

CODE-SWITCHING BOUNDARIES IN KAZAKH–RUSSIAN CONVERSATION:
THE ROLES OF PROSODY AND SYNTAX

by

Dariga Arapbayeva

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Arts in
Eurasian Studies

at

NAZARBAYEV UNIVERSITY
SCHOOL OF SCIENCES AND HUMANITIES

2026

THESIS APPROVAL FORM
NAZARBAYEV UNIVERSITY
SCHOOL OF SCIENCES AND HUMANITIES

**CODE-SWITCHING BOUNDARIES IN KAZAKH–RUSSIAN CONVERSATION:
THE ROLES OF PROSODY AND SYNTAX**

BY

Dariga Arapbayeva

NU Student Number: 201946540

APPROVED

BY

Dr. CLINTON PARKER ASSISTANT PROFESSOR

ON

The 28th of May, 2026



Signature of Principal Thesis Adviser

In Agreement with Thesis Advisory Committee

Second Adviser: Dr. Andrey Filchenko Professor

External Adviser: Dr. Giorgia Troiani Postdoctoral Fellow, University of Bologna

Abstract

This study investigates how intonation units and their syntactic structure co-occur with code-switching (CS), alternation between languages within one discourse, in Kazakh-Russian bilingual speech. Existing previous research has studied CS patterns through syntactic constraints such as the Equivalence Constraint, the Free Morpheme Constraint, and the Functional Head Constraint; however, these constraints have shown inconsistent results across typologically distant language pairs. These inconsistencies have encouraged researchers to look beyond syntactic constraints, such as prosodic approaches to CS, which remain rare, particularly for typologically distant language pairs. One of the prosodic-discourse frameworks developed by Shenk (2006) has only been applied once again to the typologically similar Spanish-English pair. And this present study applied Shenk's (2006) prosodic-discourse framework to naturally occurring Kazakh-Russian (typologically distinct language pair) speech, annotated 3,796 intonation units from 90 minutes of recordings for language, syntactic category, and switch-boundary intonation unit (SBIU), and developed two new switch-type categories: embedded switch (ES) and boundary switch (BS). As a result, prosodic and syntactic structures do not operate as competing explanations but as complementary ones. In conclusion, each of them functions at a different level of organization: IU boundaries mark where switches tend to occur, while syntactic unit type is associated with what kind of switching appears. The study also offers an empirical basis for Russian elements' integration into Kazakh speech, more precisely, what their positional behavior within IU can help distinguish borrowing from active code-switching.

Contents

Abstract	i
1 Introduction	1
1.1 Background	1
1.2 Research questions	3
2 Literature review	6
2.1 Code-switching	6
2.2 Intonation Units	11
2.3 Kazakh-Russian code-switching	13
3 Methodology	17
3.1 Data description	17
3.2 Language annotation	18
3.2.1 Researcher positionality	18
3.2.2 Further clarifications	20
3.3 Syntactic annotation	21
3.3.1 Finite clause	22
3.3.2 Non-finite clause	23
3.3.3 Phrasal units	24
3.3.4 Discourse markers	25
3.3.5 Interactional intonation units	26
3.3.6 Switch types: Embedded switch and Boundary switch	26

3.4	Switch-boundary intonation unit annotation	28
3.4.1	SBIU annotation levels	30
3.5	Data processing and quantification	32
4	Findings	33
4.1	Dataset composition	33
4.2	Syntactic composition of the dataset	35
4.3	SBIU, BS, and ES and their syntax alignment	36
4.3.1	SBIU	36
4.3.2	Switch types: Embedded switch and Boundary switch	37
4.3.3	SBIU-syntax alignment	39
4.3.4	BS and ES syntax alignment	40
4.4	Per-speaker SBIU	41
4.4.1	SBIU alignment with syntax	43
5	Discussion	44
5.1	Dataset composition	44
5.1.1	Bilingual vs monolingual IUs	44
5.1.2	IU length	45
5.2	IU boundaries	46
5.2.1	Inter-IU switching: SBIU evidence	46
5.2.2	Intra-IU switching: BS and ES evidence	48
5.3	Syntactic structure of the dataset	50
5.3.1	Finite clauses vs phrasal units	50
5.3.2	Non-finite clauses	51
5.3.3	Discourse markers	53
5.3.4	Prosody vs syntax: complementary roles	55
5.4	Russian element integration	57
5.4.1	Positional distribution or borrowing cline	57
5.4.2	Morphological integration patterns	64
5.5	Limitations	66

6	Conclusion	69
7	Research ethics	71
	References	72

Chapter 1

Introduction

1.1 Background

Grosjean (2024) writes that about 65% of people worldwide are bilingual, meaning more than half of the population speaks at least two languages. And for these speakers, code-switching (CS), alternating between two or more languages in a conversation, is common and follows systematic patterns (Poplack, 1980). These patterns can follow both the structural properties of the languages in contact and the cognitive organization of the bilingual speaker (Rayo et al., 2024). Despite the dominance of syntactic approaches, some work has examined code-switching from a prosodic perspective as well. Shenk (2006), for example, examined the alignment of IU boundaries with CS in Spanish-English bilingual speech and found a strong correlation between the two. The new approach of studying code-switching through prosody in Shenk (2006) contributed to both the framework and research focus of this study: examining how code-switching patterns in Kazakh-Russian bilingual speech align with prosodic organization. The present study does not replicate Shenk’s (2006) syntactic completion point analysis, but it does annotate the syntactic category of each IU to examine what structure can reveal about upcoming switches. The central question is how code-switching in Kazakh-Russian bilingual speech is distributed across prosodic units, and what that distribution reveals about

the organization of switching, and how syntactic structure contributes to that understanding.

In the majority of existing literature, the patterns of CS have been examined for typologically similar language pairs, such as Spanish-English bilingualism in North American contexts (Sankoff and Poplack, 1981). However, with the growing interest in these foundational constraints, when researchers have applied the same frameworks to typologically distant pairs (languages that differ fundamentally in word order, morphological type, and strategies for expressing subordination), those syntactic models started to show inconsistent results (Berk-Seligson, 1986; Djennane and Louati, 2024). This suggests that the patterns documented for some pairs may not be universal in bilingual discourse, that typologically distant languages may have different patterns for CS. Kazakh and Russian represent one of such a typologically distant pair, where the first is an agglutinative language in Turkic family, whereas the second is fusional Slavic language. And despite structural differences, dense contact between these languages has led to a bilingual community in which code-switching is common in everyday speech.

In modern-day Kazakhstan, Kazakh is the only state language but Russian still has the status of an official language (Smagulova, 2008). Despite efforts to promote the state language, Russian remains the predominant language for education, government, and many public domains of the Kazakhstani people (Smagulova, 2008). Starting with the decades of russification during Soviet times, which made Kazakhstan, among other Central Asian countries, the most russified, and continued with the 1995 constitutional status change of Russian from interethnic to official status, it remains the dominant language in the country (Terlikbayeva and Menlibekova, 2021; Smagulova, 2008). Such prolonged contact, ongoing to this day, led to bilingual communities in which CS became the norm in the region’s linguistic landscape. In the existing literature on Kazakh-Russian contact, most studies focus on the status dynamics between the two languages, particularly how Russian has influenced Kazakh vocabulary and grammar. As an instance of grammatical change, Muhamedowa’s (2016) book illustrates change in the usage of the passive-perfective suffix *-ğan* in Kazakh. Before the influence of Russian, *-ğan* was a

perfect/reportive form. In saying so, Muhamedowa (2016) states that it mostly marked events as recalled in oral communication, and example (1) shows how it was used as a spoken retelling:

- (1) Keše ciyrk-qa bar-ğam
yesterday circus-DAT go-PRF-1SG
‘I went to the circus yesterday.’ (Muhamedowa, 2016, p. 190)

However, after the influence of Russian, she demonstrates that it started to be used in official writings, such as government notes. Saying that *-ğan* forms started to get used similarly to the Russian past passive participle. For instance, *bekitil-ğan* (‘it was resolved’) corresponds to the Russian *utverždeno*. Muhamedowa (2016) states that such a change from the Russian way of using past passive constructions in such declarations.

Although Kazakh-Russian contact literature includes many works that discuss structural and status dynamics, almost no attention has been paid to code-switching in the context of prosodic interactions. Therefore, this study attempts to address CS patterns in Kazakh-Russian bilingual speech through both prosodic and syntactic organization, aligned with it, with a focus on the intonation unit (IU) as the primary unit of analysis. This thesis adopts the analytical framework developed by Shenk (2006) for Spanish-English bilingual speech, with several modifications, and applies it to the typologically distant language pair for the first time.

1.2 Research questions

Following both Shenk’s (2006) initial framework and the scope of the current study’s methodology, this thesis seeks to answer whether intonation units can explain the patterns of the Kazakh-Russian structure through three main research questions. The first asks whether IU boundaries function as switching sites at all and without this, the remaining questions have no empirical foundation. Secondly, the existing literature on syntactic constraints on CS patterns contributed to the kinds of patterns that could be anticipated or even absent. For example, whether there are grammatical differences

between Kazakh and Russian in terms of IUs’ syntactic structures that block mixing between languages. Finally, this study newly comes up with new a switching type, embedded switch IUs, which help to test the positional distribution of Russian elements to address a methodological problem that runs through the entire study: the difficulty of distinguishing active code-switching from borrowing in a contact situation as dense as Kazakh-Russian.

- Do IU boundaries function as the dominant site for code-switching in Kazakh-Russian bilingual speech, at both the inter-IU (SBIU) and intra-IU levels (ES and BS)?
- What role does syntactic unit type play in shaping code-switching patterns? And how does it interact with prosodic (IU) structure?
- What does the positional and morphological behavior of Russian elements within IUs reveal about their degree of integration into Kazakh speech?

Following the three main questions, this study also aims to cover secondary questions on each subset:

1. Does boundary preference differ between the individual speaker and at the conversation levels in the annotation of the switch metric SBIU?
2. What is the distribution of embedded switches and boundary switches across the dataset, and syntactic category-wise? And do they concentrate on specific unit types?
3. Does IU length predict switch type?
4. Can positional behavior within IUs serve as an empirical criterion for distinguishing borrowing from active code-switching?

Following the main sets and subsets of the research questions, the first objective is to create a dataset of IUs from the existing MULTICORSKL (Filchenko et al., 2023) corpus and to annotate each IU with the categories listed in the methodology chapter. The second objective is to identify the mechanism for determining switch metrics in Kazakh-

Russian bilingual speech, taking into account its differences from Shenk's framework. Lastly, qualitatively retrieve the instances that address the position of IU to observe the borrowedness cline.

Chapter 2

Literature review

Existing literature on code-switching helps shape this study around three main topics. First, what are the earlier structural/syntactic approaches to understanding code-switching, and how do their application across different language pairs reveal their non-universality? The research gap comes from why these models fail in certain bilingual contexts. Particularly in contact language situations where syntactic structures diverge significantly (different word order, agglutinative vs fusional, etc.). Second, as this thesis examines IU as the primary unit of real-life speech analysis, this literature review will discuss existing literature on it generally. Third, the body of existing research on Kazakh-Russian bilingual speech, which considers the languages' divergent origins and long-standing close contact. Across these topics, the central discussion remains that code-switching is not random but follows systematic patterns. The key question, then, is what roles syntactic structure and prosodic organization, specifically the intonation unit, each play in code-switching patterns, and whether a prosodic framework can account for switching behavior in a typologically distant language pair such as Kazakh-Russian.

2.1 Code-switching

Code-switching (CS) has been studied primarily as a sociolinguistic and grammatical structural phenomenon (Gardner-Chloros, 2009). In both cases, researchers found that

bilingual speech is not random but organized and follows certain rules. However, the latter has been approached through the lens of the syntax of contacting languages only, without sufficient attention to discourse-level and prosodic factors such as intonation, turn-taking, and conversational organization (Muysken, 2000). To identify universal CS patterns, various researchers have proposed theories that correlate syntax and code-switching, particularly in Spanish-English bilingualism. This overrepresentation raises questions about typological generalizability, since the structural similarity between Spanish and English may be the reason for the observed patterns, rather than universal properties of code-switching (Berk-Seligson, 1986). One such early influential constraint is the Free Morpheme Constraint, proposed by Poplack, as is the Equivalence Constraint. The Free Morpheme Constraint states that a switch cannot occur between a bound morpheme and a lexical stem, essentially prohibiting intra-word switching of inflectional or derivational affixes. In practice, this means switches must occur at the level of the word or above, between full sentences, clauses, or constituents, rather than within a word itself. For example, a switch like *run-ando*, where the English verb stem *run* is combined with the Spanish progressive suffix *-ando* (equivalent to English ‘-ing’), would be prohibited because the switch occurs within a single word, between a stem and a bound morpheme. By contrast, switching between full words or phrases is permitted, as in *I was muy cansada* (‘I was very tired’), where the English auxiliary *was* is followed by a complete Spanish adjectival phrase (Poplack, 2010). Another constraint, the Equivalence Constraint, states that switches mostly occur at points where the word order of both languages coincides and that such switches will be grammatically correct. For instance, she gives an example (2),

- (2) I wanted que Maria llamara
1SG_{en} want.PST_{en} COMP_{sp} Maria call.SBJV.PST.3SG_{sp}
‘I wanted Maria to call.’ (Poplack, 2010)

where the switch occurs at the boundary between the English main clause ‘I wanted’ and the Spanish subordinate clause *que María llamará* (‘that Maria would call’). Poplack writes that this switch is permitted because both English and Spanish place the subor-

dinate clause after the main verb, which allows the formed switch to be grammatically natural in both languages simultaneously. In other words, the Equivalence Constraint predicts that switches are most likely at positions where the syntactic structures of L1 and L2 run parallel, so that the utterance remains grammatical from the perspective of either language (Poplack, 2010). Another popular constraint that emerges from syntactic frameworks is the Functional Head Constraint (FHC), proposed by Belazi, Rubin, and Toribio (1994). Unlike The Free Morpheme Constraint, this constraint prohibits code-switching between a functional head such as a determiner or complementizer and its complement. For example, code-switching such as *the kitab* ('the book') where the determiner is in English and the noun is in Arabic is forbidden by this constraint, because the determiner and noun phrase must match in language.

However, despite being popular in bilingualism research, all constraints were challenged by typologically distant language pairs. For example, while testing The Equivalence Constraint on Spanish-English, researchers have found a high degree of consistency with the constraint, with fewer grammatical mistakes (Poplack, 2010). In contrast, research on typologically distant pairs such as Spanish-Hebrew (Berk-Seligson, 1986) and Korean-English (Choi, 1991) has found several counterexamples to this observation. For instance, Choi (1991) documents the Korean-English mixed utterance,

- (3) Rain *hae-lkke-ya* *onul?*
 Rain_{en} do-PROSP.Q-SFP_{kr} today_{kr}
 'Is it going to rain today?' (Choi, 1991)

in which the English noun *Rain* appears in sentence-initial position, followed by the Korean future auxiliary *hae-lkke-ya* and the Korean temporal adverb *onul*. This word order does not conform to either language fully: Korean is a verb-final (SOV) language that would normally place the verb at the end, while English would require 'Is it going to rain today?' with subject-auxiliary inversion. Despite violating the word order of both languages, the utterance is judged as natural by bilingual speakers (Choi, 1991). This demonstrates that the Equivalence Constraint fails to predict actual switching behavior when the two languages in contact have fundamentally different syntactic structures.

The same pattern of inconsistency was found in the application of the Free Morpheme Constraint. Djennane and Louati’s (2024) study of Algerian Arabic-French bilingual speech showed instances, such as *analyzer-hum*, in which the French infinitival verb stem *analyzer* (‘to analyze’) is directly inflected with the Arabic object suffix *-hum* (‘them’). As shown, it is an instance where an intra-word switch violates the constraint. Along with that, such examples are not random but frequent patterns in the data. Apart from that, they also observed that speakers produce mixed clauses and noun phrases in different grammatical environments. For example, in the following (4) example:

- (4) Les cours jbadidi:w bakri batni:n.
 DEF.ART.PL_{fr} lectures_{fr} 3PL.IMPF_{arb} early_{arb} on.Monday_{arb}
 ‘Lectures start early on Monday.’ (Djennane and Louati, 2024)

The switch occurred between subject and verb, violating the Equivalence Constraint, because it demonstrates that code-switching can occur within the clause at the subject-verb boundary. This violates the Equivalence Constraint because Arabic declarative sentences follow VSO word order, while French follows SVO, meaning the two languages do not share equivalent surface structure at the subject–verb boundary (Djennane and Louati, 2024). So if there is no structural alignment, the switch should have been blocked, but speakers freely produce it, once again demonstrating that the constraint fails to predict actual bilingual behavior.

Apart from that, Bahri (2025) also noticed similar inconsistencies regarding Arabic-English code-switching. She claims that in her data intra-sentential switches violate the Free Morpheme Constraint (34.5% of utterances) and an even larger proportion violate the Equivalence constraint (67.8%) (Bahri, 2025). Even authors who observed CS between the same language pairs and with SVO word order, have documented violations of the Free Morpheme Constraint. For example, in her study of Romanian-English CS, Ene (2007) found word examples such as:

- (5) overhead-ul
 overhead.N-DEF.ART_{ro}
 ‘the overhead’

- (6) choke-ănește
choke.V-PRS.3SG_{ro}
'he/she chokes'

In both words English stems got mixed directly with Romanian bound morphemes, which does not align with free-morpheme theory. According to Ene (2007), these are natural and frequently occurring examples. Moreover, since the English morphemes retain English pronunciation and are not phonologically adapted, they clearly count as an instance of code-switching. The Functional Head Constraint has also been extensively tested. One of the counterarguments that the FHC does not fully account for observed switching patterns were found in a Chinese-English CS study. Li et al. (2018) in their data found that, rather than showing categorical unacceptability predicted by the FHC, such as the determiner and NP can not be mixed, the observed eye-movement patterns of experiment participants showed switches with different manageable points, suggesting that functional head switches can be done within any point of bilingual speech rather than blocked by grammatical constraints.

Also, Akishev's (2020) analysis of Kazakh-Russian code-switching in reported clauses demonstrates that the complementizer language determines word order inside the reported clause, a constraint that neither the Equivalence Constraint nor the Functional Head Constraint predicts straightforwardly. This suggests that Kazakh-Russian switching requires language-specific syntactic analysis beyond the classic constraint-based frameworks.

Myers-Scotton's Matrix Language Frame states that one language acts as the matrix language and provides the grammatical frame. While the other is the embedded language that inserts its morphemes into that frame. For example, in Cree-English (Poplack, 2010) and Papiamentu-Dutch (Muysken, 2000), speakers follow the grammar of one dominant language while borrowing words or phrases from the other. However, the issue with the MLF model is that it does not address alternational code-switching, in which people change language at sentence or clause boundaries. For instance, in Greek-English alternational code-switching, bilingual speakers switch between fully formed Greek and English

sentences (Gardner-Chloros, 2009). In such cases, identifying a single matrix language becomes problematic, as neither language clearly dominates syntactically or morphologically, which makes the theory again non-universal to explain the principles of code-switching (Gardner-Chloros, 2009). After such inconsistencies of syntactic constraints, new perspectives are required to explain the complex patterns of bilingual speech. This shift moves the analysis from viewing code-switching as a static grammatical phenomenon to a more dynamic process of real-time production (Torres Cacoullos and Travis, 2019; Shenk, 2006).

Similarly, research on prosodic boundary perception in Swedish speech showed that listeners could effortlessly predict speech boundaries, which demonstrated reduced effort in both language production and processing, though this study focused on monolingual speakers, its relevance here lies in establishing that prosodic cues alone are sufficient for boundary perception, a finding that provides a baseline for understanding how such cues may function in bilingual switching contexts as well (Carlson et al., 2005). The only factors that helped listeners orient throughout the speech were prosodic cues such as pitch reset and creaky voice, which function as acoustic signals of upcoming prosodic boundaries rather than constituting boundaries themselves, and thus the authors conclude that listeners do not require syntactic knowledge to perceive and process speech boundaries effectively (Carlson et al., 2005).

Majority body of existing work suggests that prosody and discourse units may better describe where switching tends to appear in spontaneous speech (Torres Cacoullos and Travis, 2018; Carlson et al., 2005; Pattichis et al., 2023). This motivates the prosodic perspective, done by Petra Shenk and others, which emphasizes intonation units and discourse organization in relation to code-switching.

2.2 Intonation Units

Although intonation units (IUs) have been recognized as one of the primary units of prosody, little attention has been paid to how it helps us understand the language alter-

nations of bilingual speakers. All works found for the literature review of this project, as Inbar et al. (2023), Pattichis et al. (2023), and Shenk (2006) claim that their works are the one of the first to investigate IUs in speech production: one for neural mechanisms of communication, another for NLP analysis of CS, and the last for syntax-prosody interference of CS. Thus, we can see that IUs are highly understudied both as a unit in itself and for their structuring ability to organize bilingual speech in linguistics literature.

In her work that served as a framework for this study, Shenk (2006) argues that IUs are a productive unit for analyzing code-switching in bilingual discourse, because they reflect natural segmentation of speech into chunks that correspond to a speaker’s cognitive processing capacity. In saying that, she refers to Chafe’s (1994) work, which found that the IU is a unit that can maintain a certain amount of information in “active mode” within short-term memory, for example, approximately four words in one English IU. Following that, we can conclude that such cognitive limits reveal how speakers package ideas efficiently during speech production, making it meaningful and, at the same time, efficient to follow. Therefore, Chafe also claims IUs as a meaningful “thought groups” that often match changes in topic or point of view in a conversation (Chafe, 1994). In bilingual speech, IU boundaries coincide with points where language alternation tends to occur, and this co-occurrence does not appear to disrupt comprehension or increase cognitive load (Inbar et al., 2023). In fact, this feature of IU is supported by neurophysiological evidence. In their neuroscientific research, Inbar et al. (2023) used EEG to examine the brain’s response to IUs and found that responses occurred at the beginning and end of intonation units, suggesting that the brain recognizes IUs as meaningful speech chunks.

Additionally, Simpson (2016) writes that IU boundaries can be used to structure word-pair associations that form mnemonic blocks. Segmenting speech into manageable cognitive units (in this case, IUs) may facilitate the bilingual brain’s juggling of two linguistic systems simultaneously, reducing memory and processing difficulties during code-switching (Simpson, 2016). As a result, this facilitates speech planning and recall. Such features of IUs, as a unit, support their interpretation as fundamental cognitive units in language production rather than merely as acoustic phenomena (Simpson, 2016). Shenk

(2006) notes that switches tend to align with prosodic breaks rather than arbitrary points within clauses or words, and treats this alignment as the primary empirical observation rather than a consequence of prosodic organization. As it is also known from discourse studies, in addition to serving as cognitive units in prosodic cues, IUs contribute to managing interaction, turn-taking, and signaling meaning, all of which are critical in bilingual interactions in which interlocutors communicate effectively (Chafe, 1994). Such functions are particularly important in bilingual discourse, where conversational management across languages depends on both speakers tracking each other’s language choices in real time. Another mention of IUs’ potential as a CS metric, and as a direct empirical predecessor to the present study, is Troiani’s (2023) dissertation, which also draws on the MULTICORSKL corpus (Troiani, 2023). She mentions how IU boundaries were reliably identified in naturally occurring conversation by both trained and untrained listeners, which provides empirical grounding for the use of IUs as a cross-linguistic analytical unit.

Overall, these studies show that the IU is a cognitively grounded, empirically reliable, and cross-linguistically applicable unit.

2.3 Kazakh-Russian code-switching

Following the first sections, most research on code-switching has focused on languages that are similar, like English and Spanish. Because Kazakh and Russian are quite different, this study offers a new perspective on CS research. As mentioned above, earlier studies on Kazakh-Russian code-switching mainly examined sociolinguistic issues, such as language prestige and changing language policies. More recently, researchers have begun using mixed-methods approaches, combining data from language corpora and interviews, and testing well-known code-switching constraints for this language pair. For example, Muhamedowa (2005) observed how intrasentential code-switching dominates Kazakh-Russian bilingual discourse. And most specifically, switches are observed within clauses rather than at clause boundaries. In saying that, she refers to Muysken (2000)

typology. Muysken classifies code-switching into three types shown in examples from (7) to (9) (constructed following Muysken (2000));

(a) insertion (embedding elements from one language into another’s structure),

- (7) Ol vorobeydi kördi
 3SG_{kz} sparrow.ACC_{ru} see.PST.3SG_{kz}
 ‘He/she saw a sparrow.’

(b) alternation (switching between languages at clause boundaries),

- (8) Men bardym I told you
 1SG_{kz} go.PST.1SG_{kz} 1SG_{en} tell.PST_{en} 2SG_{en}
 ‘I went. I told you.’

(c) congruent lexicalization (mixing shared grammatical structure),

- (9) I saw el perro yesterday
 1SG_{en} see.PST_{en} DEF.M.SG_{sp} dog_{sp} yesterday_{en}
 ‘I saw the dog yesterday.’

And Auer and Muhamedova (2005) identifies insertion as the most common pattern, where Russian lexical items are embedded into Kazakh morphosyntactic frames. Russian words are incorporated into Kazakh lexical frames; in other words, Russian lexical items are mixed into the Russian-Kazakh contact discourse with Russian words placed into Kazakh word order and provided with Kazakh case suffixes and other late system morphemes of the Kazakh language. The quantitative analysis of Uzbek-Russian code-switching (structurally comparable to Kazakh-Russian) found that intrasentential switching accounted for 58% of all instances (Nakamura, 2025), suggesting a similar prevalence in Kazakh-Russian contexts. Thus, Russian elements undergo systematic adaptation to Kazakh morphology through several mechanisms. For example, Russian nouns take Kazakh case markers according to Kazakh grammatical rules (Nurtazina et al., 2019). Also Russian adjectives are simplified to masculine forms when preceding feminine or neuter nouns (Muhamedova, 2005). However, Kazakh elements entering Russian speech

exhibit different patterns of adaptation. Kazakh words in Russian texts undergo morphological and phonetic adaptation, forming new derivational series through Russian suffixes (Nassyrova, 2018). Kazakh words in Russian texts undergo morphological and phonetic adaptation, forming new derivational series through Russian suffixes (Nassyrova, 2018), including hybrid word formations where Kazakh roots combine with Russian affixes or Russian roots combine with Kazakh suffixes (Zharkynbekova and Chernyavskaya, 2022). Examples include formations such as *khabarit'* ('communicate') and *asarit'* ('help'), which mix Kazakh lexical roots with Russian verbal morphology (Zharkynbekova and Chernyavskaya, 2022). A related but distinct observation comes from Muhamedowa (2005), who documents how the systematic insertion of the Russian causal conjunction *potomu* ('because') and the complementizer *čto* ('that') into Kazakh clauses restructures the sentence in ways absent from monolingual Kazakh speech. Muhamedowa notes that while the borrowing of *potomu čto* is facilitated by its compatibility with Kazakh causal clause structure, where subordinate clauses can be introduced by free subjunctors, its consistent appearance at the beginning of the subordinate clause represents a structural pattern imported from Russian. She illustrates this with the following example (10), where *potomu čto* opens the causal clause while the remainder of the clause follows Kazakh morphosyntax:

- (10) Bări orındarında uže tirisip otir.
all place-PL-POSS-3-LOC yet make.an.effort-CON AUX-3

potomu čto šiq qalsa uže žumis žoq.
because go AUX-COND-3 yet work existent-3
'Everybody makes an effort at their workplaces, because if they leave their work,
they will not have a job.' (Muhamedowa, 2016, p. 47)

This example shows that Kazakh and Russian can each contribute to the grammar of the same clause simultaneously, which challenges the Matrix Language-Frame's (MLF) assumption that one language alone supplies the grammatical frame. These examples show that Kazakh and Russian sometimes both contribute to the grammar of the same clause, for example, Russian conjunctions structure the sentence while Kazakh provides case marking and morphology. Thus, it challenges the MLF, which says that in every

bilingual clause, one language provides the grammatical structure, and the other language only inserts words into it (Muhamedowa, 2009). Thus, studies on CS have found that such structural asymmetry manifests in integration patterns where Russian inflectional morphology undergoes simplification when entering Kazakh frames (Muhamedowa, 2005). Overall, such contradictions among popular syntactic constraint models underscore the need to examine code-switching patterns from additional perspectives, such as prosody. The limited existing research on Kazakh-Russian code-switching further justifies this empirical gap.

Chapter 3

Methodology

3.1 Data description

The data is drawn from the Multimedia Corpus of Modern Spoken Kazakh Language (Filchenko et al., 2023), a corpus of naturally occurring spoken discourse as family interactions, peer discussions, storytellings, and etc. The corpus was designed to represent Kazakh language use as it occurs, including its institutionalized coexistence with Russian (Troiani, 2023). To create a dataset for the present study four recordings were selected from the corpus, REC01 is Kazakh-dominant conversation of older people, and the remaining three recordings (REC02, REC03, REC04) capture conversations among younger speakers, students at Nazarbayev University and beyond. As of April 2026, the dataset contains 3,796 IUs, which is 90 minutes of speech. The recordings were already transcribed and segmented into IUs by trained annotators of the corpus following the Discourse Transcription conventions developed by Du Bois et al. (1993) and this keeps the transcription of natural spoken interaction rather than an idealized written form. Each IU corresponds to a single row in the dataset and includes the following annotated features: linguistic content; character, syllable, and word counts; language classification; syntactic category; and switch type (e.g., single-unit boundary IU, embedded switch, or boundary switch). Figure 3.1 illustrates the annotation structure applied to REC02.

G	L	M	N	O	P
iu	LA1	LA2	Syntax	SBIU	Switchtyp
Новогодняя ночь кажется	ru	ru	FC	SBIU	
отыз бірі да кешке	kz	kz	PU		
Аха	kz	kz	DM		
Сейтіп	kz	kz	DM	SBIU	
кішкентай елкамыз бар	bil	bil	FC	SBIU	ES

Figure 3.1: The structure of the dataset from REC02.

As this study did not use MULTICORSKL corpus (Filchenko et al., 2023) entirely but only four recordings and applied additional annotation to the original corpus, the annotated data is referred to as a dataset throughout this study, to avoid confusion with the source corpus.

3.2 Language annotation

As the core of IU content identification, language annotation must be highly systematized to prevent confusion, as it guides all other annotations. Each IU was classified by its language: Kazakh (kz), Russian (ru), or Bilingual (bil). Bilingual IUs contained elements from both languages, whether in lexis, morphology, or grammar. The process may seem simple at first glance. However, a key methodological challenge appeared in annotating the language of some IUs. Especially IUs that contain Russian words deeply integrated into Kazakh-Russian bilingual speech, as their status as a ‘switch’ is ambiguous. Even congruent lexicalization, the most applicable of Muysken’s (2000) three types, which describes how a foreign word gradually loses its foreign-like character through repeated use cannot straightforwardly apply here. Because it requires a degree of structural similarity between the two languages that Kazakh and Russian do not share, given their fundamental typological difference as agglutinative and fusional languages respectively (see congruent lexicalization example in section 2.3).

3.2.1 Researcher positionality

The annotator/researcher of the present study is a Kazakh-Russian bilingual with native competence in both languages, raised and educated in Kazakhstan. This background can

be an advantage and at the same time a methodological consideration that requires a careful approach. As a bilingual insider, the researcher understands the pragmatic and sociolinguistic norms of the community under study, which is valuable for interpreting naturalistic data that may not be available to annotators who speak only one or neither of the languages. Especially, awareness of which Russian items have been integrated into colloquial Kazakh speech is an important nuance when making language annotation decisions. Because certain Russian items, such as discourse markers and temporal adverbs, have become naturalized in modern Kazakh. For example, IU in example (11):

- (11) Men uzhe keldim
 1SG yet come.PST.1SG
 ‘I already came/arrived.’

Should this be classified as bilingual, since there is a presence of a Russian word, or as pure Kazakh, since ‘already’ has become part of everyday speech among native Kazakh speakers due to its high usage frequency? And the annotator’s background would be a handful in such challenging moments. However, the issue with such an approach is that it may introduce bias if not applied consistently. In saying so, an annotator’s intuition about the integration of one word may not coincide with the opinion of the reviewer or any other Kazakh-Russian bilingual, which creates subjectivity. This difficulty of proving the borrowedness of certain items is not unique to the present study: there is no consensus in the existing literature on where exactly borrowing ends and code-switching begins (Gardner-Chloros, 2009; Haspelmath, 2009). Thus, annotating in a conservative form, this study has labeled IUs for Kazakh, Russian, and Bilingual content. And lastly, following Shenk, this language annotation process has a fourth category: a truncation, sounds like ‘*aaa*’, ‘*ehh*’, and non-recognizable pauses, cuts of the word (as #2 in Table 3.1) that do not refer to any language, as well as minimal responses, ‘yes’, ‘no’, etc. Table 3.1 below illustrates the language annotation of the present study.

#	IU content	Language
1	<i>Endi salqyndap</i>	kz
2	<i>Po</i> , truncated, likely the start of the next IU	tr
3	<i>Pochti takaya zhe pogoda u nas</i>	ru
4	<i>Da</i> [ru] <i>mynandaj</i> [kz] <i>ma</i> [kz]	bil

Table 3.1: Language annotation of IUs from REC03.

3.2.2 Further clarifications

Several intonation units in the corpus required careful consideration during language identification, particularly where lexical items of non-Kazakh origin appeared within otherwise Kazakh-structured units. Since this study focuses specifically on Kazakh-Russian language contact, the annotation framework applied a principled boundary: only intonation units displaying Russian linguistic features, whether lexical, morphological, or phonological, were labelled as Russian (ru) or Bilingual (bil). The presence of loanwords or borrowings from other languages did not in itself trigger a non-Kazakh label. It is also worth noting the language annotation of IUs that contain words apart from Kazakh and Russian should be taken into consideration as well. For example, in REC02 there is a case *chalısty ğoy* ('studied'), annotated as Kazakh despite containing the verb *chalısty*, a slang term of Turkish origin among Nazarbayev University (NU) and many other, that entered the local Kazakh speech through multilingual student communities. Particularly among former Kazakh Turkic Lyceum students, majority of whom studied/study at NU. As mentioned earlier, three out of four recordings are communication of NU students with different styles of discussion (e.g story narration, casual talkings). Since *chalısty* carries no Russian features and has been integrated into the Kazakh from Turkish and *ğoy* is a Kazakh element, the intonation unit was treated as functionally Kazakh. The word's extralinguistic origin was considered irrelevant to the Kazakh-Russian contact framework under investigation, if it is not Russian.

Similarly, the unit *kaespyňĝo*, a phonologically nativised English abbreviation of CS (computer science major) was annotated as Kazakh as well. The abbreviation has been ab-

sorbed into the speakers’ Kazakh phonological system and inflected according to Kazakh morphological rules, making it a Kazakh-integrated form rather than an instance of English intrusion into the corpus. By contrast, *diskretka qursyn* was labeled bilingual because *diskretka* is a slang abbreviation of the subject name ‘discrete mathematics’, comparable in origin to *kaespynŋo*, both of which are informal academic terms derived from internationally used names. However, unlike *kaespynŋo*, it has the Russian suffix *-ka*, a morphological feature that directly indexes Russian contact, and following that, the word *diskretka* operationalizes as Russian, and *qursyn* is a Kazakh verb that created a bilingual annotation at the result. Overall, these cases illustrate that annotation remains theoretically coherent with the study’s focus on Kazakh-Russian contact by labeling IU content according to the presence or absence of Russian contact features.

3.3 Syntactic annotation

For this study, the syntactic classification follows Shenk’s (2006) groups, but it needed major changes. This is because the differences between Spanish-English and Kazakh-Russian pairs directly affect how syntax is annotated. Categories created for analytic languages’ pair by Shenk (clause, discourse marker, phrasal units, truncation, and other) did not easily fit the complex structures of Kazakh. For example, when Russian expresses subordination syntactically through finite clauses, Kazakh expresses the same function morphologically through converbs (non-finite verb forms that express subordination through suffixation but not through syntax) (Haspelmath, 2009; Muhamedowa, 2009). The following examples (12) and (13) further illustrate this structural difference:

- (12) Kogda on prišël, ja ušël.
when.CONJ he came.PST 1SG leave.PST
‘When he arrived, I left.’
- (13) Ol kel-gende, men šiq-tım.
he come-CONV.WHEN 1SG leave-PST.1SG
‘When he arrived, I left.’

In example (12), we can see that the Russian subordinate clause is introduced by the conjunction *kogda* and contains a fully inflected finite verb *prišël*. In example (13), the same statement is expressed by the converb suffix *-gende* attached directly to the verb stem *kel*, with no conjunction, which shows that the Kazakh construction functions syntactically as an adverbial phrase and lacks a finite verb. Applying Shenk’s clause category here would therefore be analytically misleading, as the construction lacks the finite predication that defines a clause in her framework. These converb constructions appear frequently in the data as independent IUs; an example of a converb IU from the corpus is given in example (14).

- (14) Sol üšin ajt-ıp
that for say-CONV
‘That is why I am saying.’

This IU contains no finite verb: the suffix *-ıp* in the verb marks a sequential converb (Ótött Kovács, 2015), making the IU syntactically dependent on a subsequent auxiliary verb *otırmyn* (following IU after example (14)) which functions as the progressive auxiliary. And the approximate translation to English would be something like, ‘That is why I am saying’ and *otırmyn* provides the finite predication. Since the converb construction cannot stand alone as a complete proposition, it was decided to subdivide Shenk’s clause category into two types for the purposes of this study: finite clauses and non-finite clauses, the latter of which was applied to constructions of this kind as in example (14). These categories are discussed in further detail below in sections 3.3.1 and 3.3.2 .

3.3.1 Finite clause

A finite clause (FC) is defined as an IU containing a fully inflected verb that bears tense and agreement marking, forming a syntactically and propositionally complete unit. Kazakh finite clauses take several forms depending on their syntactic function, including conditional, causal, concessive, and purposive clauses among others (Muhamedowa, 2016). A particularly useful feature for identifying finite clauses in this corpus is that, at the level of the IU, the expressed idea sounds complete and independent - that is, the

IU does not project a subsequent predicate to complete its meaning. Two examples from the corpus labeled as FC in the syntax category are shown below. Example (15) is a Kazakh IU from REC01:

- (15) Nege dep *sūramajsyn* *ğoj*.
 why QUOT ask.NEG.2SG PART
 ‘You don’t ask why, do you?’

The verb *sūramajsyn* (‘you don’t ask’) is fully inflected for negation (NEG), second person (2), and singular (SG). The IU expresses a complete proposition and is therefore annotated as FC. Not every FC IU in this corpus forms a context-independent statement.

- (16) Aljuminievj *qasyq* *dejdi*.
 aluminium.ADJ_{ru} spoon_{kz} say.3SG.PRES_{kz}
 ‘They say aluminium spoon.’

In example (16), a bilingual finite clause from REC01 begins with the Russian adjective *aliuminievji*. It is inserted into the Kazakh clause without changing its morphology, making the IU bilingual. Despite the mixed language, the finite verb *deidy* (‘they say’) makes the IU complete. Therefore, it is annotated as FC.

- (17) Na *pjatdesjat napisala*.
 PREP_{ru} fifty_{ru} write.PST.F_{ru}
 ‘Wrote for fifty.’

In example (17), a monolingual Russian IU contains the verb *napisala* that marks both past tense and feminine agreement. Thus, even though there is no clear subject, the subject agreement in Russian grammar indicates that the IU concerns that the speaker ‘wrote an exam for 50 points’, which forms a complete idea and is marked as an FC.

3.3.2 Non-finite clause

As noted by Ótrott Kovács (2015), Kazakh non-finite clauses (NFC) remain an under-researched area descriptively. By creating a corpus of non-finite clauses in Kazakh from

literary works, newspaper articles, and questionnaire data collected from native speakers in Kazakhstan, she concluded that Kazakh non-finite clauses are syntactically dependent constructions formed by attaching non-finite suffixes directly to the verb stem. These include converb suffixes, inflectional non-finite suffixes, etc. Syntactically, non-finite heads lack the tense and agreement marking that characterizes finite predicates (Ótrott Kovács, 2015). Following this framework, any IU that lacked a finite predicate and was syntactically dependent on a preceding or following IU for propositional completion was labeled as a non-finite clause in the present study. Examples from the corpus are given below.

- (18) Myna žaqty žina-ğanda.
 this side.ACC gather-CONV.TEMP
 ‘When gathering this side.’
- (19) Lučše žuma-j bar-ıp.
 better_{ru} wash-CONV.NEG_{kz} go.CONV_{kz}
 ‘Better go without washing.’
- (20) Eše Alualar awyr-ıp.
 still_{ru} Alua-ASSOC.PL_{kz} be.sick-CONV_{kz}
 ‘Alua and the others are sick.’

In IU (18), the suffix *-ğanda* is a temporal converb marking ‘when/while’, making this IU syntactically dependent on the following main clause predicate. IU (19) contains two converb forms - *žumaj*, marked with the negative converb suffix *-may*, and *barıp*, marked with the sequential converb suffix *-ıp*. Neither form carries finite tense or agreement marking. The IU expresses a manner or condition for an action but requires a following finite predicate to complete the proposition. In IU (20), the suffix is a sequential converb suffix, and despite the presence of the Russian temporal adverb *eše* (‘still’), the IU cannot express a complete proposition independently.

3.3.3 Phrasal units

Phrasal units (PU) have a range of syntactic subtypes, including noun phrases, postpositional phrases, and adjectival phrases, among others. In the present study, however,

phrasal units were annotated broadly as any verbless expression, without further sub-classification. This decision was made in the interest of analytical manageability, as the primary goal of syntactic annotation in this study is to identify codeswitching patterns rather than to provide an exhaustive syntactic analysis of each IU. Examples from the corpus are given in examples (21) to (23).

- (21) Muqym stolovoj-dyn.
 whole_{kz} canteen_{ru}-GEN_{kz}
 ‘The whole canteen’s / of the whole canteen.’

- (22) Segiz de üš normalno.
 eight_{kz} and_{kz} three_{kz} normally_{ru}
 ‘Eight and three — that’s fine.’

- (23) Eše men.
 also/still_{ru} 1SG_{kz}
 ‘And I also [...]’

Examples (21) and (22) illustrate PUs in which a Russian lexical item is inserted into a Kazakh nominal construction. In (21), the Russian noun *stolovoi* carries the Kazakh genitive suffix *-dyn*, integrating it morphologically into the Kazakh frame. In (22), the Russian adverb *normal’no* follows a Kazakh numeral sequence, creating a phrasal bilingual unit without a predicate. Example (23) illustrates a minimal bilingual PU.

3.3.4 Discourse markers

Discourse markers (DM) are words or short expressions that serve a discourse-pragmatic function rather than contributing to the propositional content of an utterance (Schiffrin, 1987). In the present study, they were classified as a separate syntactic category, distinct from clauses and phrasal units. Examples from the corpus include:

DM	Language	Approximate function
<i>nu</i>	ru	Hesitation/continuity ('well')
<i>tipa</i>	ru	Approximation ('sort of/like')
<i>koroche</i>	ru	Topic summary ('in short')
<i>söytip</i>	kz	Narrative continuation ('so')
<i>jañägi</i>	kz	Referential/anaphoric marker ('that one just mentioned')

Table 3.2: Examples of DM from the corpus.

3.3.5 Interactional intonation units

In this present study, Interactional IUs (IIUs) are backchannels, response tokens, and other short signals that manage the conversational floor rather than contributing propositional content. They differ from discourse markers in that they do not organize the discourse structure of the speaker’s own turn, but instead signal acknowledgement, receipt, or engagement with a preceding speaker’s contribution (Shenk, 2006). Examples from the corpus include minimal response tokens such as *mm*, *ia* ('yes'), *aaa* (a particle of recognition), *oibai* (an exclamative particle), and other truncated or paralinguistic sounds such as *brr* and etc. Since these items rarely contribute to propositional content and their primary function is interactional rather than referential, they were labeled as 'tr' (truncation) in language annotation.

3.3.6 Switch types: Embedded switch and Boundary switch

The last two annotations of the corpus are switch types and Switch-Boundary Intonation Units (SBIU). Although both help to identify points of codeswitching, they operate at different levels of analysis. Switch types: embedded and boundary were developed for the present study to capture code-switching occurring within IUs, while SBIUs (section 3.4), adopted from Shenk (2006), were used to capture code-switching between IUs. The internal codeswitching of IUs was classified into two main types. The first type, embedded switch (ES), follows Muysken’s (2000) concept of insertion, where a lexical item from one

language is placed into the grammatical frame of another. In the present study, an IU was labeled ES when a Russian element appeared surrounded by Kazakh material on both sides. As demonstrated by Auer and Muhamedova (2005) in their paper on Kazakh Russian code-switching, inserted Russian elements do not always receive host-language morphology, and this was taken into account in the annotation. The ES label was assigned based on positional embeddings rather than solely on morphological integration. Example (24) illustrates ES IU.

- (24) Endi moneta-lar stol-dyn üstinde turǵo.
 now_{kz} coin_{ru}-PL_{kz} table_{ru}-GEN_{kz} top-LOC_{kz} stand-PST_{kz}
 ‘Now the coins are standing on the table.’

Bound switches (BS) developed for this present study, on the contrary, occur at IU boundaries. In other words, IUs are annotated BS in when Russian words appear at the edges, either on the left or right, rather than within the main structure. In example (25), the IU starts with the Russian while the rest of the content in Kazakh:

- (25) Vsegda osyǵan isteimin men.
 always_{ru} this.DAT_{kz} do.1SG.PRES_{kz} 1SG_{kz}
 ‘I always do it like this.’

In this study, switch-type labels are used only for bilingual IUs. Internal IU switching, by definition, requires two languages, which occurs only in bilingual IUs; thus, monolingual IUs were not labeled for ES and BS switch types. In some intonation units, features of both ES and BS were simultaneously present. During the annotation, such conflicts raise the question of which label to assign. For example, the unit

- (26) Potom da oǵan vina zhauyp tastap
 then_{ru} and.PART_{ru} he.DAT_{kz} blame_{ru} put.upon.CONV_{kz} do.CONV_{kz}
 ‘And then went ahead and blamed him/her.’

opens IU with a Russian word, making the IU BS, but then, after the Kazakh pronoun, we also see an embedded Russian word, *vina*, and finally, IU is closed with a Kazakh converb. In such cases, a consistent decision rule was applied throughout the annotation:

whichever switch type appears first determines the label, which in this example is BS. This rule is applied consistently across all four recordings.

3.4 Switch-boundary intonation unit annotation

The identifier of switches, SBIU, marks the point of switch between intonation units. This was developed by Shenk, as her main argument is that bilingual speakers often switch languages at IU boundaries and need markers to point out those switches (shown in Table 3.3). She defines it as “the intonation unit directly preceding another intonation unit that is uttered in the other participating language” (Shenk, 2006). Thus, SBIU framing asks, “What kind of unit did the speaker just complete before switching?”

In short, annotating the corpus IUs with SBIU helps show how SBIU characterizes the conditions that lead to switching. It also reveals what makes a switch natural and interpretable to interlocutors.

Speaker	IU	Language	Syntax	SBIU
Speaker 1	<i>Közi uqsaydı qursın</i>	kz	FC	SBIU
	<i>s ochkami ne videla</i>	ru	FC	
Speaker 2	<i>On seychas</i>	ru	PU	SBIU
	<i>s ochkami khodit</i>	ru	FC	
Speaker 1	<i>Qoyshı</i>	kz	PU	

Table 3.3: Example of the SBIU from REC02.

This definition mainly instructs us on how to mark switching boundaries in Kazakh-Russian bilingual interaction. However, as noted several times, this project’s approach differs from Shenk’s due to language contact and typological differences. For example, bilingual IUs in this dataset account for 15.51% of all IUs and exhibit distinct switching patterns at boundaries. In contrast, Shenk’s corpus had only 4% bilingual IUs (Shenk, 2006, p. 189).

She defined SBIU as a switch occurring practically between two monolingual IUs in different languages. However, applying this definition to Kazakh-Russian data with almost 4 times more bilingual IU percentage has a practical problem. Thus, when using Kazakh-Russian dataset, there were challenges in strictly following her SBIU framework. To clarify how this research annotated SBIU, two real corpus examples are shown below: Figure 3.2 and Figure 3.4 show the algorithm on how it was annotated.

IU	IU content	Lang.	Annotation
IU1	<i>En</i> [kz]	kz	SBIU (switch from <i>eñ</i> to <i>uzhasnyy</i>)
IU2	<i>Uzhasnyy</i> [ru] <i>chuvstvo</i> [ru] <i>aytayyn</i> <i>ba</i> [kz] ?	bil	SBIU (switch from <i>aytayyn ba</i> to <i>ko-roche</i>)
IU3	<i>Koroche</i> [ru]	ru	

Figure 3.2: Example of SBIU annotation with boundary markers.

In Figure 3.2, *En* [kz] is annotated as SBIU because it is a monolingual Kazakh IU whose final element *en* is in direct contact with the Russian word *uzhasnyy* that opens the following IU. The language switch occurs precisely at the IU boundary.

IU	IU content	Lang.	Annotation
IU1	<i>Nu</i> [ru]	ru	SBIU (switch from <i>Nu</i> to <i>Mynandaj</i>)
IU2	<i>Mynandaj</i> [kz] <i>zherde</i> [kz] <i>otyr</i> [kz]	kz	SBIU (switch from <i>otyr</i> to <i>Vobshche</i>)
IU3	<i>Vobshche</i> [ru] <i>köröp</i> [kz] <i>otqan</i> [kz] <i>zhoq</i> [kz]	bil	

Figure 3.3: Example 2 for identifying SBIU.

Similarly, in the Figure 3.4 example, *Mynandaj* [kz] *zherde* [kz] *otyr* [kz] is annotated as SBIU because its final element *zhoq* is in direct contact with the Russian word *Vobshche*

that opens the following IU. Both third IUs in the examples are not annotated as SBIUs, as we don't know the content of the subsequent IUs. Lastly, following Shenk (2006) constraint IIUs were not labeled SBIU, as they mostly carry truncations, pauses, or simple phrases that are meaningless in the context. Thus, for this study, BS and ES are used to measure intra-IU language mixing, while SBIU is used to measure inter-IU switching, analogous to the distinction between intrasentential and intersentential switching in the traditional code-switching literature (Poplack, 1980). Table 3.4 illustrates this distinction, where BS and ES are only in bilingual IUs.

	IU	Language	SBIU	Switch types
1	<i>söyley berem</i>	kz		
2	<i>söylep jürmin</i>	kz	SBIU	
3	<i>I [ru] birkezde</i>	bil		BS
4	<i>iya</i>	kz		
5	<i>negizi</i>	kz	SBIU	
6	<i>i tak [ru] söyley berem ğo</i>	bil	SBIU	BS
7	<i>Prosto [ru]</i>	bil	SBIU	
8	<i>bilmeim</i>	kz		
9	<i>bir krugqa [ru] tüşkende</i>	bil	SBIU	ES
10	<i>prosto krugqa [ru] tüşkende</i>	bil		BS
11	<i>bir krugqa [ru] tüşsem</i>	bil		ES
12	<i>neyap qalam</i>	kz		

Table 3.4: The difference between SBIU and switch types (BS and ES).

While SBIU can be annotated as both a bilingual and a monolingual IU, marks only IUs where the language changes at their boundary. For example, the end of IU 2 is a Kazakh verb, and the beginning of IU 3 is a Russian conjunction, which marks a language switch at the IU boundary. Same with IU 5 and IU 6. From IU 6 to IU 7 and from IU 7 to IU 8. IU 10, even though it contains a Russian element, the switch is not at the IU boundary but within, thus it was labeled only to ES.

3.4.1 SBIU annotation levels

SBIU annotation was initially applied at the conversational level. Under this approach, an IU was marked as SBIU when it was immediately followed by an IU in a different

language, regardless of whether that following IU was produced by the same speaker or a different one. This approach was first motivated by Auer’s (1984) argument that language choice in bilingual conversation is not a property of individual speakers but of conversational sequences. More precisely, in his language-negotiation model:

$$A_i B_g A_i B_g \dots // A_i B_i A_i B_i \dots$$

where A and B are speakers and i and g denote different languages (Auer, 1984). As shown, the switch is not locatable within any single speaker’s turn but emerges from the sequence as a whole. So initial SBIU annotation followed this logic, treating the switching point as an interactional phenomenon located in the relationship between adjacent IUs rather than within an individual speaker’s production.

However, this approach does not account for turn-taking. Turn-taking refers to the sequential organization of conversation in which speakers alternate in producing talk: one speaker holds the floor, and another takes over at a transition-relevance place (Sacks et al., 1974). Auer’s framework addresses sequence-level language negotiation, not individual prosodic switching decisions (turn-taking), that the present study’s research question requires. A Kazakh IU from speaker A followed by a Russian IU from speaker B registers as a boundary switch under conversational-level annotation, but the switching signal is not traceable to either speaker’s prosodic production. That IU sequence is a turn transition, not a code-switch and therefore conversational-level annotation inflates the SBIU count and cannot establish whether IU boundaries triggered the switch or merely coincided with it.

To address this, a per-speaker annotation level was added. Under this approach, SBIU was marked only when the preceding and following IUs were produced by the same speaker. Additionally, a switch was not counted if a substantial time gap had spent since the speaker’s previous IU. Because long gap makes it unlikely that the two IUs form a continuous prosodic sequence that shows an active switching decision. To control such details, the start and end timestamps provided by the MULTICORSKL (Filchenko et al., 2023) corpus. Due to time constraints, per-speaker annotation was applied to

REC02 only, covering 622 IUs from the first half of that recording. This annotation level provides a more accurate measure of individual switching behavior and is the basis for the speaker-level results reported in section 4.4.

3.5 Data processing and quantification

All annotations were conducted in Microsoft Excel. Character, syllable, and word counts of all IUs' mean sizes were computed automatically using text functions (AVERAGE). Mean sizes for ES and BS subsets were computed using the AVERAGEIF function, conditioning on the switch-type column. Along with that, for the syntactic composition counts (section 4.2) and the SBIU-syntax alignment counts (section 4.3.3), the COUNTIF and COUNTIFS functions were applied to the syntactic category and SBIU columns, respectively. Raw counts were mostly done through filter function. Truncation IUs (tr) were excluded from all syntactic counts and from SBIU assessment, as described in section 3.3.5. And all proportions reported in the findings chapter are calculated as (count of category / total applicable IUs) $\times$ 100.

.

Chapter 4

Findings

4.1 Dataset composition

This chapter presents the results of a quantitative analysis of IUs in recordings from the MULTICORSKL (Filchenko et al., 2023). The data examined here consist of four recordings and 3,796 IUs. Table 4.1 provides a detailed summary of the dataset, including language-annotated IU counts for each recording.

REC/Language	kz	ru	bil	tr	Total
REC01	654	11	69	35	769
REC02	744	101	254	67	1166
REC03	669	145	168	78	1060
REC04	473	194	98	36	801
Total	2540	451	589	216	3796

Table 4.1: Language-annotated IU counts across four recordings.

As shown in Table 4.1, Kazakh is the main language in the corpus, accounting for 66.9% of all IUs. REC01 is the most Kazakh-dominant, with 85.1% of its IUs in Kazakh-only, while Russian monolingual IUs make up just 1.4% and bilingual IUs make up 9.0%. REC02 and REC03 fall in between, with Kazakh at 63.8% and 63.1%, and Russian at 8.7% and 13.7%. The similarity between REC02 and REC03 suggests these younger speak-

ers show a comparable level of code-switching. In comparison to the Kazakh-dominant older speakers of REC01 and the Russian-dominated speech of REC04 both REC02 and REC03 are in the middle between contrasting earlier mentioned recordings.

One more highlight is that, across all recordings, bilingual IUs exceed Russian monolingual IUs, except for REC04, which has the highest share of Russian-dominant IUs. The data also includes the number of truncated IUs, which is the smallest share of IUs in each recording, as they carry no content but only pauses and shortened IUs. Interestingly, REC03, the recording with the most linguistic diversity and balanced proportion of Russian and bilingual IUs, also has the highest number of truncated IUs (7.4%). This is higher than REC02 (5.8%), REC01 (4.6%), and REC04 (4.5%). The higher truncation rate in REC03 possibly suggest speakers' greater processing demands in speech where Russian and bilingual IUs are more evenly distributed, which is consistent with accounts of increased cognitive load in balanced bilingual production (Chafe, 1994).

REC	Total IU	ES	BS
<i>Characters</i>			
REC01	12.61	22.39	17.30
REC02	12.67	23.54	17.19
REC03	11.98	25.65	16.84
REC04	12.75	23.00	18.32
<i>Syllables</i>			
REC01	5.28	9.17	7.14
REC02	5.27	9.80	7.24
REC03	4.90	10.45	6.93
REC04	5.27	9.45	7.54
<i>Words</i>			
REC01	2.59	4.17	3.27
REC02	2.69	4.46	3.51
REC03	2.49	4.84	3.38
REC04	2.64	4.25	3.60

Table 4.2: Mean IU size by recording and by switch IU (ES and BS).

The corpus initially contained column data for characters, syllables, and words for each IU. Therefore, the mean IU size of the dataset for the current study was calculated as

given above. Table 4.2 shows that each metric approximately has the across-recordings range; for example, the average total IU characters range from 11.98 to 12.75, while syllable counts range from 4.90 to 5.28, and word counts range from 2.49 to 2.69. Although the numbers show no large fluctuations, REC03’s total IU results are the smallest in every category compared to the other recordings’ averages for total and BS IUs, except for ES IUs, which are the longest. Generally, ES IUs are consistently larger than BS IUs across all recordings and all three measures. ES IUs ranged from 4.17 to 4.84 words, while BS IUs ranged from 3.27 to 3.60 words, with almost one word shorter.

4.2 Syntactic composition of the dataset

A distribution of the syntactic categories across the four recordings reveals a consistent pattern of language use throughout the corpus. As illustrated in Figure 4.1, the five

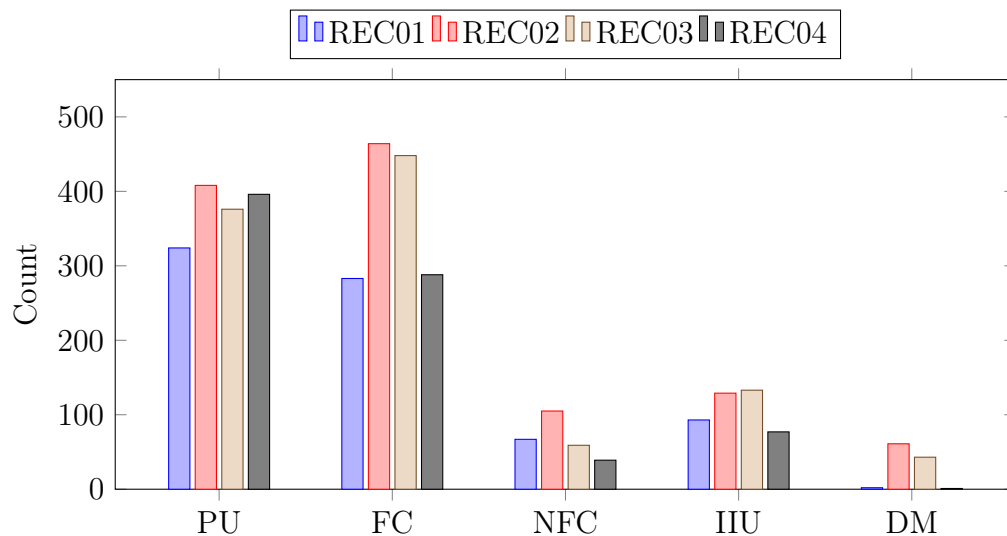


Figure 4.1: Syntactic composition of the corpus.

syntactic categories show distinct frequencies but a clear and consistent distributional hierarchy within each other. Among all categories, finite clauses (FC) and phrasal units (PU) accounted for approximately 79% of all IUs in the corpus. However, phrasal units still exceed the number of FC, making the most frequently occurring category overall, with 1,504 IUs while FC number is 1,483 IUs. The almost equal distribution of PU and

FC suggests that speech in this corpus divides evenly between propositionally complete clausal units and verbless expressions (Shenk, 2006). In Figure 4.1, FC appears more than PU, but only REC02 shows the highest number of FCs; however, overall, PUs outnumber FCs in the dataset. IIUs and NFC are secondary groups by frequency. Compared to the numbers of primary groups, they are present at considerably lower rates. However, their frequency is consistent across all recordings. For example, across recordings, 10%-13% of IUs are consistently interactional, suggesting a stable and functional role within the syntactic organization of the dataset. Despite being the least frequent of the four categories, non-finite clauses also consistently occur, but never surpass 9% in any single recording. In contrast to previous groups, the most unstable frequency across five categories was noticed in the distribution of discourse markers. DM is almost absent in REC01 and REC04, with only 2 and 1 IUs recorded, respectively. But it appears 61 and 43 times in REC02 and REC03, respectively. This represents a frequency more than 20 times higher than in REC01 and REC04. REC02 and REC03 are narrative and story sharing recordings among younger speakers, where linkers and fillers are more frequent by the nature of the interaction, which can explain their higher amount in comparison to REC01 and REC04 which have less storytelling content.

4.3 SBIU, BS, and ES and their syntax alignment

4.3.1 SBIU

SBIU is a main measure for identifying switch points in this study. Table 4.3 below shows the distribution of SBIUs across 4 recordings. The total number of SBIUs identified in the corpus is 903 and it makes up 23.8% of all IUs. In this general conversational annotation level, the distribution of SBIUs shows the overall Russian presence in each recording. For example, REC01, the most Kazakh-dominant recording, has the lowest SBIU rate at 9.1%, while REC02 shows the highest rate at 32.4%. And REC03 and REC04 fall in between, with rates of 24.8% and 24.0%, respectively.

REC	Total IUs	SBIU count
REC01	769	70
REC02	1166	378
REC03	1060	263
REC04	801	192
Total	3796	903

Table 4.3: SBIU count in each recording.

4.3.2 Switch types: Embedded switch and Boundary switch

As shown in Figure 4.2 the distribution across switch types varied with high imbalance. The BS number is considerably higher than ES, and this trend holds across recordings,

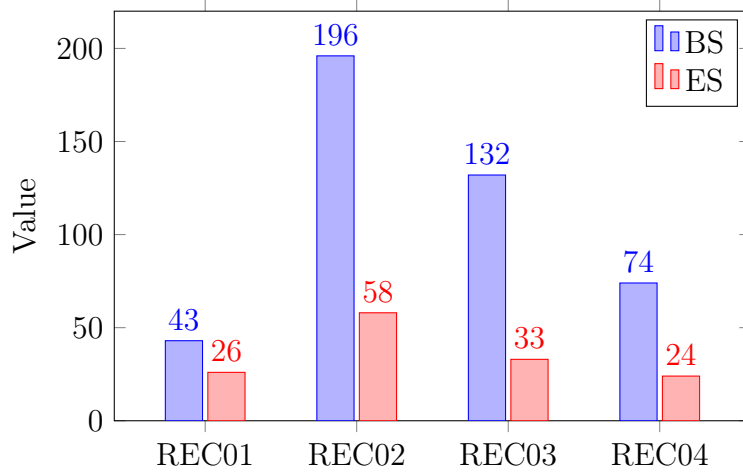


Figure 4.2: BS and ES counts per recording

rising steadily from REC01 (64.6%) through REC02 (75.7%) and REC03 (86.8%) to REC04 (90.6%). At the same time, the ES proportion (ES:total bil) drops from 35.4% in REC01 to 9.4% in REC04. The steady rise in BS and fall in ES from REC01 to REC04 suggests that as Russian becomes more dominant in a recording, switches increasingly occur at IU edges rather than internally, possibly because more Russian-origin items are functioning as active switches rather than integrated loans in those contexts. In terms of ES REC01 stands out because it has a relatively high count (23) compared to its total switch count (65). Consequently, it has the highest ES proportion of 35.4%. This is more than three times higher than in the REC04 percentage of ES. The BS-to-ES ratio also

varies: REC01 is roughly 2:1, while REC02, REC03, and REC04 are approximately 3:1, which means that the gap between BS and ES is also consistent in recordings with more overall switching. To calculate whether code-switching is happening at IU boundaries, this study used the above-mentioned 3 switch metrics (SBIU, BS and ES). An event was counted as a boundary switch if it was annotated as SBIU, meaning the language change occurred at the edge between two IUs. An event was counted as a non-boundary switch if it was annotated as ES, meaning the Russian element was embedded inside a bilingual IU, surrounded by Kazakh on both sides. Since every switching event in the dataset falls into one of these two categories, the proportion of boundary switching was calculated as $SBIU / (SBIU + ES)$.

Across all four recordings, 86.5% of switches occurred at an IU boundary. This means that in over eight out of ten switching events across the dataset, the language change occurred at an IU boundary rather than within one, supporting the study’s central hypothesis that IU boundaries are the preferred site for code-switching in Kazakh-Russian bilingual speech. REC02 had the highest rate at 86.1%, followed by REC04 at 88.9% and REC03 at 89.5%. REC01 was the lowest at 75.3%.

REC	BS	ES
REC01	43	26
REC02	196	58
REC03	135	33
REC04	74	24

Table 4.4: Switch type distribution across recordings.

To observe the same pattern at the intra-IU level, the calculation used BS and ES only. Because BS and ES are the only metrics that show the position of the Russian element in IU. Therefore $BS / (BS + ES)$ reveals the time it appears at the edge rather than the interior. As a result, within bilingual IUs, 76.2% of Russian elements appeared at the IU edge (BS) rather than the interior (ES). REC03 had the highest rate at 81.2%, followed by REC02 at 76.0% and REC04 at 75.8%. REC01 was the lowest at 64.6%. Across all four recordings, boundary-adjacent placement consistently outnumbered internal embedding.

The fact that this boundary preference holds even within already-mixed IUs show that edge placement reflects a property of how speakers distribute different languages within a single prosodic unit.

4.3.3 SBIU-syntax alignment

The SBIU-syntax alignment counting represents the most challenging part of the annotation, particularly when applied to high-contact language pairs such as Kazakh-Russian. The bar chart in Figure 4.3 illustrates that FC and PUs account for the great majority of SBIUs in the corpus.

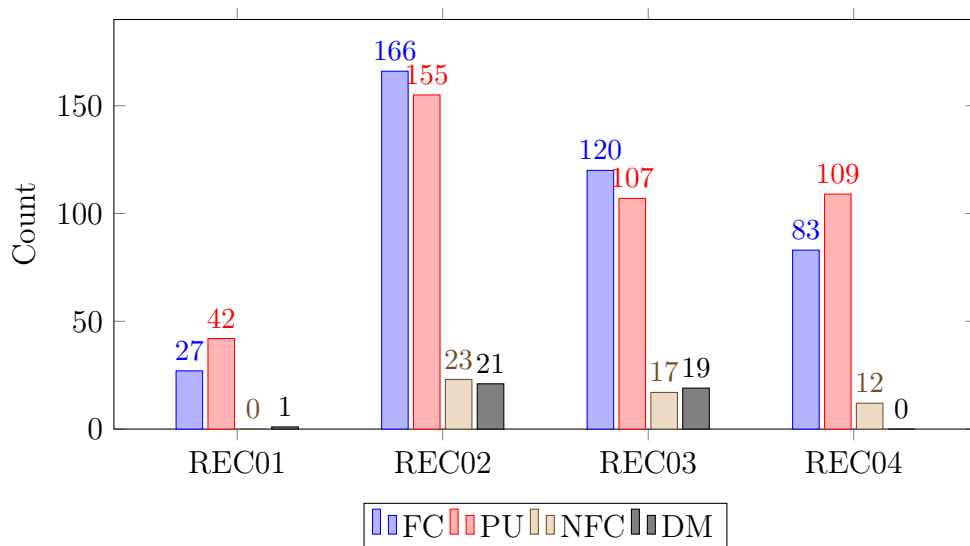


Figure 4.3: SBIU type distribution across recordings.

To be precise, FC aligned with 396 SBIU and PU with 413 SBIU instances combined across all recordings. Together, these two categories represent approximately 80% of all switching boundary sites. This concentration partly reflects the overall frequency of FC and PU in the dataset, but the near-equal split between the two categories at switching boundaries also anticipates a distinction that becomes apparent in the ES and BS breakdown below in section 4.3.4, that FC and PU align with switching through different structural mechanisms.

Despite the almost equal number of SBIU for both, PU still outnumbers FC by 17 ad-

ditional IU sites. In comparison, non-finite clauses appear at SBIU sites at considerably lower rates, with approximately 52 instances across the three recordings in which they were annotated. Discourse markers represent the smallest category at SBIU sites, appearing only in REC02 and REC03, consistent with their overall low frequency in the dataset.

4.3.4 BS and ES syntax alignment

The most popular syntax category in the dataset that aligns with switch types is the finite clause, accounting for 58.9% and 70.4% in BS and ES types, respectively. The higher FC share in ES than in BS indicates that finite clauses not only host switching at their boundaries but accommodate internal Russian embedding at a proportionally higher rate than any other syntactic category, suggesting that structural completion makes finite clauses the most permissive environment for intra-IU mixing.

The next category is phrasal units, which were considerably higher in BS (154) than in ES (27). The NFC has the same frequency distribution in the BS and ES switches as PU, but is less popular than the previous categories by almost 4 times. No IIUs were included in this review since they were all labeled as truncations, and as mentioned earlier, only bilingual IUs had switch types.

An interesting case was found in the DM, which showed only 4 BS instances and no ES instances across the whole corpus. The complete absence of ES in discourse marker IUs indicates that Russian-origin discourse markers never appear embedded within a Kazakh frame on both sides because they occur only at IU edges, a pattern examined further in section 5.3.3. All discourse marker instances are Kazakh-Russian mixed connectives and the reason can be the small size of the DMs in the dataset compared to other syntactic categories etc. Other reasons for such a result can also be explained more prominently in the discussion chapter (see section 5.3.3).

Figure 4.4 demonstrates the distribution of switch types in each syntactic category per recording. The main pattern is that BS are more frequent than ES across all syntactic categories and recordings. Additionally, we can see that FC and PU account for most of

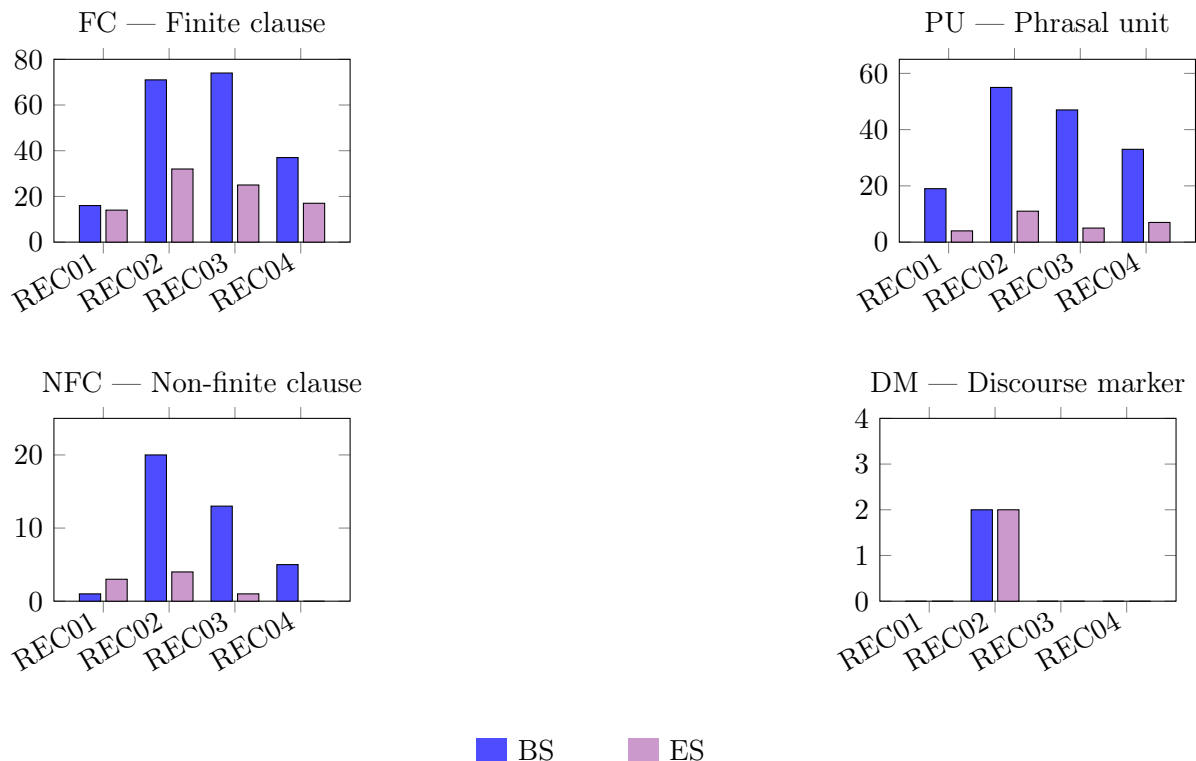


Figure 4.4: Syntactic alignment of BS and ES switch types across recordings.

the BS and ES switches, which makes sense since these categories appear most often in the data. However, as shown in the bar charts, PU alignment is about 3.3 times lower than FC in ES, whereas in BS it is only 1.3 times lower than FC. So, in ES switches, the drop from FC to PU is much steeper, whereas in BS, the two categories are close. This suggests that phrasal units are considerably less prone to internal-IU language mixing than finite clauses but are nearly as likely as finite clauses to serve as boundary-level switching points. Non-finite clauses have very few ES switches in all recordings. There is only one in REC03 and none in REC04.

4.4 Per-speaker SBIU

To address the methodological issue in section 3.4.1, a corrective re-annotation of SBIU was conducted on the first 622 IUs of REC02, as this recording had the largest number of SBIU, which would be informative to test. Table 4.5 represents language distribution

of each speaker. This annotation covers three speaker IUs: Speaker A (181 IUs), Speaker B (381 IUs), and Speaker C (60 IUs). Truncations are not listed in the table.

Speaker	kz	ru	bil	Total
A	121	12	43	181
B	275	22	60	381
C	31	4	15	60

Table 4.5: Language distribution of REC02 chunk.

And Figure 4.5 shows the distribution of BS and ES in each speaker’s IU.

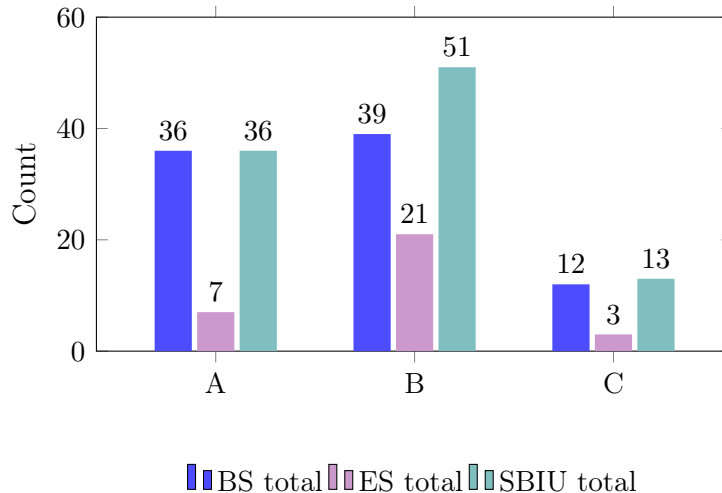


Figure 4.5: BS, ES, and SBIU totals per speaker in REC02 chunk.

Following the formula in section 4.3.2, it was calculated that across all switching events, 76.2% occurred at an IU boundary (SBIU) rather than inside one (ES). In the same way, the percentage of boundary switches in each speaker was calculated. Thus, Speaker A had 83.7%, Speaker B 70.8%, and Speaker C 81.3%. All three speakers exceed 70%, meaning boundary preference is consistent at the individual level regardless of each speaker’s overall switching rate, which confirms that the conversational-level SBIU pattern is not an indicator of speaker’s CS. At the intra-IU level, results of the calculation show that Speaker A had 83.7%, Speaker B 65%, and Speaker C 80%, making the total 73.7%.

Speaker B’s lower rate of 65% compared to Speakers A and C suggests this speaker produces proportionally more internal Russian embedding within bilingual IUs, making their speech the strongest case for examining intra-IU switching as a separate term for this study.

4.4.1 SBIU alignment with syntax

As shown in Figure 4.6, FC dominates across in all three speakers, followed by PU.

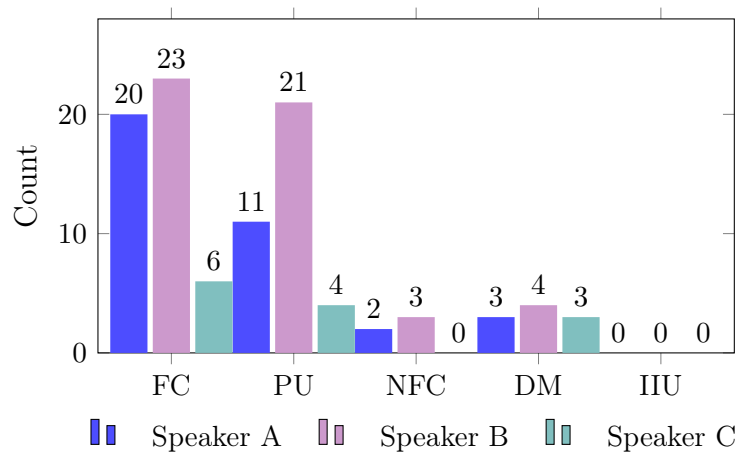


Figure 4.6: SBIU syntactic category distribution per speaker in REC02 chunk.

Following the pattern as in conversational level annotated data, NFC and DM show the lowest number in all cases. IIU contributes nothing as it was not annotated for switch types from the start. Speaker B shows the highest PU count (21), which is close to their FC count (23), more balanced than Speakers A and C. Speaker B’s more even spread across FC and PU at switching boundaries, compared to the FC dominance seen in Speakers A and C, suggests individual differences in the syntactic environments to which speakers switch most often, even if the general preference for boundaries remains unchanged. And as can be seen, Speaker C has the smallest counts overall, which shows smaller switching rate.

Chapter 5

Discussion

5.1 Dataset composition

5.1.1 Bilingual vs monolingual IUs

The consistent relationship between bilingual and Russian monolingual IUs is shown in Table 4.1, where across REC01, REC02, and REC03, bilingual IUs outnumber Russian monolingual IUs in every recording. This suggests that in three of the four recordings, speakers prefer to switch languages within a single IU rather than produce a fully Russian IU. This particular finding shows the high level of contact between the languages. Such observation once again confirms that, the Russian language has been integrated into the Kazakh language deeply enough, that mixed IU is a more common result of language contact, rather than a pure alternation between monolingual IUs. However such tendency has not been observed in REC04, which is the only exception where one of the speakers speaks mostly in Russian.

One such finding, as shown in Table 4.1, is that REC03 demonstrates the most balanced bilingual data (ru and bil IUs) and the highest number of truncated IUs. And this rate is consistent with cognitive load accounts of bilingual speech production. Torres Cacoullos and Travis (2018) state that speakers switching between two languages within IUs may require more processing time and experience greater hesitation. Thus, the high number of

truncations and high number of Russian elements containing IUs (both pure and mixed) shows that the statement is consistent with the IU examination of present study.

Additionally, it can be noted that the four recordings in this dataset were selected in a way that creates a gradient, not recordings with the same content. For example, Kazakh IUs gradually decline moving from REC01 to REC04 (85.5% to 59.2%), while Russian monolingual IUs rise from 1.4% to 24.2%. Thus, given the diversity of recordings in terms of bilingual IU content, the results discussed in the following paragraphs show firm conclusions. Furthermore, even though truncations are most of the time excluded from primary counts and data analysis, their presence in real-life speech remains important to the findings above.

5.1.2 IU length

As shown in Table 4.2, the consistency of total IU size across all four recordings was the same. The average per IU was approximately 12 characters, 5 syllables, and 2.5 words. The consistency of results shown in Table 4.2 across all IUs and switch type/bilingual IUs, shows that IU size remains a consistent prosodic unit feature across speakers, recordings, and degrees of code-switching.

These results align with the main theoretical framework of IU discussed in this study’s literature review. Chafe (1994) observed that IU represents a cognitive unit with approximately fixed informational capacity. He notes that, generally, the average in English speech is 4 words per IU. In the Kazakh-Russian bilingual corpus, however, it ranged from 2.49 to 2.69 words per IU. And such difference can be explained by Kazakh’s agglutinative morphology, where a single verb form in Kazakh can encode tense, agreement, negation, and aspect (Muhamedowa, 2016). As a result, propositional content is compressed into fewer words than would be required in English.

According to the same label, embedded switch (ES) IUs are consistently 25–50% larger than boundary switch (BS) IUs across all recordings and measures. Both are bilingual IUs: ES IUs require Russian material surrounded by Kazakh material on both sides (Kazakh-Russian-Kazakh), while BS IUs only require a single Russian word at a Kazakh

utterance boundary. Observing the structure of both, there is a simpler reason ES IUs are longer: they require more words by its definition. For example,

- (27) Mīnau tochno pīskēn
 this.DEM_{kz} definitely.ADV_{ru} cooked.PART_{kz}
 ‘This is definitely cooked.’

will always produce longer IUs than a BS, such as,

- (28) mēn tam.
 1SG_{kz} there.ADV_{ru}
 ‘I am there...’

Simply put, the length difference between ES and BS IUs can be explained by their structures. ES IUs are longer because, by definition, a Russian element surrounded by Kazakh material on both sides requires at least three words, the Russian item plus Kazakh material on each side.

As shown in Figure 4.4, this is consistent with ES mostly aligning with FCs and PUs, which are structurally complete units, the type of IU where internal switching is most frequently observed in the dataset. But for a while, NFCs, despite being shorter units, resist it. In this sense, IU length demonstrates the syntactic context in which switching occurs (more in section 5.3.2).

5.2 IU boundaries

5.2.1 Inter-IU switching: SBIU evidence

Across all four recordings, 903 SBIUs were identified using the conversational-level annotation method. And their results showed 75.3%, 86.1%, 89.5%, and 88.9% switching preferences in REC01, REC02, REC03, and REC04, respectively. These results suggest a strong surface tendency for switches to coincide with IU boundaries. However, as discussed in section 3.4.1, the conversational level method of annotation would not distinguish whether the IU boundary is the site at which switching events tend to happen, but

would only reveal the surface pattern of CS through IUs' organization. Such an approach does not trace the production of CS or whether the IU aligns with CS sites, because a Kazakh IU from speaker 1 may be followed by a Russian IU from speaker 2, which is not CS within a discourse. Therefore, a corrective re-annotation at the individual speaker level would address this directly; the findings presented in section 4.4 will be discussed further.

The results of corrective annotation show boundary preference of 76.3% overall, ranging from 70.8% to 83.7% across speakers. Shenk's work, which served as the basis for the present study, found that speakers in her Spanish-English corpus switched at IU boundaries 96% of the time. The gap between 96% and 76.3% is worth discussing as well. One contributing factor to this difference is the number of bilingual IUs, which reflects the contact situation. At both the conversational and speaker levels, bilingual IUs account for 15.31% and 19%, respectively. Such findings of high bilingual IUs across recordings and speakers suggest it is a feature of how these speakers have internalized code-switching, as written in Chernyavskaya and Zharkynbekova (2024), who claim that language switching in modern Kazakhstan is part of usual communication rather than an unusual case.

Torres Cacoullos and Travis (2018), observe that switching tends to align with prosodic boundaries and associate this co-occurrence with reduced cognitive load. In the present Kazakh-Russian data, Russian is available throughout the conversation, which is consistent with switches appearing at a wider range of positions, including mid-IU, without the same cognitive cost.

Even in dense contact and extensive borrowing between languages (all borrowings annotated as Russian, with no exceptions), speakers mostly switch to boundaries three-quarters of the time. As the results of the corrective annotation across all three speakers indicate, IU boundaries are a consistent site of code-switching as Speaker A (83.7%), B (70.8%), and C (81.3%), and no speaker approaches the 50% mark. And Speaker B's comparatively lower rate (70.8%) shows that high contact between Kazakh and Russian with embedded switch IUs occurs more often than in other cases, highlighting the importance of examining intra-IU switching in this study. (section 5.2.2)

In Kazakh-Russian bilingual speech, Russian items function as both active switches and fully integrated loans (Chernyavskaya and Zharkynbekova, 2024). For example, in the latter case, a speaker retrieving *kompot* or *yolkamyz* (section 5.4.1) within a Kazakh IU has no cognitive reason to impose a prosodic boundary before it. The result is a lower boundary preference rate, not because IU boundaries are less relevant as switching sites, but because the density of contact is consistent with switches appearing at a wider range of positions, not exclusively at IU boundaries. The 76% is therefore best explained as a contact-conditioned result of the boundary preference. The syntactic distribution of the remaining 23.7% of non-boundary switches (ES), as observed in Figure 4.4 and discussed in Section 5.2.2, mostly aligns with finite clauses and phrasal units. They are, compared to non-finite and discourse markers, structurally independent enough to accommodate an internal Russian element without disrupting the IU’s coherence as a cognitive unit. And only because of form-based language annotation, even though those ES IUs mostly contain borrowed Russian words, it reduced the number of boundary switches.

This means that the only boundary-preference percentage cannot explain the reasons for such patterns. Therefore, syntax provides a complementary explanation. Saying so, when a switch does not occur at an IU boundary, the syntactic structure of the IU is associated with those non-boundary positions. Section 5.2.2 examines this syntactic distribution in detail, showing that intra-IU switching is a structurally conditioned phenomenon.

5.2.2 Intra-IU switching: BS and ES evidence

Unlike SBIU counts, the findings for switch types do not inflate results across recordings, because each IU’s content is independent of the others. This discussion, therefore, includes results at both the across-recording and across-speaker levels.

At 73.7% overall across speakers and 76.2% across recordings, the boundary is the preferred position even within bilingual IUs. Even in IUs that are already mixed, the Russian element tends to sit at the edge rather than the interior.

This observation rules out additional factors that might contribute to an SBIU preference claim. For example, at the SBIU level, boundary alignment could, in principle, reflect

external factors such as speaker alternation, topic shifts, or discourse organization. However, the intra-IU level examination eliminates it all. Intra-IU examination is within a single bilingual IU, produced by a single speaker in a single prosodic unit. And even in such an independent chunk, where the above-mentioned influences of the topic or other speakers cannot always govern, the Russian element still appears at the edge rather than the interior. The counts in the dataset, across recordings (448/589 vs 141/589) and across speaker level (87/118 vs 31/118), confirm that the number of BS always outnumbered the number of ES in bilingual IUs. Other turn-taking or discourse-level explanations cannot explain such a high ratio of BS. The intra-IU boundary preference, therefore, shows how individual speakers package language-specific material within a mini-unit. The SBIU results reflect this same pressure operating at a larger scale.

For example, in speaker-level SBIU annotation, Speaker A’s SBIU is the same as her BS IUs 36 and 36, which means that in all her switches, she is switching at the boundaries, as shown in Figure 4.5. The same is true for Speaker C, who has a total of 15 bilinguals, 13 SBIUs, and 12 BS, meaning she is always switching between languages at the boundary except once.

Referring to the IU length explanation of switch types: ES IUs are consistently longer than BS IUs across all four recordings, averaging approximately 1 full word longer. Apart from the above-mentioned speaker’s tendency to switch IU boundaries both at inter- and intra-IU levels, such an abundance of BS may be related to the mean IU length of 2.5 words, as edge positions are already statistically frequent by chance; in a three-word IU, two or three positions are edges. Internal embedding requires at least three words to place the Russian element in the center. And the consistent 3:1 ratio of BS to ES, regardless of IU length, cannot be explained by chance alone. Taken together, the SBIU, BS, and ES data confirm that IU boundaries are consistent and systematic sites for code-switching in this corpus.

5.3 Syntactic structure of the dataset

5.3.1 Finite clauses vs phrasal units

Syntactic categories can indicate code-switching in two ways when aligned with IUs. Firstly, it determines the likelihood of a switch at any given boundary. Secondly, it determines what kind of switch the boundary permits. The distribution of finite clauses (FC) with phrasal units (PU) can illustrate these points by aligning switch metrics (SBIU, BS, and ES). A finite clause of the present study, as mentioned in methodology, is annotated FC when the IU ends with a fully inflected verb. Having said that the verb of the IU content should carry the tense, agreement, and negation, and resolving every internal dependency (Muhammedowa, 2016; Ótrott Kovács, 2015). Mostly, after such IUs, nothing projects forward into the next IU, so that the boundary after an FC is structurally free, and switching co-occurs with this position more frequently than with structurally dependent sites (Shenk, 2006; Torres Cacoullós and Travis, 2019). In the conversational level SBIU annotated dataset, 396 FC SBIU instances across all recordings, representing a second large share of all switching boundary sites. However, unlike conversational level SBIU annotation, speaker level annotated SBIUs primarily align with FC as shown in Figure 4.6, for all speakers.

Unlike finite clauses, phrasal units demonstrate switching through not completion but through detachment. PU 413 slightly outnumbers FC as a switching boundary site overall. And the reason is the broad scope of what is described under phrasal units. Namely, noun, postpositional, and adjectival phrases, the subdivisions of PU, are frequently placed at the edge of clausal structure e.g., topics, tails, etc.(Chafe, 1994). These positions are already syntactically free from the core, making them transfer points without interrupting any ongoing dependency. At the speaker level, PU alignment with SBIU is consistent across all three speakers, though it never exceeds FC in any individual speaker.

Such high alignment with SBIU results can be disputed with the initial high number of FC and PU in the dataset. But the intra-IU switch rules out this by showing more information. At BS sites, the FC-to-PU ratio is 1.3:1, and phrasal units are equal as finite

clauses tend to precede a boundary switch. However, at ES sites, the ratio considerably increases to 3.3:1 in favor of FC, which means that FC can accommodate internal Russian elements more frequently than PU. And this distinction of ES to be more aligned with FC than PU shows that these two categories license switching through different mechanisms: FC through completion, which creates room for both boundary and internal switching, and PU through detachment, which allows exit but not entry. What the per-speaker data shows is not about this FC vs PU contrast, but about what happens inside FC specifically when internal embedding is structurally available, but the speaker still switches the boundary. Among Speaker A’s finite clauses, every bilingual FC is labeled BS with no ES cases. For example:

- (29) I tak söyley berem ğo
COORD_{ru} so.ADV_{ru} speak-CONV_{kz} give-PRES.1SG_{kz} PART_{kz}
‘And so, I just keep talking anyway.’
- (30) Će-to ne-to bar ğo ana Ayaz atamen bar emes pa
something NEG.that exist PART that Ayaz ata.COM exist NEG Q
‘Something feels off — wasn’t there something with Grandpa Ayaz?’

Thus, the syntax’s structural permission and the behavioral default are not the same thing. Because, as it is seen, FC permits internal embedding, but the default is still BS. Russian material tends to appear at the IU edge even when the syntax could accommodate it internally, and the dataset results show this conclusion.

5.3.2 Non-finite clauses

The non-finite clauses are a new category in such a framework of observing IU syntax. They are, as mentioned in the methodology section, syntactically dependent constructions whose structural position is defined by their relationship to a following finite predicate. And such dependency contributes interesting insights into the study. Because of their dependence, they are not frequent in the dataset as an IU, generally.

Internal code-switching is rare in non-finite clauses and instead concentrate switching at their boundaries, a pattern that follows directly from their structural properties. As we

can see, across the entire dataset, ES switches occur in NFC annotation 3 and 7, and only once in REC01, REC02, and REC03, respectively. And not at all in REC04 (see Figure 4.4).

As the Matrix Language Frame (MLF) model suggests, the morphosyntactic frame of a bilingual speech is set by the matrix language, which provides the morphemes and governs the placement of embedded language material (Myers-Scotton, 1993). In Kazakh-Russian speech, Kazakh functions as the matrix language in non-finite clauses, and its non-finite suffixes, such as the converbs described in section 3.3.2, have no Russian equivalents. Saying that, every internal position within these clauses carries active morphosyntactic dependencies that Russian, as a fusional language, may not easily meet (Ótót Kovács, 2015). The results also support Muysken’s Muysken argument that structural differences between contact languages constrains mixing. Therefore, in those little cases where Russian-origin material does appear internal to a non-finite clause, it means that this element is systematically integrated into Kazakh morphophonology: forms like *zvan-dap* (Muhamedowa, 2016), changed to *-da* through regular voicing assimilation after the sonorant, with converb suffix *-(i)p* in example:

- (31) Srochno papağa zvandap
urgently_{ru} dad_{ru}-DAT_{kz} call_{ru}-VBLZ_{kz}-CONV_{kz}
‘Urgently calling dad...’

or nouns like *papa* in (31) and *krug* in examples (32)

- (32) Bir krugqa tüskende
one_{kz} circle_{ru}-DAT_{kz} fall-NMLZ-LOC_{kz}
‘When falling into one circle.’

used within a Kazakh syntactic frame reflects not active switching but borrowing and would be annotated as Kazakh IU in the absence of conservative language annotation of this study’s methodology.

5.3.3 Discourse markers

The distribution of discourse markers across the dataset have tendencies. First, across all recordings, Kazakh DMs outnumber Russian DMs, which means that even in high-contact speech, the discourse markers tend to stay Kazakh. The content of only four IUs is annotated as both DM and bilingual such as:

- a. *söytiip* [kz] *etot* [ru]
- b. *i* [ru] *söytiip* [kz]
- c. *vot* [ru] *anau* [kz]
- d. *tak* [ru] *myna* [kz]

also confirming the previous statement. Because all four are BS – Russian material at the initial edge, the Kazakh predicate closes the IU, which also aligns with a pattern established in 5.2.2: even in the smallest and most pragmatically lightweight bilingual IUs, the Kazakh frame holds and the Russian element stays at the boundary. Second, interestingly, REC01 and REC04 each contain three IUs, while REC02 contains 43 and REC03 contains 61. Such a large gap between the numbers suggests paying closer attention to what is happening in the recordings. Apparently, both REC02 and REC03, being narrative-heavy storytelling exchanges among university students/friends, showed the context-sensitive feature of DM IUs. Namely, the high number of linkers and fillers used for connecting the sequence of the events automatically ensures the appearance of discourse marker based words and constructions in the speech.

Despite their smaller number as a single word in the dataset, they attach to finite and non-finite clauses and phrasal units, and discourse markers performs different pragmatic work depending on its position within the IU. And raw counts do not capture the fact that DMs rarely constitute IUs on their own, for example only the count of Russian origin coordinator *i* (‘and’) attached to clauses and other units approximately contains 50 IUs. Such an appearance of DMs is consistent with Matras (2000), who states that discourse markers and pragmatic particles are among the earliest items to cross lan-

guage boundaries because they operate at the discourse-management level rather than the propositional level and require no morphosyntactic integration into the host language (Matras, 2000), which makes them easily integrate into any type of syntactic category IU.

One of such frequently noted DMs is *tipa* ('like/kind of') with the most functional mobility. It both appears as a single-word IU, and as a part of other syntactic structures, FC, PU, and occasionally NFC. This item performs different pragmatic work depending on its position and context. For example, when it appears as a single-word IU, *tipa* functions as a filler (therefore labeled as DM syntax) between two IUs of the speaker's situation being described. The following Russian turn of one speaker illustrates this:

(33) Kapets interesno
damn very.interesting
'Damn so interesting.'

(34) Tipa
like.DM
'Like.'

(35) Nu ya ne znayu skol'ko ya nabirayu konechno
well 1SG NEG know.PRES.1SG how.much 1SG score.PRES.1SG of.course
'Well, I don't know how much I'm scoring of course.'

As shown, the speaker uses *tipa* as an excitement filler between IU1 and IU2. Examining the same MULTICORSKL corpus, Troiani (2023) writes about the quotative function of *tipa*, where it frames reported content as non-verbatim and therefore distances the reporter from responsibility for its accuracy. Such function of the element aligns with its origins in Russian where it developed from an approximative meaning before extending to reported speech ((Kolyaseva, 2018), as cited in Troiani, 2023). The retrieved dataset contains the case of the approximative function as well:

(36) Tipa sonday siyaqtı ğoy
like like.that similar PART
'(I was) like, it's something like that, you know.'

where *tipa* co-occurs with the native Kazakh approximative particle *sïyaqtï* (‘it seems/like’). In this example *tipa* has not displaced its Kazakh equivalent *sïyaqtï* but still coexists with it in the bilingual register. This observation is consistent with Torres Cacoullos and Travis (2019) discussion that contact does not necessarily lead to displacement of native forms, but they can coexist with parallel structures.

Unlike other Russian-origin DMs, *koroche* (‘in short’) shows an interesting pattern, position-wise in IU. It appears at IU onset as a summary or topic-closing device

- (37) Koroche qorqyp otyrmin zhestko
 shortly.DM_{ru} fear-CONV_{kz} sit-PRES.1SG_{kz} hard.ADV_{ru}
 ‘So (in short), I am seriously scared.’

Or in the middle as a filler:

- (38) Bizde koroche byl vybor.
 we-LOC_{kz} shortly.DM_{ru} be.PST.M_{ru} choice_{ru}
 ‘We, (in short), had a choice.’

The speaker inserts a *koroche* and then carries on. The inserted word is not part of the grammatical structure of the sentence but fills the gap before structuring the next IU. And such medial position suggest *koroche* has been absorbed into the bilingual register to the point where it no longer parses as a discrete insertion from Russian but as part of the phonological flow of the utterance. (Andersen, 2014) describes this as a later stage of pragmatic borrowing, where an item that entered the recipient language as a code-switch becomes conventionalized to the point of losing its identity as a transfer from a specific source language.

5.3.4 Prosody vs syntax: complementary roles

Following what has been discussed above, prosody and syntax answer different questions about the same phenomenon of code-switching. In this study, prosody measured CS through SBIU, BS, and ES at IU boundaries, consistently across speakers and recordings. The more precise result of which the SBIU percentage annotated per speaker showed that

speakers were switching in 76.3% of the switches at the boundary of IU overall and always more than 70% for any speaker (section 5.2.1). And in this study syntax was labeled across IUs to determine what kind of syntax allows a switch, and to understand how it can influence the code-switching and can be reflected when prosodical organization is applied. The discussion chapter of the current study has examined the three syntactic categories in sections 5.3.1 through 5.3.3.

As findings showed, the boundary preference can be seen not only through inter-IU level but also the frequent alignment of BS with finite clauses. The almost same proportion of FC and PU at BS sites (1.3:1) against their big difference of proportions at ES sites (3.3:1) is the clearest demonstration that prosody and syntax are operating at different levels: both FC and PU are frequent switching sites as identified through prosodic annotation, but the syntactic category is associated with what kind of switch occurs there. As FCs align with higher numbers of ES IUs, it means its greater ability to host internal Russian embedding as the structural completion and independency allows it.

The absence of internal embedding in NFC IUs is consistent with the complementary relationship between prosodic and syntactic structure observed elsewhere in the data. The low number of embedded switch IUs can be explained by Kazakh's converb dependency that can be disrupted when insertion happens. And this is consistent with the MLF's principle that if the matrix language's morphosyntactic frame cannot be maintained, embedded language material cannot appear.

DMs never align with ES, not because their syntactic environment blocks insertion, but simply DM IUs usually had single words only. Therefore, bilingual DM IUs were very rare, and all times happened as BS IUs. However them being always at the edge of longer IUs or as a single word IU, is not associated with the cline explained in section 5.4.1. Most of the discourse markers of Russian origin in the data are located at the edges of the IU, either as separate IUS or attached to the periphery of a longer IU, but this arrangement does not indicate an active switching. As Appel and Muysken (1987) point out, it is precisely the peripheral role of DMs in grammar that makes them susceptible to borrowings. And their edge position at IU is a syntactic default, not a

switching point. They can be separated from the main sentence regardless of language contact; then this separability is reflected in their tendency to appear at IU boundaries. This is also confirmed by the fact that bilingual cases of DM were rare in the data, and when they did occur, they were borrowings from a single word rather than complete substitutions. Many Russian-origin DMs used by speakers of this study’s dataset has been referenced in literature as borrowings as borrowed element. And from Andersen’s point of view (Andersen, 2014), item is a borrowing when it is a part of the mental lexicon of native speakers of the recipient language and can appear in contexts where switching does not occur at all. Thus, the association between DM position and IU boundaries can be explained by syntax again. Whenever examining an IU that has a boundary Russian element and it is DM, it is more likely that the speaker is not in fact code-switching, but organizing their speech by using DM.

These complementary roles of syntax and prosody have a direct implication for how the present findings relate to earlier syntactic constraint models. The Free Morpheme Constraint, the Matrix Language Frame, etc., all treat switching as a syntactic phenomenon and locate its constraints in the grammar of the languages in contact. The present data do not contradict those constraints; the FC vs NFC distinction is a syntactic observation, and the MLF accounts for why NFC resists internal switching. However, what the data add: syntax is associated with what kind of switching occurs at a given site, while the distribution of switching co-occurs with prosodic structure at the IU level. The prosody and syntax are not competing explanations but more complementary ones, each answering a different question about the same concept of code-switching.

5.4 Russian element integration

5.4.1 Positional distribution or borrowing cline

One of the main challenges at the designing of methodology for this project was how to annotate the language content of IUs. As it was discussed before several times, the contact between Kazakh and Russian is quite dense. Therefore, many Russian words

have penetrated Kazakh speech so deeply that they no longer feel foreign. They sound Kazakh, but there is no standardized way to demonstrate this, due to the lack of studies that distinguish the level of borrowedness. During the first annotation, such a distinction existed only at the level of intuition. But as discussed in the methodology section, this was changed to a form-based annotation approach to remove the bias effect from the annotator. This choice has a direct analytical outcome here as well; rather than labeling certain Russian words as no longer feeling foreign to Kazakh speakers, the study could treat foreignness as an empirical question. A Russian word appearing at an IU edge co-occurs with a boundary; that positional pattern may reflect a lower degree of integration. A Russian word embedded inside a Kazakh IU, surrounded by Kazakh material on both sides, does not carry such a marking. The speakers process it within the same intonation unit as the surrounding Kazakh words. Perhaps, if those words that do not feel foreign were annotated as Kazakh, the percentage of boundary switching would be higher than currently observed.

The decision to label IUs with internal switches as a separate embedded switch (ES) category was motivated by the high degree of contact between Kazakh and Russian, which makes it difficult to establish the boundary between borrowing and active code-switching through form-based annotation alone. ES cases are, therefore, the more analytically telling group in this section. By the heuristic developed for this study, they suggest that for at least some Russian-origin items, speakers produce no switch at the boundary between the item and its Kazakh context, and the absence of that switch is itself evidence of integration status. If code-switching tends to produce IU boundaries, then an item appearing mid-IU without triggering a boundary is not being processed as an active switch. It is part of the same unit of thought as the surrounding Kazakh material.

To distinguish integrated loans from active insertions within the ES data, the heuristic asks two questions. First, does the item carry Kazakh morphology? Second, does the item have a direct alternative/translation in Kazakh? Items satisfying both criteria behave like loans; items satisfying neither behave like insertions; items satisfying only one sit in

between. Applying this to the ES data reveals a cline rather than a binary distinction, which is precisely what the positional analysis wants to identify.

One such word, *monetka*, illustrates a possible borrowedness case. In REC02, the word appears six times, four of which are in ES IUs, where it adopts Kazakh morphological suffixes, including:

1. plural: *monetka-lar*
2. accusative: *monetka-ny*
3. first-person possessive: *monetka-m*

integrating fully into the Kazakh nominal paradigm. Before answering the second heuristic question, whether a Kazakh alternative exists, it is worth providing a context for its use.

In REC02, students are sharing everyday stories, and the speakers who use *monetka* are students of Nazarbayev University (NU) living in its dormitories. The researcher is a former NU dormitory resident with direct familiarity with the infrastructure described in the recording, a position established as a methodological resource in section 3.2.1. In NU dormitories, residents need coins to use laundry machines and water dispensers. The names of both facilities are used among residents in Russian (*stiralka*, *vodomat*) or in English (*laundry*), but never in Kazakh, mostly because student communication in dormitory Telegram chats is in Russian even among Kazakh-dominant speakers. But spoken communication can be mixed or Kazakh-dominant (personal observation of the researcher). There is an explanation of why *monetka* and both those facilities are referred to in their Russian names. When an object or practice is introduced through a particular language, people adopt that language's word for it and use it even when switching back to their other language (Haspelmath, 2009; Matras, 2009). Since the machines have no Kazakh name in this community, neither does the coin used to operate them *tiyn*, the general Kazakh word for coin, does not substitute for *monetka* because it does not carry the institutional meaning the word holds in this context and this answers the second heuristic question: no Kazakh alternative exists for this referent in this community, which

places *monetka* on the borrowing end of the cline rather than the active switching end. Its consistent appearance in ES position with full Kazakh morphological integration reflects the fact that speakers do not experience it as a Russian insertion requiring a prosodic reset – it is simply the name of the thing. *Komnata* presents a more contextually split pattern than *monetka*. Across 6 occurrences in the dataset, it always carries Kazakh morphological suffixes, satisfying the first heuristic question positively in every case. However, the second criterion, whether a Kazakh alternative exists, depends on context. In REC01, *komnata* appears once in ES position in a flat-describing context, as in a one-room flat.

- (39) Bır komnata-lıq-ta tūrıp pa
 one_{kz} room_{ru}-NMLZ-LOC_{kz} live-CONV_{kz} Q_{kz}
 ‘Did they live in a one-room (flat)?’

While the Kazakh alternative *bölme* exists and is used quite frequently as well, *komnata* remains common in informal speech, particularly in contexts from Soviet-era housing terminology. This single example from REC01, fully embedded with Kazakh morphology, places it near the borrowing end of the cline in that context.

In REC02, the picture is more mixed. Out of six occurrences, five of them are ES and one is BS, as in the example (40):

- (40) komnata-ğa bar-ğasın
 room_{ru}-DAT_{kz} go-CONV.AFTER_{kz}
 ‘After going to the room.’

The majority of occurrences are embedded, as in the example (41):

- (41) prosto men komnata-da bol-maym
 just_{ru} 1SG_{kz} room_{ru}-LOC_{kz} be-NEG.PRES.1SG_{kz}
 ‘I am just not in the room.’

It stands to mention that, in both ES and BS occurrences, *komnata* carries Kazakh suffixes. The dormitory context is relevant here in the same way it was for *monetka*: NU

dormitory rooms are discussed among students mostly in Russian, and *komnata* is the habitual term for them.

Istoriya presents a different integration case from the words discussed above. Across REC02, it appears seven times, three as BS and four as ES, but carries Kazakh morphological suffixes in only two of those cases. This makes it the least morphologically integrated item in the ES data and places it closer to the active switching end of the cline. The semantic context is consistent across the REC02: speakers use *istoriya* for sharing personal life stories among each other as in example (42).

- (42) sondej istoriya bol-dy
 such_{kz} story_{ru} be-PST.3SG_{kz}
 ‘There was such a story.’

For this word, as well, the Kazakh alternatives *oqiğa* and *hikaya* exist but are not in present in this register among this speaker group. There is also one example of *istoriya* use in REC04 which is in BS IU and with a Kazakh suffix. And in comparison to semantic use of *istoriya* in REC02, it has completely different function as the speaker by saying:

- (43) istoriya-ğa sal-ğim kel-ip tūr
 story_{ru}-DAT_{kz} put-DES.1SG_{kz} come-CONV_{kz} stand.PRES_{kz}
 ‘I want to post it as a story.’

is referring to Instagram function of a temporary post, story. In modern Kazakh speech there is no alternative in Kazakh, therefore it is quite frequent when people use an English original ‘story’ in its plural form ‘stories’ or even more frequently as speaker of REC04 uses in example (43) in Russian.

Russian words embedded in IUs of REC01 demonstrate a set of cultural borrowing consistent with the older speaker profile of that recording. First, *Kompot* entered Kazakh through decades of Soviet cultural cuisine contact. According to Mack and Surina, the use of the word *kompot* mechanism was institutional: Soviet-standardized catering served the same menus across the entire Union, and Kazakh schoolchildren consumed *kompot* alongside Russian ones, together with its Russian name (Mack and Surina, 2005, pp. 150, 170).

The word *kompot* appears ten times, and five times of it is in ES position, fully embedded within Kazakh IUs with Kazakh morphological marking, never triggering a prosodic boundary. Doing so, speakers retrieve the word as part of the Kazakh lexical frame because no Kazakh alternative was made for the concept as it arrived through that specific cultural channel.

Word	ES	BS	Suffixation	Total
<i>monetka</i>	4	2	4	6
<i>komnata</i>	5	1	6	6
<i>istoriya</i>	2	4	2	7
<i>kompot</i>	5	5	5	10

Table 5.1: Positional distribution and suffixation of selected Russian elements.

Table 5.1 summarizes the positional distribution and suffixation counts for four Russian elements discussed above. *Monetka* and *komnata* both have a higher ES than BS ratio, and both carry Kazakh suffixes consistently that show a borrowing profile according to both the heuristics and cline.

Istoriya sits at the opposite end, more BS than ES, and low suffixation, with both criteria pointing toward active switching rather than borrowing. *Kompot* is the exception: suffixation is consistent, but ES and BS are equal, suggesting positional stabilization has not kept pace with morphological integration. Reviewing the dataset as a whole, the positional criterion, whether a Russian element appears as ES or BS, provides additional information about integration status that morphological suffixation alone can not explain. The data from Table 5.1 illustrates this.

Monetka and *komnata* show high ES ratios alongside consistent suffixation, placing them near the borrowing end of the cline. *Istoriya*, with predominantly BS position and low suffixation, sits closest to the active switching end, and crucially, both criteria agree on this placement. In comparison to others, *Kompot* is the most analytically interesting case: its suffixation rate is high, but its BS occurrences match its ES occurrences exactly. This may suggest that morphological integration has proceeded further than positional distribution, and this dissociation between the two criteria points to a broader problem

with using suffixation counts as the primary measure of integration. In many occurrences, the Russian element occupied a syntactic position that does not require case marking in Kazakh, meaning the absence of a suffix reflects grammatical context rather than foreignness. The example:

- (44) *kompot* *tura-p* *qoyu* *kerek*
 compote_{ru} grate-CONV_{kz} put.down_{kz} necessary_{kz}
 ‘One needs to grate the compote’

illustrates this directly: *kompot* appears without any Kazakh suffix, but the phrase is grammatically natural and structurally identical to a fully Kazakh equivalent such as,

- (45) *bashbarmaq* *pisir-ip* *qoyu* *kerek*
 bashbarmaq cook-CONV put.down necessary
 ‘One needs to cook the bashbarmaq.’

where the noun appears bare in the same position. Replacing *kompot* with a Kazakh noun requires no structural adjustment whatsoever. A word that consistently appears in positions that do not have case marking will show low suffix frequency regardless of how deeply it has been absorbed into the lexicon. As a native Kazakh-Russian bilingual, the researcher’s intuition consistently aligned with the positional pattern rather than suffix frequency alone, which supports treating ES/BS distribution as the more sensitive indicator of integration.

Another culturally integrated word is *yolka*. REC03 opens with a story set during New Year celebrations. The Soviet New Year holiday was constructed from 1935 as a secular substitute for Christmas, packaging its symbols, including the fir tree, as tools of Soviet identity-building across all republics of the Union (Smolkin, 2018). The word arrived in Kazakh as part of a standardized Soviet festive word imposed uniformly across the USSR (Mack and Surina, 2005). A Kazakh alternative exists, *shyrsha*, but it has rarely displaced the Russian form in informal speech, and in this dataset shows no sign of doing so. *Yolka* appears in ES position as *yolkamyz*, carrying the Kazakh first-person plural suffix *-myz* inside a complete Kazakh existential clause:

- (46) Kishkentai yolka-myz bar.
small_{kz} fir.tree_{ru}-1PL.POSS_{kz} exist_{kz}
 ‘We have a small fir tree.’

The word is grammatically Kazakh in every respect except its phonological origin, which places it at the borrowing end of the cline described in section 5.4.1.

What unites all the listed words is not their Russian origin but the conditions under which they entered Kazakh speech. Each arrived as part of a semantic package, institutional, culinary, or festive, for which no Kazakh alternative was installed at the point of introduction. Their appearance in the ES position with Kazakh morphological integration reflects the fact that speakers retrieve these words from the same lexical store as their Kazakh material. They are not insertions from another language requiring a prosodic boundary, as they are simply the words for those things in this community.

5.4.2 Morphological integration patterns

The positional distribution of Russian elements within IUs raises a further question: are all embedded Russian items structurally equivalent, or do they represent different degrees of integration into Kazakh? The morphological evidence is to answer the latter.

Across all four recordings, ES instances fall into at least three structurally distinct patterns, each reflecting a different point on the borrowing cline introduced in section 5.4.1. Moving from the most to the least morphologically integrated, these patterns show Russian material being absorbed into Kazakh at progressively deeper levels of grammar.

The most common pattern is a Russian noun taking Kazakh case morphology. Forms like:

- (47) holodilnik-ke
 refrigerator.N_{ru}-DAT_{kz}
 ‘to the refrigerator.’

- (48) komnata-n
 room.N_{ru}-2SG.POSS_{kz}
 ‘your room’

- (49) mantovarka-ndy
 steamer.N_{ru}-ACC_{kz}
 ‘the steamer (acc.)’

treat the Russian noun as a bare stem and attach Kazakh suffixes exactly as they would to a native Kazakh noun. A subset goes further, stacking multiple Kazakh suffixes onto the same Russian root:

- (50) podruzhka-lar-ym
 girlfriend.N_{en}-PL-1SG.POSS_{kz}
 ‘my girlfriends’

- (51) raznitsa-sy-n
 difference.N_{ru}-3SG.POSS-ACC_{kz}
 ‘its difference (acc.)’

Stacked suffixes indicate the Russian noun has entered not just the case system but the full nominal paradigm: number, possessive agreement, and case simultaneously. These are not switches in any meaningful sense; the Russian root is behaving as a Kazakh stem.

A second pattern involves Russian adverbs serving as modifiers inside Kazakh clauses without taking any Kazakh morphology:

- (52) Zhestko zhanbyr zhaydy
 hard.ADV_{ru} rain_{kz} fall.PST.3SG_{kz}
 ‘It rained hard.’

Unlike boundary-positioned pragmatic operators such as *tipa* and *koroche*, these items fill a syntactic slot inside the clause rather than framing it from the edge. They integrate structurally but not morphologically, and this is possible because adverbs in Kazakh are invariable. They carry no inflectional morphology, meaning a Russian adverb can occupy an adverbial slot in a Kazakh clause without any formal adaptation being required by the grammar.

A third pattern involves two constructions in REC03. The first is Russian item with a Kazakh quotative frame built on *deidi* (‘says’) or *dedi* (‘said’):

- (53) Zarazhish'sya deidi
 get.infected.2SG.FUT_{ru} say.3SG.PRES_{kz}
 'They say you will get infected.'

In each case *deidi* carries all grammatical marking while Russian supplies the reported content. Quotative frames are a natural site for code-switching because they report an original utterance rather than the speaker's own formulation. The second construction moves in the opposite direction. Several turns in REC03 show Russian verb stems combined with the Kazakh auxiliary *etu* ('to do') to form a fully conjugated Kazakh predicate:

- (54) Gnobit' etip tastaidy
 bully.INF_{ru} do.CONV_{kz} do.completely.3SG.PRES_{kz}
 'They go ahead and bully.'

The Russian infinitive provides the lexical content; *etu* supplies all inflectional morphology, tense, person, aspect, mood. The structural basis of this pattern has been identified in Akishev's (2020) where he writes that Russian infinitives are absorbed into Kazakh auxiliary constructions because Kazakh requires the auxiliary to carry all inflectional morphology, leaving the lexical verb as a bare stem. The *etu* construction observed in this dataset shows the same.

The semantic domains also add a context to understand: *gnobit'* is a colloquial Russian, *shipperit'* and *otmetit'* ('to tag') are digital media vocabulary for which no established Kazakh verb exists in the active repertoire of these speakers. Together, the three patterns confirm that Russian elements inside Kazakh IUs are not uniform instances of code-switching.

5.5 Limitations

The main limitation of this study is its methodology. The study of Shenk's (2006) whose framework served as a basis for the current study had a central argument, that IU boundaries are stronger predictors of code-switching than syntactic boundaries, but this current

study cannot test that claim directly. The syntactic annotation here describes the category of each IU; it does not identify syntactic boundaries that fall at different points than IU boundaries. For a typologically distant pair like Kazakh-Russian, where IU and syntactic boundaries are less likely to coincide than in Spanish-English, this comparison would be especially informative. Both time constraints and lack of replicating Shenk’s syntactic completion point analysis for Kazakh-Russian remain the limitations of this study, which later made it into different study.

The second limitation is dataset size and speaker diversity. Three of the four recordings come from university-linked speakers. REC01 is the only exception, and it contains the fewest IUs both overall and monolingual Russian. The corrective per-speaker SBIU annotation, which is the more accurate of the two annotation levels, covers only 622 IUs from one recording. Results may not generalize to rural speakers, less Russified communities, or any other cases of CS of spoken data.

Third, the form-based annotation scheme labels any IU containing Russian-origin material as bilingual or Russian, which means borrowings and active code-switches are labeled the same. This has been chosen to remove annotator bias and, at the same time, inflates the number of ES IUs that possibly contain borrowed Russian words, which, in the end, created a circularity. This circularity cannot be fully resolved within the current design. However, it does not make the positional analysis meaningless, but just reframes what the analysis can claim. What it shows is that certain Russian-origin items behave differently from others within the same annotation scheme: they appear mid-IU with consistent Kazakh morphological integration, in positions any Kazakh word of the same grammatical category would occupy, without triggering a prosodic boundary. That behavioral pattern is worth documenting, even if the label that put those items in the dataset was applied conservatively. The positional framework is therefore best understood not as a borrowing detector but as a gradient measure of how differently Russian-origin items behave within Kazakh prosodic structure.

Last but not least, the typological distance question raised in the literature review is not answered here. Answering it properly requires a comparative design, so that the same framework applied to a typologically similar pair under equivalent conditions. This

study establishes that Shenk's framework is applicable to Kazakh-Russian data and that boundary preference is present. The question of whether boundary preference differs systematically between typologically distant and similar pairs requires a different study with total comparative perspective with another pair of typologically similar languages.

Chapter 6

Conclusion

This thesis applied Shenk's (2006) prosodic-discourse framework to Kazakh-Russian bilingual speech and extended it with two switch-type categories: embedded switch (ES) and boundary switch (BS) to capture code-switching patterns within IUs. The central finding is that IU boundaries are a consistent site of code-switching in this dataset, at both the inter-IU and intra-IU levels, and that syntactic structure and prosodic organization are complementary rather than competing explanations.

The results of the study showed that, at the inter-IU level, 76.3% of switching events occurred at an IU boundary across individual speakers. The same preference holds within bilingual IUs: 76.2% of Russian elements appeared at IU edges more than interiors. Along with that, two independently annotated metrics, SBIU and the BS/ES ratio, have almost the same rate, which is the strongest evidence the study offers for boundary preference as a structural property of this bilingual speech.

Along with the answer to the central question, the gap between this study's results and Shenk's 96% is also a significant result, which showed the density of Kazakh-Russian contact, where Russian items appear in Kazakh speech as loans as well as active switches, and a borrowed item retrieved mid-IU produces no prosodic boundary. The 76.3% is therefore lower, not because IU boundaries less effective in tracing contact situations, but because fewer switching events occur - many Russian items are simply the Kazakh words for particular things. The ES data positional distribution makes this case visible:

monetka, *komnata*, *yolka*, and *kompot* appear mid-IU with full Kazakh morphological integration because speakers retrieve them from the same lexical store as their Kazakh material; they are not insertions from another language. The ES/BS positional distinction turns out to be a more sensitive diagnostic of integration status than suffixation counts alone. Because, as showed discussion a word can carry Kazakh morphology and still pull its own prosodic boundary when speakers accept it as foreign.

The syntactic findings add complementary explanations on how Kazakh-Russian code-switching is structured. For example, finite clauses align with internal Russian embedding at a ratio of 3.3:1 over phrasal units, while non-finite clauses resist it almost entirely. This finding was explained by the converb dependency structure of Kazakh, which leaves no position inside the IU where a Russian element can sit without disrupting active morphosyntactic dependencies. And discourse markers behaved completely differently. Because Russian-origin items like *tipa* and *koroche* appear at IU edges by syntactic default, not because speakers are switching and their edge position is a property of discourse markers in general.

The most original contribution of this study is the ES/BS framework as a tool for approaching the borrowing question empirically mentioned above. Instead of labeling that certain Russian words feel native to Kazakh speakers, the positional distribution of Russian elements within IUs makes integration status visible in the structure of speech itself. Despite the limitations of the methodology of this question, future work could develop this into a quantitative criterion by using independent judgments of integration tested against ES/BS distribution.

Chapter 7

Research ethics

This study involves secondary analysis of an existing corpus. The methodology required only manual annotation of recordings from the MULTICORSKL corpus. All ethical procedures governing the original data collection, including IREC approval, were carried out by the corpus builders, including the secondary advisor, Andrey Filchenko, and the external advisor, Giorgia Troiani. No additional ethical approval was required for the present study. To protect anonymity, recording identifiers were replaced with the labels REC01-REC04. Speakers of REC02 were relabeled as Speaker A, B, and C as well.

References

- Akishev, T. (2020). The syntax of kazakh-russian intrasentential code-switching in reported clauses. *Language. Text. Society*, 7(1).
- Andersen, G. (2014). Pragmatic borrowing. *Journal of Pragmatics*, 67:17–33.
- Appel, R. and Muysken, P. (1987). *Language Contact and Bilingualism*. Edward Arnold, London.
- Auer, P. (1984). *Bilingual conversation*. John Benjamins, Amsterdam.
- Auer, P. and Muhamedova, R. (2005). “embedded language” and “matrix language” in insertional language mixing: Some problematic cases. *Rivista di Linguistica*, 17(1):35–54.
- Bahri, S. (2025). *Challenging the universality of Poplack’s syntactic constraints: Insights from Arabic-English code-switching among Saudi Arabian women*. PhD thesis, University of New Mexico.
- Belazi, H. M., Rubin, E. J., and Toribio, A. J. (1994). Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237.
- Berk-Seligson, S. (1986). Linguistic constraints on intrasentential code-switching: A study of Spanish/Hebrew bilingualism. *Language in Society*, 15(3):313–348.
- Carlson, R., Hirschberg, J., and Swerts, M. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46(3–4):326–333.

- Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.
- Chernyavskaya, V. and Zharkynbekova, S. (2024). Code switching patterns in Kazakh-Russian hybrid language practice: An empirical study. *Training, Language and Culture*, 8:9–19.
- Choi, J.-O. (1991). Korean-English code switching: Switch-alpha and linguistic constraints. *Linguistics*, 29(5):877–902.
- Djennane, T. and Louati, F. (2024). Challenging the free morpheme and the equivalence constraints: Counter-evidence from Algerian Arabic/French code-switching. *AL-Lisaniyyat*, 30(1):147–158.
- Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., and Paolino, D. (1993). Outline of discourse transcription. In Edwards, J. A. and Lampert, M. D., editors, *Talking data: Transcription and coding in discourse research*, pages 45–89. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Ene, E. (2007). Romanian-English code-switching: A preliminary study. *Linguistic and Philosophical Investigations*, 6:225–233.
- Filchenko, A., Troiani, G., Du Bois, J. W., Sarseke, G., Akanov, A., Bizhanova, M., Mikhailov, N., Temirbekova, T., Seitak, B., and Turaliyeva, Z. (2023). Multimedia corpus of spoken Kazakh language.
- Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.
- Grosjean, F. (2024). The statistics of bilingualism. In *On bilinguals and bilingualism*, pages 138–147. Cambridge University Press.
- Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. In Haspelmath, M. and Tadmor, U., editors, *Loanwords in the world’s languages: A comparative handbook*, pages 35–54. De Gruyter Mouton.

- Inbar, M., Genzer, S., Perry, A., Grossman, E., and Landau, A. N. (2023). Intonation units in spontaneous speech evoke a neural response. *Journal of Neuroscience*, 43(48):8189–8200.
- Kolyaseva, A. (2018). The ‘new’ Russian quotative *tipa*: Pragmatic scope and functions. *Journal of Pragmatics*, 128.
- Li, R., Zhang, Z., Ni, C., Xiao, W., Wei, J., and Dai, H. (2018). Examining the functional category in Chinese–English code-switching: Evidence from eye-movements. *Journal of Psycholinguistic Research*, 47(1):1–28.
- Mack, G. R. and Surina, A. (2005). *Food culture in Russia and Central Asia*. Greenwood Press.
- Matras, Y. (2000). Fusion and the cognitive basis for bilingual discourse markers. *International Journal of Bilingualism*, 4(4):505–528.
- Matras, Y. (2009). *Language contact*. Cambridge University Press.
- Muhamedowa, R. (2005). Kasachisch-russisches code-mixing: Ein fall von morphologischer vereinfachung. *Zeitschrift für Sprachwissenschaft*, 24(2):263–285.
- Muhamedowa, R. (2009). The use of Russian conjunctions in the speech of bilingual Kazakhs. *International Journal of Bilingualism*, 13(3):331–356.
- Muhamedowa, R. (2016). *Kazakh: A comprehensive grammar*. Routledge.
- Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
- Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press.
- Nakamura, M. S. (2025). Beyond bilingualism: A discourse analysis of Uzbek-Russian code-switching in contemporary Uzbekistan. *Turkophone*, 2(1).
- Nassyrova, A. (2018). Functioning Kazakh lexical units in the Russian speech of the residents of Kazakhstan. *Przegląd Wschodnioeuropejski*, 9(2):201–207.

- Nurtazina, M., Alefirenko, N., Kesner, J., and Shakhputova, Z. (2019). Linguistic and extralinguistic factors of code-switching as a process of contacting languages. In *Proceedings of the 2nd International Conference on Education Science and Social Development (ESSD 2019)*. Atlantis Press.
- Pattichis, R., LaCasse, D., Trawick, S., and Torres Cacoullos, R. (2023). Code-switching metrics using intonation units. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16840–16849.
- Poplack, S. (1980). 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, S. (2010). Syntactic structure and social function of code-switching. In Durán, R. P., editor, *Latino language and communicative behavior*, pages 169–184. Ablex. Original work published 1981.
- Rayo, W., Barrita, A. M., Martinez, L. C., and Carbajal, I. (2024). The complexity in bilingual code-switching research: A systematic review. *Languages*, 9(6).
- Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
- Sankoff, D. and Poplack, S. (1981). A formal grammar for code-switching. *Papers in Linguistics*, 14(1):3–45.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press.
- Shenk, P. S. (2006). The interactional and syntactic importance of prosody in Spanish-English bilingual discourse. *International Journal of Bilingualism*, 10(2):179–205.
- Simpson, H. E. (2016). *The role of intonation units in memory for spoken English*. PhD thesis, University of Edinburgh.
- Smagulova, J. (2008). Language policies of Kazakhization and their influence on language attitudes and use. *International Journal of Bilingual Education and Bilingualism*, 11(3–4):440–475.

- Smolkin, V. (2018). *A sacred space is never empty: A history of Soviet atheism*. Princeton University Press.
- Terlikbayeva, N. and Menlibekova, G. (2021). The dynamics of language shift in Kazakhstan: Review article. *Journal of English Language Teaching and Applied Linguistics*, 3(2):12–22.
- Torres Cacoulios, R. and Travis, C. E. (2018). *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge University Press.
- Torres Cacoulios, R. and Travis, C. E. (2019). Code-switching and bilinguals’ grammars. In Adamou, E. and Matras, Y., editors, *The Routledge handbook of language contact*. Routledge.
- Troiani, G. (2023). *Representing a language in use: Corpus construction, prosody, and grammar in Kazakh*. PhD thesis, University of California, Santa Barbara.
- Zharkynbekova, S. K. and Chernyavskaya, V. E. (2022). Kazakh-Russian code mixing in metacommunicative perspective. *Vestnik of Saint Petersburg University: Language and Literature*, 19(3):512–528.
- Ótott Kovács, E. (2015). *The syntax of non-finite clauses in Kazakh*. PhD thesis, University of Szeged.