* Vol. 2. Stanford, CA: Stanford University Press.

Langacker, Ronald W. 2000. A dynamic usage-based model. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 1–63. Stanford, CA: CSLI Publications, Center for the Study of Language & Information.

Lantto, Hanna. 2015. Conventionalized code-switching: Entrenched semantic-pragmatic patterns of a bilingual Basque-Spanish speech style. *International Journal of Bilingualism* 19(6). 753–768.

Lanza, Elizabeth. 2004. *Language mixing in infant bilingualism: A sociolinguistic perspective.* Oxford; New York: Oxford University Press.

Lapteva, Olga Alekseevna. 1976. *Russkij razgovornyj sintaksis [Russian spoken syntax].* Moscow: Nauka.

Larson, Richard K. 1985. Bare-NP adverbs. *Linguistic inquiry* 16(4). 595–621.

Law, Danny. 2014. *Language contact, inherited similarity and social difference: The story of linguistic interaction in the Maya Lowlands* (Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory 328). Amsterdam; Philadelphia: John Benjamins.

Le Page, Robert Brock & Andrée Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity.* Cambridge, England; New York; Melbourne: Cambridge University Press.

Lederer, Harald W. 1997. *Migration und Integration in Zahlen. Ein Handbuch im Auftrag der Beauftragten der Bundesregierung für Ausländerfragen.* Bamberg: europäisches forum für migrationsstudien.

Lehtinen, Meri Kaisu Tuulikki. 1966. An analysis of a Finnish-English bilingual corpus. Indiana University, Bloomington.

Levelt, Willem J. M & Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology* 14(1). 78–106.

Levelt, Willem J. M. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levkovych, Nataliya. 2012. *Po-russki in Deutschland: Russisch und Deutsch als Konkurrenten in der Kommunikation mehrsprachiger Gruppen von Personen mit postsowjetischem Hintergrund in Deutschland* (Diversitas linguarum 33). Bochum: Brockmeyer.

Libben, Maya R. & Debra A. Titone. 2008. The multidetermined nature of idiom processing. *Memory and Cognition* 36(6). 1103–1121.

Lieven, Elena. 2010. Input and first language acquisition: Evaluating the role of frequency. *Lingua* 120(11). 2546–2556.

Lieven, Elena, Heike Behrens, Jennifer Speares & Michael Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language* 30(02). 333–370.

Lieven, Elena, Dorothé Salomo & Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics* 20(3). 481–507.

Lieven, Elena V. M., Julian M. Pine & Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language* 24(1). 187–219.

Loebell, Helga & Kathryn Bock. 2003. Structural priming across languages. *Linguistics* 51(5). 791–824.

Lorenz, David. 2014. *Contractions of English semi-modals: The emancipating effect of frequency* (New ideas in human interaction (NIHIN)). Freiburg: Freiburg University Library.

MacDonald, Maryellen C. 1993. The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language* 32. 692–715.

MacSwan, Jeff. 2000. The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: Language and Cognition* 3(1). 37–54.

MacWhinney, Brian. 1997. Second language acquisition and the Competition Model. In Judith F Kroll & Annette M. B. de Groot (eds.), *Tutorials in Bilingualism*, 113–142. Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, Brian. 2005. Extending the Competition Model. *International Journal of Bilingualism* 9(1). 69–84.

MacWhinney, Brian. 2014. Conclusions: Competition across time. In Brian MacWhinney, Andrej Malchukov & Edith Moravcsik (eds.), *Competing moti-*

vations in grammar and usage, 364–386. Oxford; New York: Oxford University Press.

MacWhinney, Brian, Andrej Malchukov & Edith Moravcsik (eds.). 2014. *Competing motivations in grammar and usage*. Oxford; New York: Oxford University Press.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marten, Heiko F., Michael Rießler, Janne Saarikivi & Reetta Toivanen (eds.). 2015. *Cultural and linguistic minorities in the Russian Federation and the European Union* (Multilingual Education 13). Cham: Springer International Publishing.

Maschler, Yael. 1998. On the transition from code-switching to a mixed code. In Peter Auer (ed.), *Code-switching in conversation: Language, interaction and identity*, 125–149. London; New York: Routledge.

Matras, Yaron. 2009. *Language contact* (Cambridge textbooks in linguistics). Cambridge, England; New York: Cambridge University Press.

Matusevych, Yevgen, Ad Backus & Martin W. C. Reynaert. 2013. Do we teach the real language?: An analysis of patterns in textbooks of Russian as a foreign language. *Dutch Journal of Applied Linguistics* 2(2). 224–241.

McCauley, Stewart & Morten H. Christiansen. 2015. Individual differences in chunking ability and predict on-line sentence processing. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings & P. P. Maglio (eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 1553–1558. Austin, TX: Cognitive Science Society.

McDonald, Scott A. & Richard C. Shillcock. 2003. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43(16). 1735–1751.

Meechan, Marjory & Shana Poplack. 1995. Orphan categories in bilingual discourse: Adjectivization strategies in Wolof-French and Fongbe-French. *Language Variation and Change* 7(2). 169–194.

Melinger, Alissa & Christian Dobel. 2005. Lexically-driven syntactic priming. *Cognition* 98(1). B11–B20.

Melton, Arthur W. 1963. Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior* 2(1). 1–21.

Meng, Katharina. 2001. *Russlanddeutsche Sprachbiografien: Untersuchungen zur sprachlichen Integration von Aussiedlerfamilien*. Tübingen: Narr.

Meng, Katharina & Ekaterina Protassova. 2017. Young Russian-German adults 20 years after their repatriation to Germany. In Ludmila Isurin & Claudia Maria Riehl (eds.), *Integration, identity and language maintenance in young immi-*

grants: Russian Germans or German Russians (IMPACT: Studies in Language and Sociology), 159–196. Amsterdam; Philadelphia, PA: Benjamins.

Meunier, Fanny & Juan Segui. 1999. Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language* 41(3). 327–344.

Meyer, David E. & Roger W. Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology* 90(2). 227–234.

Migge, Bettina & Isabelle Léglise. 2013. *Exploring language in a multilingual context: Variation, interaction and ideology in language documentation*. Cambridge, England; New York: Cambridge University Press.

Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2). 81–97.

Miller, Jim. E. & Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford; New York: Oxford University Press.

Milroy, Lesley. 1980. *Language and social networks*. Oxford; New York: Blackwell.

Moravcsik, Edith. 1978. Language contact. In Joseph Harold Greenberg (ed.), *Universals of human language*, 93–122. Stanford, CA: Stanford University Press.

Morton, John. 1969. Interaction of information in word recognition. *Psychological Review* 76(2). 165–178.

Muhamedowa, Raihan. 2006. *Untersuchung zum kasachisch-russischen Code-mixing (mit Ausblicken auf den uigurisch-russischen Sprachkontakt)* (LINCOM Studies in Language Typology 12). München: Lincom Europa.

Mukhina, Irina. 2007. *The Germans of the Soviet Union*. Abingdon, England; New York: Routledge.

Münz, Rainer. 2003. Ethnic Germans in Central and Eastern Europe and their return to Germany. In Rainer Münz & Rainer Ohliger (eds.), *Diasporas and ethnic migrants: Germany, Israel, and post-Soviet successor states in comparative perspective*, 261–271. London; Portland, OR: Frank Cass.

Mürkhein, Vera V. 1970. Nabljudenija nad sistemoj sklonenija imën suščestvitel'nyx v odnom iz russkix govorov Èstonskoj SSR [observations concerning the noun declension system in one of the Russian dialects in Estonian SSR]. In *Trudy pribaltijskoj dialektologičeskoj konferencii 1968 [proceedings of Baltic dialectology conference 1968]*, 108–117. Tartu: University of Tartu.

Muysken, Pieter. 1995. Code-switching and grammatical theory. In Lesley Milroy & Pieter Muysken (eds.), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, 177–198. Cambridge, England; New York: Cambridge University Press.

Muysken, Pieter. 1997. Code-switching processes: Alternation, insertion, congruent lexicalization. In Martin Pütz (ed.), *Language choices: Conditions, constraints, and consequences* (Impact: studies in language and society 1), 361–380. Amsterdam; Philadelphia, PA: John Benjamins Pub. Co.

Muysken, Pieter. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge, England; New York: Cambridge University Press.

Muysken, Pieter, Hetty Kook & Paul Vedder. 1996. Papiamento/Dutch code-switching in bilingual parent-child reading. *Applied Psycholinguistics* 17(04). 485–505.

Myers-Scotton, Carol. 1993. *Duelling languages: Grammatical structure in code-switching*. Oxford; New York: Clarendon Press; Oxford University Press.

Myers-Scotton, Carol. 1995. A lexically based model of code-switching. In Lesley Milroy & Pieter Muysken (eds.), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, 233–256. Cambridge, England; New York: Cambridge University Press.

Myers-Scotton, Carol. 2001. The matrix language frame model: Development and responses. In Rodolfo Jacobson (ed.), *Codeswitching worldwide II* (Trends in linguistics 126), 23–58. Berlin; New York: Mouton de Gruyter.

Myers-Scotton, Carol. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford; New York: Oxford University Press.

Myers-Scotton, Carol. 2006. Natural codeswitching knocks on the laboratory door. *Bilingualism: Language and Cognition* 9(02). 203–212.

Myers-Scotton, Carol & Janice L. Jake. 1995. Matching lemmas in a bilingual language competence and production model: Evidence from intrasentential code-switching. *Linguistics* 33. 981–1024.

Myers-Scotton, Carol & Janice L. Jake. 2000. Testing the 4-M model: An introduction. *International Journal of Bilingualism* 4(1). 1–8.

Myers-Scotton, Carol & Janice L. Jake. 2016. Revisiting the 4-m model: Code-switching and morpheme election at the abstract level. *International Journal of Bilingualism* 21(3). 340–366.

Naiditch, Larissa. 2008. Tendencii razvitija russkogo jazyka za rubežom: Russkij jazyk v izraile [development trends of the Russian language abroad: Russian in Israel]. *Russian Linguistics* 32(1). 43–57.

New, Boris, Marc Brysbaert, Juan Segui, Ludovic Ferrand & Kathleen Rastle. 2004. The processing of singular and plural nouns in French and English. *Journal of Memory and Language* 51(4). 568–585.

Newell, Allen. 1990. *Unified theories of cognition* (The William James lectures 1987). Cambridge, MA: Harvard University Press.

Nortier, Jacomine. 1990. *Dutch-Moroccan Arabic code switching among Moroccans in the Netherlands*. Dordrecht, Holland; Providence, RI: Foris.

Nortier, Jacomine & Henriette Schatz. 1988. Van éénwoordwisseling naar ontlening, een vergelijkend onderzoek [From one-word switching to derivation: A comparative study]. *Forum der Letteren* 29. 161–178.

Oakes, Michael P. 1998. *Statistics for corpus linguistics* (Edinburgh textbooks in empirical linguistics). Edinburgh: Edinburgh University Press.

Oldfield, Richard Ch. & Arthur Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17(4). 273–281.

Olshtain, Elite & Shoshana Blum-Kulka. 1989. Happy Hebrish: Mixing and switching in American-Israeli family interactions. In Susan M. Gass, Carolyn Madden, Dennis R. Preston & Larry Selinker (eds.), *Variation in second language acquisition*, 59–83. Clevedon, England; Philadelphia, PA: Multilingual Matters.

Parafita Couto, Maria del Carmen, Margaret Deuchar & Marika Fusser. 2015. How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution. In Gerald Stell & Kofi Yakpo (eds.), *Code-switching between structural and sociolinguistic perspectives* (linguae & litterae 43), 65–84. Berlin; Boston, PA: De Gruyter.

Paul, Hermann. 1920[1880]. *Prinzipien der Sprachgeschichte*. 5th. Halle: Max Niemeyer.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication* (Applied linguistics and language study), 191–225. London; New York: Longman.

Pfaff, Carol. 1979a. Constraints on language mixing: Intrasentential code-mixing and borrowing in Spanish/English. *Language* 55. 291–318.

Pfaff, Carol W. 1979b. Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language* 55(2). 291–318.

Pickering, Martin J. & Holly P. Branigan. 1998. The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language* 39(4). 633–651.

Pickering, Martin J. & Holly P. Branigan. 1999. Syntactic priming in language production. *Trends in Cognitive Sciences* 3(4). 136–141.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee & Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure* (Typological studies in language 45), 137–57. Amsterdam; Philadelphia, PA: John Benjamins.

Pierrehumbert, Janet. 2002. Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory phonology 7*, 101–39. Berlin; Boston, PA: Mouton de Gruyter.

Pine, Julian M. & Elena V. Lieven. 1993. Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language* 20(3). 551–571.

Pine, Julian M. & Elena V. M. Lieven. 1997. Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics* 18. 123–138.

Pizzuto, Elena & Maria Cristina Caselli. 1992. The acquisition of Italian morphology: Implications for models of language development. *Journal of Child Language* 19(03). 491–557.

Poplack, Shana. 1980a. Deletion and disambiguation in Puerto Rican Spanish. *Language* 56(2). 371–385.

Poplack, Shana. 1980b. Sometimes i'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. *Linguistics* 18(7/8). 581–618.

Poplack, Shana. 2004. Code-switching / Sprachwechsel. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics / Soziolinguistik: An international handbook of the science of language and society*, vol. 1, 589–596. Berlin: Mouton De Gruyter.

Poplack, Shana. 2011. What does the Nonce Borrowing Hypothesis hypothesize? *Bilingualism: Language and Cognition* 15(03). 644–648.

Poplack, Shana. 2018. *Borrowing: Loanwords in the speech community and in the grammar*. New York: Oxford University Press.

Poplack, Shana & Nathalie Dion. 2012. Myths and facts about loanword development. *Language Variation and Change* 24(03). 279–315.

Poplack, Shana & Marjorie Meechan. 1995. Patterns of language mixture: Nominal structure in Wolof-French and Fongbe-French bilingual discourse. In Lesley Milroy & Pieter Muysken (eds.), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, 199–232. Cambridge, England; New York: Cambridge University Press.

Poplack, Shana & David Sankoff. 1984. Borrowing: The synchrony of integration. *Linguistics* 22(1). 99–136.

Poplack, Shana, David Sankoff & Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26(1). 47–104.

Poplack, Shana & Sali A. Tagliamonte. 1996. Nothing in context: Variation, grammaticization and past time marking in Nigerian Pidgin English. In Philip Baker & Anand Syea (eds.), *Changing meanings, changing functions: Papers relating*

to grammaticalization in contact languages, 71–94. London: University of Westminster.

Poplack, Shana & Rena Torres Cacoullos. 2014. Linguistic emergence on the ground: A variationist paradigm. In Brian MacWhinney & William O'Grady (eds.), *The handbook of language emergence*, 267–291. Hoboken: Wiley-Blackwell.

R Core Team. 2010. *R: A language and environment for statistical computing. R foundation for statistical computing*. Vienna, Austria. http://www.R-project.org.

Ramers, Karl Heinz & Heinz Vater. 1992. *Einführung in die Phonologie*. 3rd edn. (Kölner linguistische Arbeiten, Germanistik 16). Hürth-Efferen: Gabel.

Reali, Florencia & Morten H. Christiansen. 2007. Word chunk frequencies affect the processing of pronominal object-relative clauses. *The Quarterly Journal of Experimental Psychology* 60(2). 161–170.

Remennick, Larissa. 2003. What does integration mean? Social insertion of Russian immigrants in Israel. *Journal of International Migration and Integration / Revue de l'integration et de la migration internationale* 4(1). 23–49.

Remennick, Larissa. 2017. Generation 1.5 of Russian-speaking immigrants in Israel and Germany: An overview of recent research and a German pilot study. In Ludmila Isurin & Claudia Maria Riehl (eds.), *Integration, identity and language maintenance in young immigrants: Russian Germans or German Russians* (IMPACT: Studies in Language and Sociology), 69–98. Amsterdam; Philadelphia, PA: Benjamins.

Richards, Jack C. 1970. A psycholinguistic measure of vocabulary selection. *IRAL - International Review of Applied Linguistics in Language Teaching* 8(2). 87–102.

Riehl, Claudia Maria. 2017. Russian-Germans: Historical background, language varieties, and language use. In Ludmila Isurin & Claudia Maria Riehl (eds.), *Integration, identity and language maintenance in young immigrants: Russian Germans or German Russians* (IMPACT: Studies in Language and Sociology), 11–40. Amsterdam; Philadelphia, PA: Benjamins.

Rijkhoff, Jan. 2009. On the co-variation between form and function of adnominal possessive modifiers in Dutch and English. In William B. McGregor (ed.), *The expression of possession* (The expression of cognitive categories 2), 51–106. Berlin; New York: Mouton de Gruyter.

Roll, Heike. 2003. Young ethnic German immigrants from the Former Soviet Union: German language proficiency and its impact on integration. In Rainer Münz & Rainer Ohliger (eds.), *Diasporas and ethnic migrants: Germany, Israel, and post-Soviet successor states in comparative perspective*, 272–288. London; Portland, OR: Frank Cass.

Russian National Corpus. 2003–2014. no address: no publisher. http://www.ruscorpora.ru/en/ (8 October, 2014).

Salamoura, Angeliki & John N. Williams. 2006. Lexical activation of cross-language syntactic priming. *Bilingualism: Language and Cognition* 9(03). 299–307.

Sankoff, David, Shana Poplack & Swathi Vanniarajan. 1990. The case of the nonce loan in Tamil. *Language Variation and Change* 2(1). 71–101.

Schäfer, Michael. 2014. *Phonetic reduction of adverbs in Icelandic: On the role of frequency and other factors* (New ideas in human interaction (NIHIN)). Freiburg: Freiburg University Library.

Schiffman, Harold F. 1999. *A reference grammar of spoken Tamil*. Cambridge, UK: Cambridge University Press.

Schmitt, Norbert (ed.). 2004. *Formulaic sequences: Acquisition, processing, and use* (Language learning and language teaching 9). Amsterdam; Philadelphia: John Benjamins.

Schmitt, Norbert, Sarah Grandage & Svenja Adolphs. 2004. Are corpus-derived recurrent clusters psychologically valid? In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing, and use* (Language learning and language teaching 9), 127–148. Amsterdam; Philadelphia, PA: John Benjamins.

Schneider, Ulrike. 2014. *Frequency, chunks and hesitations: A usage-based analysis of chunking in English* (New ideas in human interaction (NIHIN)). Freiburg: Freiburg University Library.

Schoonbaert, Sofie, Robert J. Hartsuiker & Martin J. Pickering. 2007. The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language* 56(2). 153–171.

Schröder, Peter. 1997. *Schlichtungsgespräche. ein textband mit einer exemplarischen analyse*. Berlin; New York: De Gruyter.

Schwitalla, Johannes. 2006. *Gesprochenes Deutsch. Eine Einführung*. 3., neu bearbeitete Aufl. (Grundlagen der Germanistik 33). Berlin: Erich Schmidt Verlag.

Sebba, Mark. 2009. On the notions of congruence and convergence in code-switching. In Barbara E Bullock & Almeida Jacqueline Toribio (eds.), *The Cambridge handbook of linguistic code-switching*, 40–57. Cambridge, England; New York: Cambridge University Press.

Seiler, Hansjakob. 1976. Determination: A universal dimension for inter-language comparison. *Arbeiten des Kölner Universalien-Projekts* 23. 1–29.

Sereno, Joan A. & Allard Jongman. 1997. Processing of English inflectional morphology. *Memory & Cognition* 25(4). 425–437.

Shastri, S. V. & Gerhard Leitner. 2002. *The international corpus of English: Indian corpus*. CD-ROM. London: Survey of English Usage, University College.

Sherwood, Lauralee. 2012. *Fundamentals of human physiology*. 4th ed. Belmont, CA: Brooks/Cole Cengage Learning.

Simpson-Vlach, Rita & Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4). 487–512.

Sinclair, John M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford university press.

Sinclair, John McHardy. 2003. *Reading concordances: An introduction*. London; New York: Pearson/Longman.

Siyanova-Chanturia, Anna, Kathy Conklin & Walter J. B. van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(3). 776–784.

Sosa, Anna Vogel & James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language* 83(2). 227–236.

Stammers, Jonathan R. & Margaret Deuchar. 2012. Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition* 15(03). 630–643.

Stemberger, Joseph Paul & Brian MacWhinney. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition* 14(1). 17–26.

Stenson, Nancy. 1990. Phrase structure congruence, government, and Irish-English code-switching. In Randall Hendrick (ed.), *The syntax of the modern Celtic languages* (Syntax and semantics. 23), 167–197. San Diego: Academic Press.

Švedova, Natalija Ju. (ed.). 2005 [1980](a). *Russkaja grammatika [Russian grammar]*. Vol. 1: Fonetika, Fonologija, Udarenie, Intonacija, Slovoobrazovanie, Morfologija [Phonetics, Phonology, Stress, Intonation, Word-Formation, Morphology]. Moskva: Institut russkogo jazyka im. V.V. Vinogradova.

Švedova, Natalija Ju. (ed.). 2005 [1980](b). *Russkaja grammatika [Russian grammar]*. Vol. 2: Sintaksis [Syntax]. Moskva: Institut russkogo jazyka im. V.V. Vinogradova.

Szabó, Csilla Anna. 2010. *Language shift und Code-mixing: Deutsch-ungarisch-rumänischer Sprachkontakt in einer dörflichen Gemeinde in Nordwestrumänien* (Variolingua. Nonstandard – Standard – Substandard 38). Frankfurt/Main; Berlin; Bern; Bruxelles; New York; Oxford; Wien: Lang.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis, syntactic persistence in spoken English* (Trends in linguistics 177). Berlin; New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2013. The great regression: Genitive variability in late modern English news texts. In Kersti Borjars, David Denison & Alan Scott (eds.), *Morphosyntactic categories and the expression of possession*, 59–88. Amsterdam; Philadelphia, PA: John Benjamins.

Tabossi, Patrizia, Rachele Fanari & Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2). 313–327.

Taft, Marcus. 1979. Recognition of affixed words and the word frequency effect. *Memory & Cognition* 7(4). 263–272.

Taft, Marcus & Kenneth I. Forster. 1976. Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior* 15(6). 607–620.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(02). 135–178.

Tajsner, Przemysław. 1997. Licensing of bare NP adjuncts. In Jacek Fisiak, Raymond Hickey & Stanisław Puppel (eds.), *Language history and linguistic modelling: A festschrift for Jacek Fisiak on his 60th birthday*, vol. 2 (Trends in linguistics 101), 1231–1244. Berlin; New York: Mouton de Gruyter.

Tannen, Deborah. 1989. *Talking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge, England: Cambridge University Press.

Taylor, John R. 2012. *The mental corpus: How language is represented in the mind*. Oxford; New York: Oxford University Press.

The "Five Graces Group", Clay Beckner, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59. 1–26.

Thomason, Sarah G. 2008. Social and linguistic factors as predictors of contact-induced change. *Journal of Language Contact* 2(1). 42–56.

Thomason, Sarah Grey. 2001. *Language contact*. Edinburgh: Edinburgh University Press.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Timberlake, Alan. 2004. *A reference grammar of Russian*. Cambridge, UK: Cambridge University Press.

Tolts, Mark. 2009. Post-Soviet aliyah and Jewish demographic transformation. In *15th world congress of Jewish studies, jerusalem, August 2-6*. Berman Jewish Policy Archive at New YorkU Wagner. http://www.bjpa.org/Publications/details.cfm?PublicationID=11924 (19 April, 2015).

Tomasello, Michael. 1992. *First verbs: A case study of early grammatical development*. Cambridge, England; New York: Cambridge University Press.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA; London, England: Harvard University Press.

Torres Cacoullos, Rena. 2015. Gradual loss of analyzability: Diachronic priming effects. In Aria Adli, Göz Kaufmann & Marco García García (eds.), *Variation in language: System- and usage-based approaches* (linguae & litterae 50), 265–288. Berlin; Boston, PA: De Gruyter.

Torres Cacoullos, Rena & James A. Walker. 2009. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47(1). 1–43.

Travis, Catherine E. 2005. The *yo-yo* effect: Priming in subject expression in Colombian Spanish. In Randall Gess & Edward J. Rubin (eds.), *Theoretical and experimental approaches to Romance linguistics* (Current Issues in Linguistic Theory 272), 329–349. Amsterdam: John Benjamins.

Travis, Catherine E. & Rena Torres Cacoullos. 2016. Two languages, one effect: Structural priming in spontaneous code-switching. *Bilingualism: Language and Cognition* 4(19). 733–753.

Treffers-Daller, Jeanine. 1994. *Mixing two languages: French-Dutch contact in a comparative perspective* (Topics in Sociolinguistics 9). Berlin; New York: Mouton de Gruyter.

Treffers-Daller, Jeanine. 2005a. Code-switching / Sprachwechsel. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics / Soziolinguistik: An international handbook of the science of language and society*, vol. 2, 1469–1482. Berlin: Mouton De Gruyter.

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

Tremblay, Antoine & Harald R. Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 151–173. London; New York: Continuum.

Tremblay, Antoine, Bruce Derwing & Gary Libben. 2009. Are lexical bundles stored and processed as single units? *Working Papers of the Linguistics Circle of the University of Victoria* 19. 258–279.

Tremblay, Antoine, Bruce Derwing, Gary Libben & Chris Westbury. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61(2). 569–613.

Trudgill, Peter. 1974. *The social differentiation of English in Norwich*. Cambridge, England; New York: Cambridge University Press.

Trueswell, John C., Michael K. Tanenhaus & Christopher Kello. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 19(3). 528–553.

Vahtin, Nikolaj, Oksana Žironkina, Irina Liskovec & Ekaterina Romanova. 2003. *Novye jazyki novyh gosudarstv: Javlenija na styke blizkorodstvennyh jazykov na postsovetskom prostranstve [New languages of new states: Contact phenomena in genetically related languages in the post-Soviet space]*. https://web.archive.org/web/20130503063312/http://old.eu.spb.ru/ethno/projects/project3/ukraine/007/012.htm (21 August, 2020).

van Hout, Roeland & Pieter Muysken. 1994. Modeling lexical borrowability. *Language Variation and Change* 6(01). 39–62.

van Hell, Janet G. & Annette M. B. de Groot. 1998. Disentangling context availability and concreteness in lexical decision and word translation. *The Quarterly Journal of Experimental Psychology Section A* 51(1). 41–63.

Verhagen, Véronique, Maria Mos, Ad Backus & Joost Schilperoord. 2018. Predictive language processing revealing usage-based variation. *Language and Cognition* 10(2). 329–373.

Verschik, Anna. 2004. Aspects of Russian-Estonian codeswitching: Research perspectives. *International Journal of Bilingualism* 8(4). 427–448.

Verschik, Anna. 2008. *Emerging bilingual speech: From monolingualism to code-copying*. London, New York: Bloomsbury Publishing.

Vogt, Hans. 1954. Rev. of *Languages in Contact: Findings and Problems*, by Uriel Weinreich. *Word* 10(1). 79–82.

Voigt, Herman A. 1994. Code-wisseling, taalverschuiving en tallverandering in het Melaju Sini. Tilburg University.

Weiner, E. Judith & William Labov. 1983. Constraints on the Agentless Passive. *Journal of Linguistics* 19(1). 29–58.

Weinreich, Uriel. 1979 [1953]. *Languages in contact: Findings and problems*. Ninth Printing. The Hague; Paris; New York: Mouton.

Weinreich, Uriel, William Labov & Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics*, 95–195. Austin, TX: University of Texas Press.

Whitney, William Dwight. 1881. On mixing in languages. *Transactions of the American Philological Associations* 12. 5–26.

Wiechmann, Daniel. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.

Worbs, Susanne, Eva Bund, Martin Kohls & Christian Babka von Gostomski. 2013. *(Spät-)Aussiedler in Deutschland. Eine Analyse aktueller Daten und Forschungsergebnisse* (Forschungsberichte 20). Nürnberg: Bundesamt für Migration und Flüchtlinge. https://www.bamf.de/SharedDocs/Anlagen/DE/Forschung/Forschungsberichte/fb20-spaetaussiedler.html (19 October, 2020).

Wortschatz: Corpus français. 1998–2020. Abteilung für Sprachverarbeitung, Universität Leipzig. https://corpora.uni-leipzig.de/de?corpusId=fra_mixed_2012&word= (21 August, 2020).

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge, England; New York: Cambridge University Press.

Wray, Alison. 2008. *Formulaic language: Pushing the boundaries*. Oxford; New York: Oxford University Press.

Wurzel, Wolfgang Ullrich. 1984. *Flexionsmorphologie und natürlichkeit*. Berlin: Akademie Verlag.

Zabrodskaja, Anastassia. 2009. Evaluating the Matrix Language Frame model on the basis of a Russian-Estonian codeswitching corpus. *International Journal of Bilingualism* 13(3). 357–377.

Zaliznjak, Andrej Anatol'evič. 2002 [1967]. *"Russkoe imennoe slovoizmenenie" s priloženiem izbrannyx rabot po sovremennomu russkomu jazyku i obščemu jazykoznaniju ["Russian nominal inflection" with a supplement of selected papers on modern Russian and general linguistics]*. Moscow: Jazyki slavjanskoj kul'tury.

Zaliznjak, Andrej Anatol'evič. 2009 [1977]. *Grammaticeskij slovar' russkogo jazyka: Slovoizmenenie [Dictionary of Russian grammar]*. 5th edn. Moscow: AST-Press Kniga.

Ždanova, Vladislava & Dmitrij Trubčaninov. 2001. Nekotorye osobennosti rečevogo povedenija russkojazyčnoj diaspory v Germanii [some specific speech characteristics of the Russian diaspora in Germany]. In Katharina Böttger, Sabine Dönninghaus & Robert Marzari (eds.), *Beiträge der europäischen slavistischen Linguistik (POLYSLAV) 4*, vol. 12 (Die Welt der Slaven. Sammelbände/Sborniki [edited volumes]), 274–285. München: Otto Sagner.

Zemskaja, Elena A. 1979. *Russkaja razgovornaja reč': Lingvističeskij analiz i problemy obučenija [colloquial Russian speech: A linguistic analysis and teaching issues]*. Moskva: Russkij jazyk.

Zenner, Eline, Dirk Speelman & Dirk Geeraerts. 2015. A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show. *International Journal of Bilingualism* 19(3). 333–346.

Zeschel, Arne. 2010. Exemplars and analogy: Semantic extension in constructional networks. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (Cognitive linguistics research 46), 201–219. Berlin; New York: De Gruyter Mouton.

Zifonun, Gisela, Ludger Hoffmann, Bruno Strecker, Joachim Ballweg, Ursula Brauße, Eva Breindl, Ulrich Engel, Helmut Frosch, Ursula Hoberg & Klaus Vorderwülbecke. 1997. *Grammatik der deutschen Sprache*. Vol. 3 (Schriften des Instituts für Deutsche Sprache Bd. 7). Berlin; New York: W. de Gruyter.

Explaining Russian-German code-mixing

The study of grammatical variation in language mixing has been at the core of research into bilingual language practices. Although various motivations have been proposed in the literature to account for possible mixing patterns, some of them are either controversial, or remain untested. Little is still known about whether and how frequency of use of linguistic elements can contribute to the patterning of bilingual talk. This book is the first to systematically explore the factor usage frequency in a corpus of bilingual speech. The two aims are (i) to describe and analyze the variation in mixing patterns in the speech of Russian German adolescents and young adults in Germany, and (ii) to propose and test usage-based explanations of variation in mixing patterns in three morphosyntactic contexts: the adjective-modified noun phrase, the prepositional phrase, and the plural marking of German noun insertions in bilingual sentences. In these contexts, German noun insertions combine with either Russian or German words and grammatical markers, thus yielding mixed bilingual and German monolingual constituents in otherwise Russian sentences, the latter also labelled as embedded-language islands. The results suggest that the frequency with which words are used together mediates the distribution of mixing patterns in each of the examined contexts. The differing impacts of co-occurrence frequency are attributed to the distributional and semantic specifics of the analyzed morphosyntactic configurations. Lexical frequency has been found to be another important determinant in this variation. Other factors include recency, or lexical priming, in discourse in the case of prepositional phrases, and phonological and structural similarities and differences in the inflectional systems of the contact languages in the case of plural marking.