# Borrowing and contact intensity

Evangelia Adamou, Walter Breu, Lenka Scholze, Rachel X. Shen

Adamou et al. 2016, Borrowing and contact intensity, *Journal of Language Contact*.

# Borrowing and contact intensity: A corpus-driven approach from four Slavic minority languages

Evangelia Adamou, CNRS (French National Centre for Scientific Research)
adamou@vjf.cnrs.fr
7, rue Guy Moquet
94801 Villejuif, France

Walter Breu, University of Konstanz
Walter.Breu@uni-konstanz.de
Fachbereich Sprachwissenschaft
Fach D 179
D-78457 Konstanz, Germany

Lenka Scholze, University of Konstanz
lenka.scholze@uni-konstanz.de
Fachbereich Sprachwissenschaft
Fach D 179
D-78457 Konstanz, Germany

Rachel Xingjia Shen, University Paris Diderot and Labex EFL
rachel.shen@univ-paris-diderot.fr
Case 7031 – 5, rue Thomas Mann,
75205 Paris cedex 13, France

Adamou et al. 2016, Borrowing and contact intensity, *Journal of Language Contact*.

# Borrowing and contact intensity: A corpus-driven approach from four Slavic minority languages

## Abstract

Numerous studies on language contact document the use of content words and especially nouns in most contact settings, but the correlations are often based on qualitative or questionnaire-based research. The present study of borrowing is based on the analysis of free-speech corpora from four Slavic minority languages spoken in Austria, Germany, Greece, and Italy. The analysis of the data, totalling 34,000 word tokens, shows that speakers from Italy produced significantly more borrowings and noun borrowings than speakers from the other three countries. A Random Forests analysis identifies 'language' as the main predictor for the ratio of both borrowings and noun borrowings, indicating the existence of borrowing patterns that individual speakers conform to. Finally, we suggest that the patterns of borrowing that prevail in the communities under study relate to the intensity of contact in the past, and to the presence or absence of literary traditions for the minority languages.

## Keywords

Borrowing, Slavic, corpus-driven, variationist, minority languages.

## 1. Introduction

Numerous studies on language contact document the use of content words and especially nouns in most contact settings (e.g., Thomason and Kaufman, 1988; Muysken, 2000; Myers-Scotton, 2002; Matras and Sakel, 2007; Matras, 2009; Haspelmath and Tadmor, 2009). A vast literature on language contact agrees that the degree of borrowing largely depends on extra-linguistic factors such as the intensity and type of language contact, as well as on language attitudes (e.g., Thomason and Kaufman, 1988; Muysken, 2000; Winford, 2003; Matras, 2009; Haspelmath and Tadmor, 2009). But the correlations are often based on qualitative or questionnaire-based research, and not on quantitative analysis of naturally occurring speech.

In this paper we provide evidence relevant to this discussion based on the analysis of free-speech corpora from four Slavic minority languages in contact with three Indo-European languages from three different branches, namely German (Germanic), Greek (Greek), and Italian (Romance). The corpora were collected among fluent speakers of the Slavic languages in a language documentation perspective. The analysis of the corpora, totalling approximately 34,000 word-tokens, is conducted in a variationist perspective with respect to the study of language contact (see Poplack, 1993).

The main finding of the study is that individual speakers of Slavic minority languages conform to patterns of borrowing which prevail in their bilingual community (also see Poplack, 1985; Adamou and Granqvist, 2014; Travis and Torres Cacoullos, in press). Moreover, most communities under study use surprisingly low proportions of word-tokens from the current-contact language despite fluency of the speakers in both languages. These communities can be dubbed 'low borrowers' following the typology in Tadmor (2009). In contrast, the Molise Slavic speakers from Italy are producing significantly higher proportions of contact word-tokens, and thus fall under the category of 'high borrowers' (Tadmor 2009: 57). As we will show, the ratio of current contact-language words seems to be related to the intensity and type of contact as well as to language attitudes related to literary traditions established in the past.

In this paper, we first provide an overview of the theoretical background in 2. We then present some background on the Slavic minority languages under study, in 3, followed by the methodology, in 4, and the results of the corpus analysis, in 5. Lastly, in 6, we discuss the results with respect to sociolinguistic explanatory factors.

## 2. Theoretical background

Borrowing can be defined as the transfer of sound and form-meaning units (Heine and Kuteva, 2005) while codeswitching may be broadly defined as the alternation of languages within a conversation (Matras, 2009: 101). Historical and comparative linguistic methods allow us to identify borrowings which have been introduced in past contact settings. When, however, these items come from a current-contact language, determining whether they should be treated as borrowings or as codeswitches becomes more difficult.

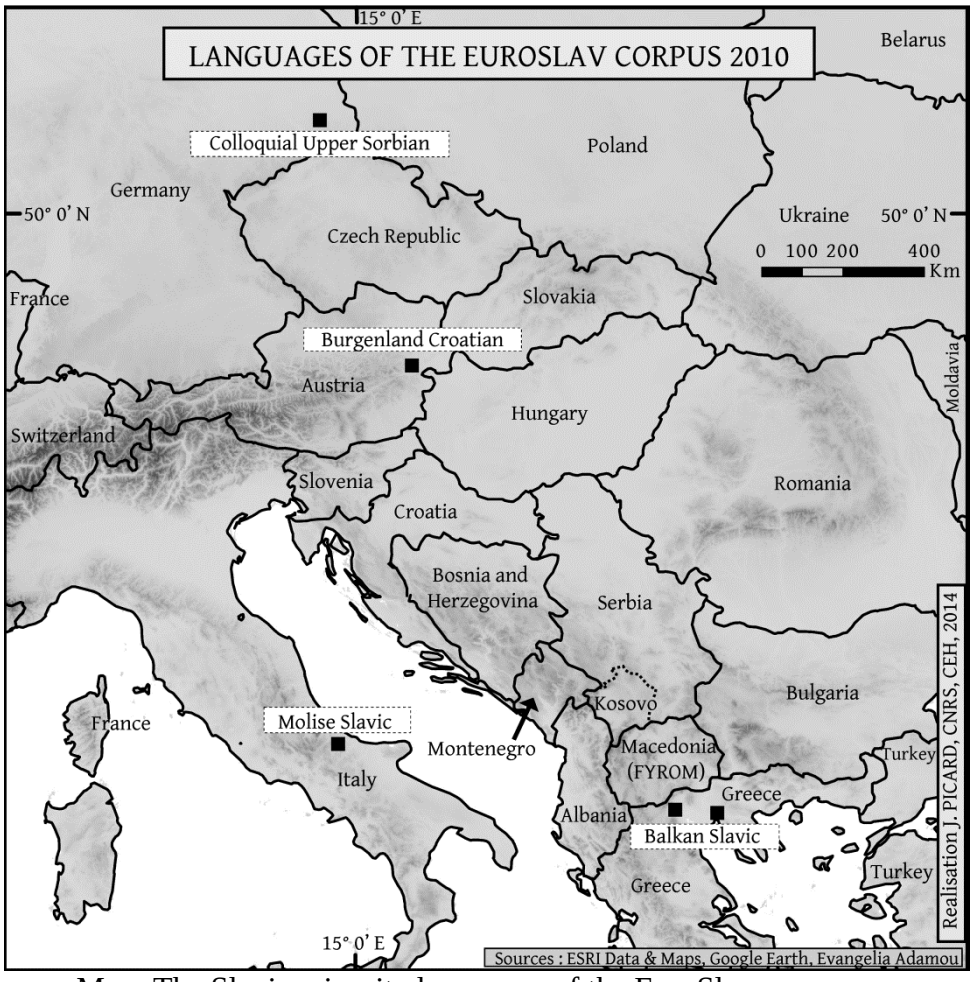
The view that borrowing and codeswitching form a continuum is expressed among others in Matras (2009: 110–111), including an updated list of criteria which allows us to distinguish between the two: the speaker's degree of bilingualism (monolingual vs. bilingual), the item's composition (utterance vs. single lexeme) and functionality (stylistic vs. default use), the unique character of the referent (lexical vs. para-lexical), the item's operationality (core vocabulary vs. grammatical operations), the regularity of the process (single vs. regular occurrence), and structural integration (non-integrated vs. integrated).

The degree of structural integration was long considered as a dominant feature for the identification of borrowings: "Despite etymological identity with the donor language, established loanwords assume the morphological, syntactic, and often, phonological, identity of the recipient language" (Poplack, 2001: 2063). In practice, all levels do not show the same degree of integration and Poplack and Dion (2012) argue that all single-word tokens of a current-contact language should be treated as borrowings. This proposal is based on the analysis of a three-million word corpus from French-English bilinguals of Canada, in which single words are structurally integrated independent of their frequency and diffusion across the speakers, while multi-word tokens always behave as codeswitches by not being structurally integrated. In contrast, other researchers suggest a distinction between core borrowings and cultural borrowings, where cultural borrowings should be treated as codeswitches (see Myers-Scotton, 1993).

As far as the criteria that trigger borrowing are concerned, Matras (2009: 161–162) identifies the 'utilitarian' motivation, i.e. the fact that words associated with a particular social activity in the contact language are more borrowable: e.g., unique referents > general/core vocabulary; nouns > non-nouns; numerals in formal contexts > numerals in informal contexts; higher cardinal numerals > lower cardinal numerals; days of week > times of day (Matras, 2009: 161). Matras further identifies the need to reduce the processing load, whereby the types of relatively less accessible words are more borrowable than those characterized by 'frequency', 'routine', and 'casualness': e.g., more remote kin > close kin; peripheral local relations > core local relations (Matras, 2009: 161). Finally, the degree of the speaker's control in the communicative negotiation is relevant for borrowing, as weaker control may interfere with the speaker's language selection mechanism, thus favouring borrowing: e.g., but > or > and; superlative > comparative > (positive); indefinites > interrogatives > (other) deixis, anaphora; focus and modal particles > other particles; already > still; only > too (Matras, 2009: 161).

## 3. Slavic minority languages in Europe

The present study is based on four Slavic minority languages spoken in four European countries, namely Austria (Burgenland Croatian), Germany (Colloquial Upper Sorbian), Greece (Balkan Slavic), and Italy (Molise Slavic); see Map. In all these cases, language contact takes place between a Slavic minority language in contact with a non-Slavic language which is the official, state language and has a strong literary written tradition, and which is used in administration, schools, and media, but also in the everyday life of the communities.[1] This type of 'total (or absolute) language contact', where all the speakers of a minority language are competent speakers of the majority language, is typical of twentieth-century European states. Such settings are marked by a deep imbalance between minority and majority languages, with the majority language covering almost all the language domains, and a tendency to shift to the majority language is frequently observed. In total language contact, "everybody can freely introduce elements of the majority language, functioning as an umbrella language, into the local minority language" (Breu, 2011: 431). Contact with the official languages is rendered more complex through contact with the regional varieties of the official languages and eventually through contact with the most-closely related standard Slavic languages.




MAP. The Slavic minority languages of the EuroSlav corpus

[1] These settings cannot be qualified as diglottic as an anonymous reviewer suggested to us.

### *3.1. Molise Slavic, Italy*

Molise Slavic or Na-Našu/Na-Našo, literally 'in our (language)', is a South Slavic language (Štokavian-Ikavian) which has been spoken in Italy for 500 years, following earlier migration from Dalmatia. Molise Slavic is spoken in three bordering villages, Acquaviva Collecroce, Montemitro, and San Felice del Molise. The dialectal differences of Molise Slavic between these three villages are considerable on all linguistic levels, especially in phonology and lexicon, and to a lesser extent morphology (Breu, 2011: 434).

Molise Slavic has been under the influence of the Molisian dialect of Italian and several varieties of regional Italian for centuries, joined by the influence of Standard Italian about 150 years ago. Nowadays, the Molise Slavic speakers do not know Molise Italian but they only use Standard Italian and several regional varieties of Italian. The Molise Slavic dialects are no longer transmitted: half of the roughly 2,000 inhabitants of the villages of Acquaviva, Montemitro, and San Felice are speakers of Molise Slavic, and these include only a few children. Each village shows different degrees of language preservation: Acquaviva Collecroce is the village with the highest number of Slavic speakers who nowadays predominantly use Italian; Montemitro is the smallest and most conservative community; and San Felice has only a few remaining fluent speakers of Molise Slavic. In Krauss's scale of endangerment (Krauss, 2006) the Molise Slavic varieties fall under the category B, as they are almost exclusively spoken by the parental generation and up, with differing degrees of vitality depending on the locality. The existence of bilingual education in Italian and Molise Slavic has a very limited impact on language transmission and revitalization.

### *3.2. Balkan Slavic, Greece*

The Balkan Slavic varieties of present-day Northern Greece, closely-related to Bulgarian and Macedonian, have been spoken in the area since the seventh century. During the Ottoman times, from the fifteenth up until the beginning of the twentieth century, contact with Greek, at least for the Christian Orthodox populations, was restricted to church-related activities while contact with Turkish was generally restricted to trade. A shift to Greek, which started when the area became part of the Greek State in 1912–1913, is nowadays almost complete for most of the Orthodox Balkan Slavic-speaking populations of Greece.[2] Contact with Greek became increasingly important through education and administration during the first part of the twentieth century, namely with the literary Greek language *Katharevousa*. More importantly, for most Orthodox Slavic-speaking communities contact took place with the spoken variety of Modern Greek, *Demotiki*, namely through intermarriages and everyday interactions at work.

Balkan Slavic is represented in this study by two sub-corpora. One corpus was recorded in 1976, in a locality close to the city of Edhessa, Hrisa. The other corpus comes from Nashta, literally 'our (language)', a variety spoken in the small town of Liti, which is located 10 km from the city of Thessaloniki recorded in the years 2000s from among the last speakers. Both corpora were recorded with fluent speakers in order to document the Slavic varieties, but the corpus of the 1970s was recorded when the local Slavic variety was still in use in the community together with Modern Greek.

In the scale of endangerment (Krauss, 2006), the Nashta variety is thus critically endangered (D), but the variety of Hrisa at the moment of the recordings can be qualified as severely (C) endangered. The Balkan Slavic varieties of Greece are not officially recognized by the Greek state and there are no revitalization efforts.

[2] Muslim Balkan Slavic-speaking populations living in Greek Thrace, known as 'Pomaks', are nowadays shifting to Turkish, the local minority language.

### *3.3. Colloquial Upper Sorbian, Germany*

Colloquial Upper Sorbian is a Slavic minority language spoken in Germany, and is the everyday language of the Catholic Sorbian population living in Upper Lusatia. Most Sorbian speakers live in the former district of Kamenz (Sorbian: Kamjenc), now part of the *Landkreis* of Bautzen (Budyšin), including the rural communities of Crostwitz (Chróśćicy), Ralbitz/Rosenthal (Ralbicy/Róžant) and their immediate surroundings (Breu, 2000; Scholze, 2008). Only in this region does Sorbian continue to be used in everyday life and is transmitted to children which may also be monolingual in Sorbian. While German is used mainly through modern media, Sorbian remains the dominant language, both in exchanges with the local authorities and in schools. Older generations still use dialect variants but people under fifty are native speakers of Colloquial Upper Sorbian (Breu, 2000; Scholze, 2008). Colloquial Upper Sorbian may therefore be considered to be an unstable endangered language, $A^-$ in Krauss's scale (Krauss, 2006).

A Standard Upper Sorbian language, *hornjoserbska spisowna rěč,* evolved out of both Protestant and Catholic traditions from the sixteenth up until the middle of the nineteenth century. Since then, a widely unified norm has been used, based essentially on the Protestant tradition and strongly influenced by the work of grammarians and purists (Faska, 1998). In the core region we face a situation of diglossia between the two Sorbian varieties: Standard Upper Sorbian is restricted to formal speech, school lessons, church services, newspapers, as well as radio and television broadcasting, and is also used in communicating with Sorbian speakers from outside the region, whereas Colloquial Upper Sorbian is used in everyday life.

### *3.4. Burgenland Croatian, Austria*

The territory where Burgenland Croatian is spoken consists of separate enclaves, namely Burgenland and its bordering areas in Lower Austria (where the language is nowadays practically extinct), Slovakia, and Hungary (Neweklowsky, 1978: 19–21). We estimate the Slavic minority at approximately twenty thousand, but there are clearly fewer speakers of Burgenland Croatian as a shift to German is generalized (Szucsich, 2000: 874). Burgenland Croatian can be rated as $A^-$, unstable (Krauss, 2006), with children still speaking it in some localities.

A literary variety of Burgenland Croatian has been elaborated since the sixteenth century. Nowadays, the Burgenland Croatian Standard is mainly based on the Čakavian varieties, but it has been adapted in many respects to Štokavian-based Standard Croatian by language reformers and purists despite criticisms by the local communities (Szucsich, 2000: 861–869). The Burgenland Croatian Standard variety is used in schools, church services, and in the mass media, but its role in everyday life is limited (Szucsich, 2000: 869–874). In contrast, German and its Austrian variants are more present in everyday life. Notice that German was an important contact language through education since the times of the Austro-Hungarian Empire, when the Burgenland was part of Hungary (Breu J., 1970). In contrast, Hungarian was mainly used in daily interactions and in the administration. In spite of some remaining Hungarian borrowings, the influence from Hungarian seems to have been fairly limited.

## 4. Method

### *4.1. Data collection*

In this paper, we adopt a quantitative analysis of naturalistic speech which allows us to evaluate what bilinguals are actually doing when speaking (see among others Myers-Scotton, 1993; Poplack, 1993; van Hout and Muysken, 1993; Treffers-Daller, 1994; Travis and Torres Cacoullos, 2013; Adamou and Granqvist, 2014; Adamou, in press).

For the data collection we used interviews conducted by the researchers,[3] providing a total of 107 texts. Interviews targeted the collection of a variety of registers, including tales, stories about local traditions, and life-stories, with the goal of documenting the endangered Slavic languages under study. Eight texts of semi-spontaneous speech are based on visual stimuli (Pear story, Frog story, Policeman's story) for cross-linguistic comparability. The resulting corpora, with a total of 33,894 words, comprise several sub-corpora of 3,000-7,000 words each; see Figure 1.

The Molise Slavic (Na-Našu) corpus was collected and annotated by Walter Breu in the years 2000–2012. The corpus is constituted by data from three localities of the Province of Campobasso in the Region of Molise in Italy: Acquaviva Collecroce, Montemitro, and San Felice. The three sub-corpora are quite similar in size: the Acquaviva corpus is made up of 7,106 tokens, the Montemitro of 4,855 tokens, and the San Felice corpus of 5,318 tokens.

Balkan Slavic from Greece is composed of two sub-corpora. A corpus of 5,301 word tokens comes from the Balkan Slavic variety called Nashta, collected and annotated by Evangelia Adamou. The recordings took place at home, bringing together acquaintances who were asked to speak Nashta, although Greek would normally have been the language used in these circumstances and Nashta would have been used only punctually for cryptic purposes, or in order to emphasize the humoristic aspect of a story. Interviews with the researcher and one text based on the Pear Story complete the corpus. The second Balkan Slavic sub-corpus, of a total of 3,302 tokens, was collected and annotated by Georges Drettas in the 1970s in Hrisa. It is mainly composed of traditional tales signalling the vitality of the oral tradition which is lost among the last speakers of the Liti community.

The Colloquial Upper Sorbian corpus is composed of a total of 4,348 tokens. It was collected in 2012 by Lenka Scholze and Maria Utschitel and annotated by Walter Breu, Lenka Scholze, and Maria Utschitel. It consists of a variety of texts, including tales, everyday conversations, and one task of semi-spontaneous speech. All speakers use the same Colloquial Upper Sorbian variety based on the 'Catholic' dialect of the south-western Upper Sorbian area.

The corpus of Burgenland Croatian (Central South Slavic), has a total of 3,664 tokens, and was collected by Maria Utschitel and Lenka Scholze in 2012, mainly through interviews and using three tasks eliciting semi-spontaneous speech. The corpus was annotated by Walter Breu, Jasmin Meinzer, Lenka Scholze, and Maria Utschitel. The speakers were all from the Northern and Central Burgenland area.

[3] The researchers had different types of relation to the speech communities: for Molise Slavic, Balkan Slavic at Hrisa, Sorbian, and Burgenland Croatian, researchers were either community-outsiders with long-term presence in the community and excellent knowledge of the language, conducting the interviews in presence of in-group members, researchers or not; and for Balkan Slavic at Liti, the researcher was a community member with adult acquisition of the minority language, conducting some of the interviews in presence of in-group semi-speakers.

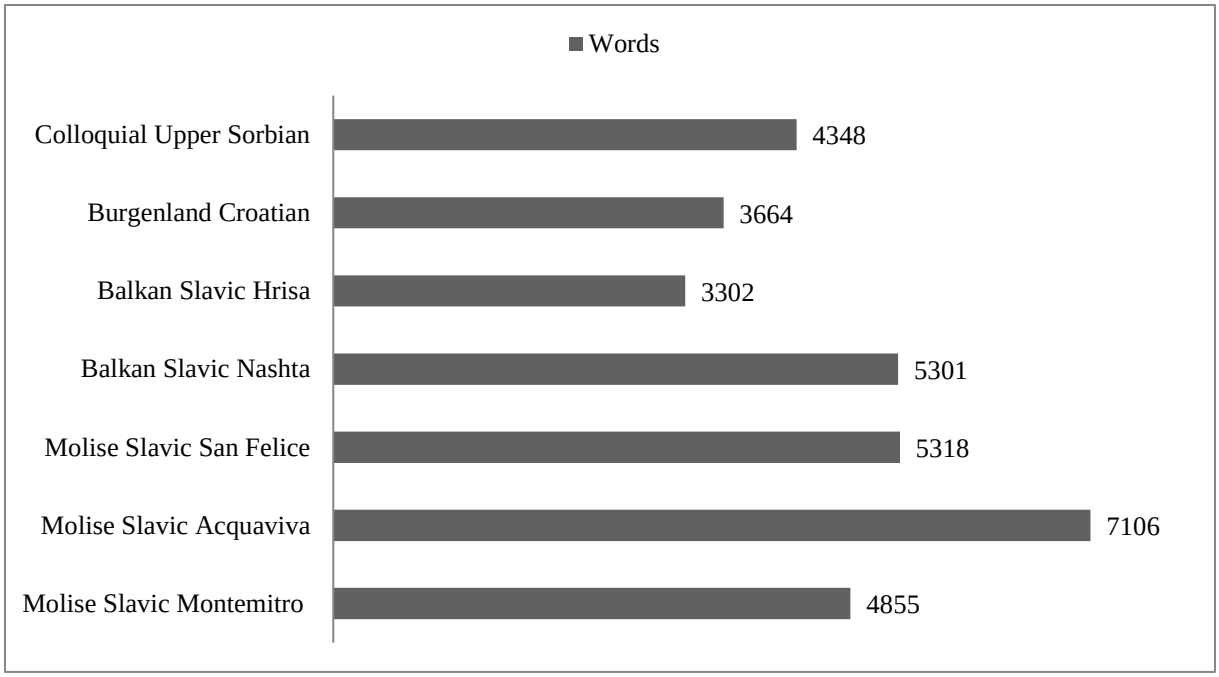



FIGURE 1. Number of words per sub-corpus (based on locality)

### *4.2. Participants*

The sample, presented in Table 1, was shaped by the acceptance of the community members to participate in the study. Participants were part of existing social networks established during previous research conducted in these villages. There was no financial compensation for participation in the study.

The Molise Slavic corpus includes data from 13 fluent speakers, 7 female and 6 male speakers, aged from 30 to 80.

The Balkan Slavic corpus is based on the speech of three fluent speakers of Nashta, aged from 75 to 85 years, two females and one male, as well as one female semi-speaker aged 56, and of two fluent female speakers from Hrisa, 40 and 60 years old.

The Colloquial Upper Sorbian corpus results from 8 different speakers, 3 female and 5 male, aged from 26 to 83, all fluent speakers of the language.

Lastly, the Burgenland Croatian corpus comprises data from 6 female and 4 male speakers, aged from 11 to 82, all fluent speakers of the language.

As the degree of bilingualism was high for all the speakers, with a single exception, this criterion was not coded for the analysis of the data. We note that age and education are strongly correlated, with older speakers being less educated than younger speakers. Education was thus not treated as a separate factor and age was preferred instead since it allowed for a fine-grained classification of the speakers in three groups (young, middle, and old).

TABLE 1. The sample of the EuroSlav corpus

| | MALE | FEMALE | YOUNG ≤ 49 | MIDDLE 50-69 | OLD ≥ 70 |
|---|---|---|---|---|---|
| MOLISE SLAVIC | 6 | 7 | 2 | 4 | 7 |
| BALKAN SLAVIC | 1 | 5 | 1 | 2 | 3 |
| BURGENLAND CROATIAN | 4 | 6 | 2 | 3 | 5 |
| COLLOQUIAL UPPER | 5 | 3 | 6 | 1 | 1 |

| | | |
|---|---|---|
| SORBIAN | | |

### *4.3. Annotation*

The corpora were transcribed, glossed, and translated into English and partly into French, as well as into the respective contact-languages, namely Italian, Greek, and German. We used the authoring tool Interlinear Text Editor (ITE), developed by Michel Jacobson at LACITO-CNRS. The files were then synchronised with SoundIndex, a tool for time-aligning XML-formatted text annotation developed by Michel Jacobson.

The annotated corpora and their sound files are all available online at the Pangloss collection,[4] a language archive hosted at the research centre *Oral Tradition Languages and Civilizations* (LACITO) of the French National Centre for Scientific Research (CNRS). The audio recordings may be listened to in their entirety or sentence by sentence. The visitor may set parameters so that the transcription appears in parallel with the sound recordings, either sentence by sentence, or with the transliteration and morphosyntactic glosses provided by the grammatical analysis of the language.

All codeswitching insertions and lexical borrowings from the current-contact languages, either from the standard or the regional variety, were tagged based on four diagnostics: length of insertion, phonological, syntactic, and morphological integration. Words which could come from either the current-contact language or a past-contact language were treated separately and were tagged as 'multiple', i.e., words present in more than one contact languages. The phonologically and morphologically non-integrated or multi-word insertions from the current-contact languages were tagged as 'code-switching', as shown in (1).

(1) Excerpt from the Nashta corpus; Slavic (in plain), Greek (in italics), Turkish (underlined); codeswitching in angle brackets

Female fluent speaker:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| a. | *aˈla* | uˈno | saˈxat | ta | a= | ˈtʃini-ʃe | ˈona | |
| | but | DEM.DIST.SG.N | time/hour.M | DEM.MID.SG.F | ACC.3SG.F- | do-IPRF.3SG | DEM.DIST.SG.F | |

'But at that moment, she was doing,'

| | | | | |
|---|---|---|---|---|
| b. | a= | ˈtʃeʃ-ʃe | *to* | *ˈxtenize* |
| | ACC.3SG.F- | card-IPRF.3SG | ACC.3SG.N | card.IPRF.3SG |

'she was carding it, <she was carding it.>'

Female semi-speaker:

| | | | | |
|---|---|---|---|---|
| c. | *to* | *ˈxtenize* | *to* | *maˈli* |
| | ACC.3SG.N | card.IPRF.3SG | ART.3SG.N | wool |

'<She was carding the wool.>'

Female fluent speaker:

| | | | | | | |
|---|---|---|---|---|---|---|
| d. | *na* | *sas* | *ˈδoso* | *na* | *kataˈlavete* | *ˈtora* |
| | to | 2PL.ACC | give.1SG | to | understand.2PL | now |

'<So that you really understand,'

[4] http://lacito.vjf.cnrs.fr/pangloss/

e. *to* *ˈxtenize* *to* *maˈli*
ACC.3SG.N card.IPRF.3SG ART.3SG.N wool
'<She was carding the wool.>'

Researcher:
f. *ne* *kala*
yes fine
<Yes...it's OK.>
(Adamou, 2013, 'Le petit Dimitro', sentence 11. Accessed online at http://lacito.vjf.cnrs.fr/pangloss)

The phonologically, syntactically, and morphologically integrated, single or two-word insertions were tagged as 'borrowings'. For example, in Molise Slavic, one and the same speaker alternates for 'Little Red Riding Hood' between an integrated form, tagged as a borrowing, i.e., [kapuˈtʃeto ˈros] (in sentences 1 and 41), and a phonologically and morphologically non-integrated form, with Italian gemination and morphology, tagged as a codeswitch insertion, i.e., [kapputˈtʃetto ˈrosso] (in sentence 3) http://lacito.vjf.cnrs.fr/pangloss/ Na-našu, Acquaviva, Text 'Petit chaperon rouge'.

An example illustrating borrowings from each corpus is shown in (2)–(5), but the corpora are all available on line in open access.

(2) Molise Slavic < Slavic (in plain), Italian (in italics)

| | | | | |
|---|---|---|---|---|
| a. | ˈaje-*ka* | ˈbi-x-u | *fan ˈdaːzm-a* | |
| | because-COMP | be-IPRF-3PL | ghost-NOM.PL.M | |
| b. | *ma* | káːkạ | ˈbi-x-u | *fan ˈdaːzm-a* |
| | but | how | be-IPRF-3PL | ghost-NOM.PL.M |
| c. | je= | ˈrek-l-a | mo | ˈmat |
| | be.PRS.3SG-AUX.PRF(1) | say.PFV-PTCP-SG.F PRF(1) | my.NOM.F | mother.NOM.F |

'Because there were ghosts! What? There were ghosts? said my mother.'
(Breu, 2013. Excerpt from 'L'âne et les fantômes', sentences 12 and 13, Accessed online at http://lacito.vjf.cnrs.fr/pangloss/index.htm#europe)

(3) Balkan Slavic Nashta (Greece) < Slavic (in plain), Greek (in italics), Multiple (underlined)

| | | | | |
|---|---|---|---|---|
| a. | tʃuˈban-at | ˈpasʲ-ʃe | ˈvoftsʲ-e-te | diˈʎeko |
| | shepherd-ART.SG.M | graze-IPRF.3SG | sheep-ART.PL | far |
| b. | i | *cini ˈsa* | iˈdno | den |
| | and | get.going.AOR.LVM.3SG | one.SG.N | day.SG.M |

'A shepherd was grazing his sheep far away, and one day, he got going...'
(Adamou, 2013. Excerpt from 'Le berger et son ombre', sentences 2 and 3, Accessed online at http://lacito.vjf.cnrs.fr/pangloss/index.htm#europe)

(4) Colloquial Upper Sorbian < Slavic (in plain), German (in italics)

| | | | | | |
|---|---|---|---|---|---|
| a. | *alzɔ* | pɔ | nas | wɛ | swʊ̯ə̯jb-ɛ |
| | well | at | we.GEN | in | family-LOC.SG.F |
| b. | da-w-ɛ | jɛn | nawɔʃk | | |
| | give-IPFV-3SG | ART.INDF.NOM.SG.M | custom.NOM.SG.M | | |

'So, in our family, there is a custom.'

(Breu, Scholze, and Utschitel 2013. Excerpt from 'Pâques chez les sorabes', sentence 1, Accessed online at http://lacito.vjf.cnrs.fr/pangloss/index.htm#europe)

(5) Burgenland Croatian < Slavic (in plain), German (in italics)

| | | | | | |
|---|---|---|---|---|---|
| a. | tako | ˈpraːv-o | ko=se | ur | ˈpi̯etak |
| | so.MID | custom-NOM.N | PTL-REFL | already | Friday.ACC |
| b. | ˈnuo̯t-iː | mblaˈdiːn-a | *ˈstref-i* | | |
| | night-LOC.SG.F | youth-NOM.SG.F | meet.PFV-PRS.3SG | | |

'...that (is) the custom. So on Friday at night the youth already meet...'
(Breu et al., 2013. Excerpt from 'Le marriage en Burgenland', sentence 2, Accessed online at http://lacito.vjf.cnrs.fr/pangloss/index.htm#europe)

### *4.4. Statistical analyses*

For the analysis of the Slavic data, we examined the following dependent and independent variables: the independent variables (fixed or random) are Speaker, Sex (male or female), Age, Language (Molise Slavic, Balkan Slavic, Burgenland Croatian, and Colloquial Upper Sorbian), Locality (13 locations), Text (spontaneous or semi-spontaneous through elicitation), and Type (nouns or non-nouns); and the dependent variables are whether the participants successfully ('success') or not ('failure') produced a borrowing.

The statistical analyses performed on the two dependent variables are identical. We first ran linear mixed models to explore the data in detail and we then ran Random Forests (Breiman, 2001; Tagliamonte and Baayen, 2012) to explore the most effective factor(s). For the sake of visual inspection, we present the data using the proportion (percentage) of borrowed words, but for the statistical analyses, we used the actual counts of borrowings and non-borrowings as the response variable.

## 5. Results

In the following sections we present the results of the quantitative and statistical analysis of the corpora for the borrowings, in 5.1., and the noun borrowings, in 5.2.

### *5.1. Borrowings*

To determine the proportion of borrowings from the current-contact language, whether content words or grammatical words, we counted all the words shared with the languages in contact and calculated the percentage they represent of the total number of words in the corpus. As can be seen in Figure 2, the Colloquial Upper Sorbian, Burgenland Croatian, and Balkan Slavic corpora show less than 5% tokens from their respective current-contact languages. The corpus of Molise Slavic shows 22.6% borrowings from Italian.

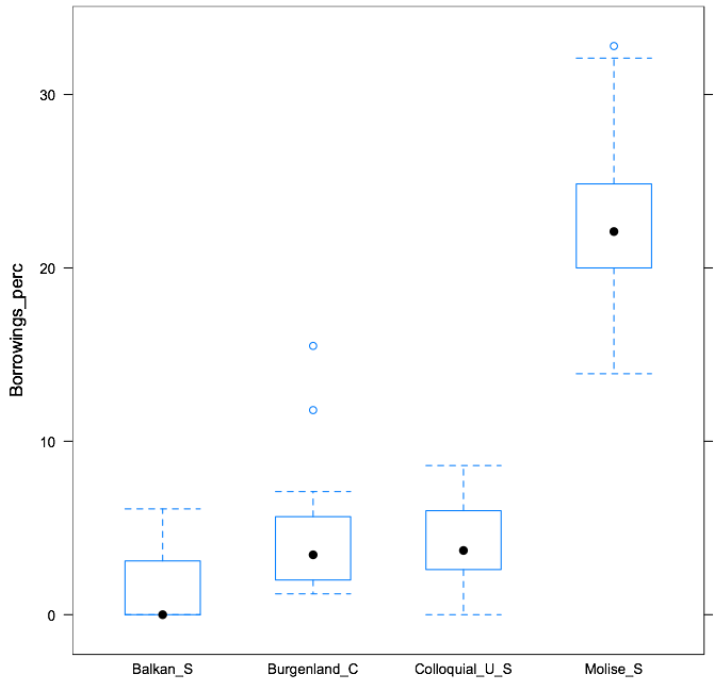



FIGURE 2. Borrowings with respect to Language

Visual inspection of the graph in Figure 3 shows clearly that the rates of Italian borrowings in the three Molise Slavic corpora of Acquaviva, Montemitro, and San Felice are similar, on average 22–23%. Furthermore, among the Balkan Slavic varieties of Greece, the Nashta corpus shows 6% Greek borrowings, while the corpus of the 1970s from Hrisa shows practically no Greek tokens. As the two corpora were recorded more than 30 years apart, the result is more likely to be related to the moment of recording rather than to the locality.

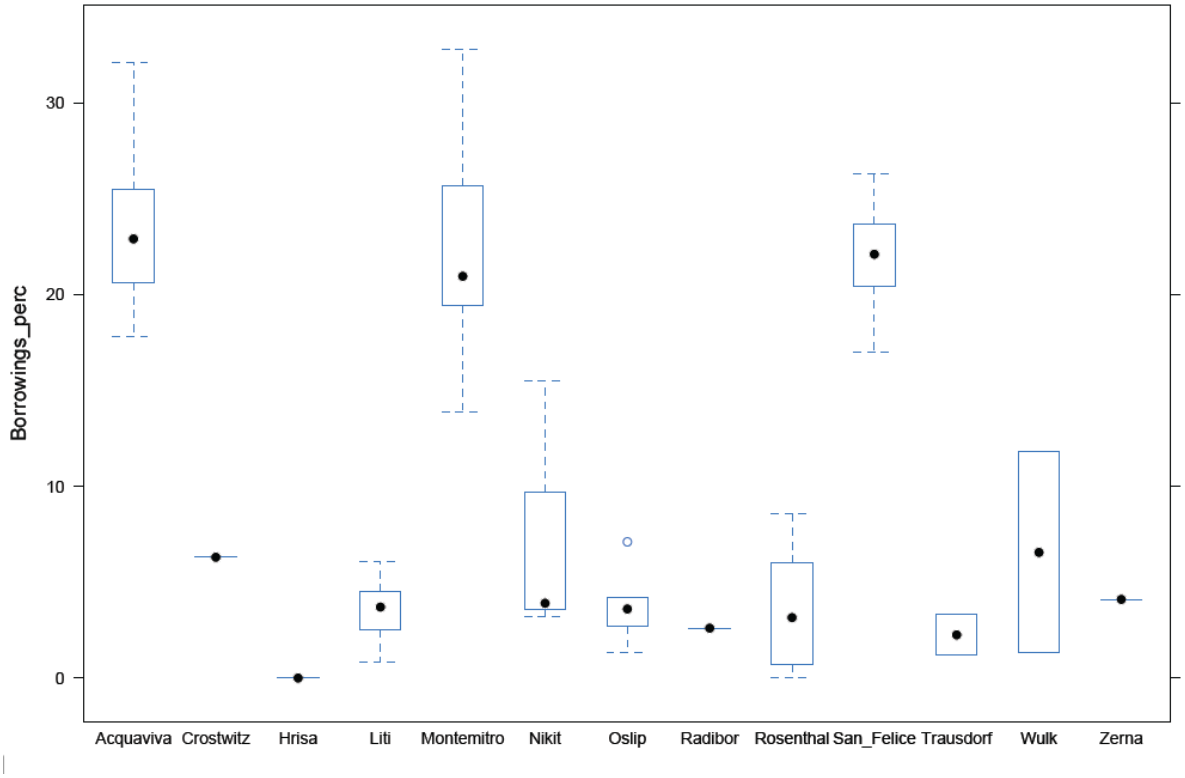



FIGURE 3. Borrowings with respect to Locality

We also examined the relation of the production of Borrowings and extra-linguistic factors such as Sex and Age. Figure 4a presents the rates of borrowings within each community with respect to Sex, and Figure 4b with respect to Age. It can be seen that there is no clear difference for the speakers within each community with respect to their sex and age.

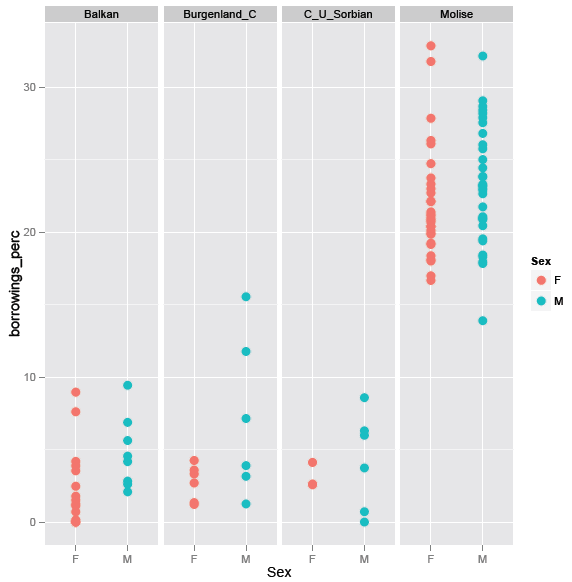



FIGURE 4a. Borrowings with respect to Language and Sex

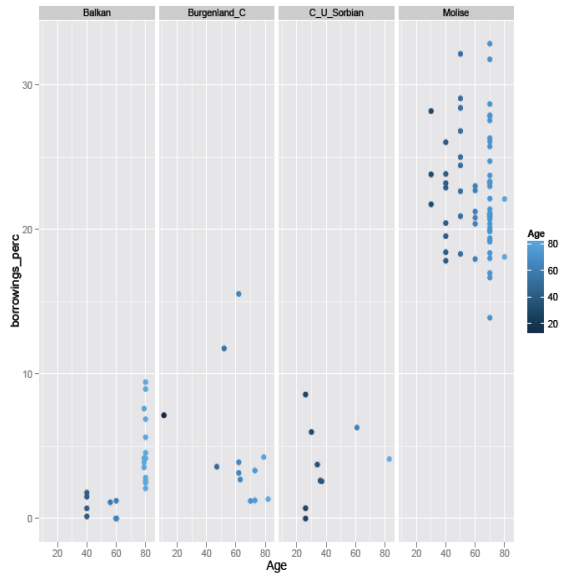



FIGURE 4b. Borrowings with respect to Language and Age

In order to explore more detailed relationships between various factors and the production of Borrowings, several (general) linear mixed models with binomial family were constructed. The dependent variable is the odd ratio of Borrowings to Non-borrowings. The fixed factors are Sex, Age, and Language (Balkan Slavic, Burgenland Croatian, Colloquial Upper Sorbian, and Molise Slavic), Type (nouns or non-nouns) and Text (spontaneous or elicited). The random factors are Speaker and Locality. They are treated as random categorical factors as they do not exhaust the population and are nonsystematic, idiosyncratic, and unpredictable in the current study. From simple to complex, we built several models, adding each fixed factor and possible interaction one at a time. ANOVA showed that Language alone increased the model's predictability significantly ($\chi^2(3) = 31.1$, $p < .001$). There were significantly more Non-noun Borrowings than Noun Borrowings ($\chi^2(1) = 1218.7$, $p < .001$). There is a significant interaction of Type by Language ($\chi^2(3) = 77.1$, $p < .001$).

TABLE 2. The mean values and standard deviation of Types of Borrowings and Non-borrowings in the four Slavic languages

| LANGUAGE | BORROWINGS NOUNS | BORROWINGS NON-NOUNS | NON-BORROWINGS NOUNS | NON-BORROWINGS NON-NOUNS |
|---|---|---|---|---|
| BALKAN SLAVIC | 5.09 (5.4) | 5.91 (5.2) | 91.13 (79.5) | 246.87 (164.2) |
| BURGENLAND CROATIAN | 8.75 (6.4) | 2.91 (2.1) | 43.00 (22.5) | 250.67 (121.3) |
| COLLOQUIAL UPPER SORBIAN | 10.11 (11.5) | 10.56 (11.1) | 59.00 (24.9) | 403.44 (173.8) |
| MOLISE SLAVIC | 20.32 (15.1) | 41.65 (29.1) | 22.79 (17.4) | 190.98 (133.7) |

Molise Slavic speakers produce the highest ratio of Borrowings ($Z = 7.2$, $p < .001$), and Burgenland Croatian speakers produce more Borrowings than Balkan speakers ($Z = 2.6$, $p < .007$). However, the differences between Colloquial Upper Sorbian and Burgenland Croatian speakers, and the differences between Burgenland Croatian and Balkan Slavic speakers, respectively, are not significant ($ps > .1$).

For Molise Slavic speakers, there are significant interactions of Age by Type (Noun Borrowings or Non-noun Borrowings) ($\chi^2(0) = 5.6$, $p < .001$). This means that the differences in the production of Noun and Non-noun Borrowings are larger for younger speakers than those of older speakers ($z=2.62$, $p = .008$); see Table 3.

TABLE 3. The mean values and standard deviation of Types of borrowings with respect to Age in Molise Slavic

| | 30 Y.O. | 40 Y.O. | 50 Y.O. | 60 Y.O. | 70 Y.O. | 80 Y.O. |
|---|---|---|---|---|---|---|
| BORROWINGS NON-NOUNS | 76.3 (43.8) | 27.3 (8.9) | 36.3 (15.0) | 46.5 (44.1) | 41.9 (30.0) | 53 (8.5) |
| BORROWINGS NOUNS | 34.3 (14.6) | 10.9 (4.3) | 16.3 (7.5) | 28.0 (31.9) | 20.6 (13.5) | 27 (5.7) |

For Burgenland Croatian speakers, there are significant interactions of Age by Type of borrowings ($\chi^2(0) = 1.7$, $p < .001$), and three-way interactions of Age by Type and by Sex ($\chi^2(3) = 8.8$, $p = .03$). In younger age groups, the differences for Noun Borrowings and Non-noun Borrowings are larger for male speakers than for female speakers ($z=2.6$, $p = .009$).

TABLE 4. The mean values of Types of borrowings with respect to Age and Sex in Burgenland Croatian[5]

| | | 11 Y.O. | 47 Y.O. | 52 Y.O. | 62 Y.O. | 63 Y.O. | 70 Y.O. | 73 Y.O. | 79 Y.O. | 82 Y.O. |
|---|---|---|---|---|---|---|---|---|---|---|
| FEMALE | BORROWINGS NON-NOUNS | NA | 1 | NA | NA | 2 | 0 | 2 | 3 | 5 |
| | BORROWINGS NOUNS | NA | 11 | NA | NA | 8 | 3 | 5 | 8 | 1 |
| MALE | BORROWINGS NON-NOUNS | 5 | NA | 2 | 2 | NA | NA | 7 | NA | NA |
| | BORROWINGS NOUNS | 15 | NA | 18 | 11 | NA | NA | 1 | NA | NA |

For Colloquial Upper Sorbian speakers, there are significant interactions of Age by Type of borrowings ($\chi^2(0) = 1.7$, $p < .001$), and the interaction of Type by Sex ($\chi^2(1) = 156$, $p < .001$). This means that the differences for Noun Borrowings and Non-noun Borrowings are smaller for male speakers than for female speakers ($z=2.6$, $p= .01$).

TABLE 5. The mean values and standard deviation of Types of borrowings with respect to Sex in Colloquial Upper Sorbian

| | FEMALE | MALE |
|---|---|---|
| BORROWINGS NON-NOUNS | 8.3 (7.6) | 11.7 (13.1) |
| BORROWINGS NOUNS | 4.3 (1.2) | 13.0 (13.5) |

For Balkan Slavic Nashta speakers, there are no significant interactions or main effects between the subject groups ($\chi^2$s < 1).

Finally, data exploration with Random Forests (Breiman 2001) was used to explore the relevant factors for all the communities. Following Tagliamonte and Baayen (2012), the percentage of Borrowings is the dependent variable and the random predictors are: Language (four sub-language groups), Locality (13 different locations of recording), Speaker, Sex, Text (elicited or spontaneous), Recording sessions, and Age. Based on Tagliamonte and Baayen (2012), a model of Random Forests using the function of *cforest* (Hothorn et al., 2006; Strobl et al., 2007; Strobl et al., 2008) was built with *R* (R Core Team, 2013) package *Party*. The model has a Concordance Index of 0.83. The rank of the importance of the variables was calculated and is presented in Figure 5 which shows that the most important factor is Language and to a lesser extent Locality. However, we note that the factor Locality, graphed in Figure 3, is not a very good fixed grouper, as there are some localities represented by a very small number of speakers.

[5] There are some NA values because of the imbalance for male and female speakers. Also for this reason, the SD cannot be computed.

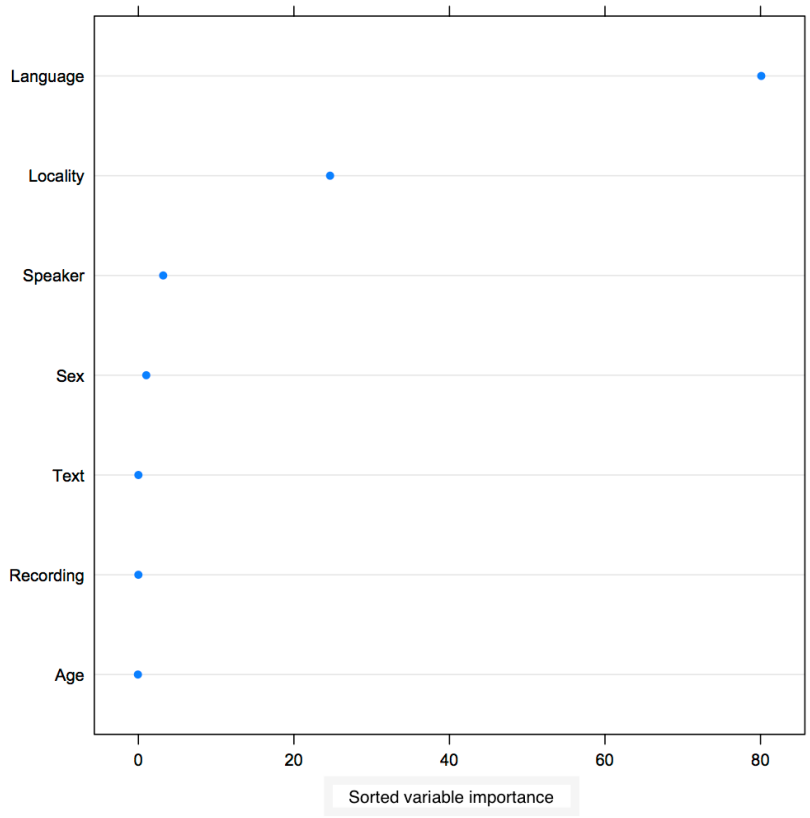



FIGURE 5. Importance ranking of variables (Random Forests) for Borrowings

### *5.2. Noun borrowings*

As the use of borrowed content words typically relies on the topic of conversation and is expected to show more variability than that of borrowings including all the word-classes, we conducted a study of inter-speaker variation to explore the usage of noun borrowings from the current-contact languages. Analyses similar to those conducted for borrowings were also carried out for the noun borrowings.

As Figure 6 shows, Balkan Slavic speakers used 10% Noun Borrowings within the noun word-class, Colloquial Upper Sorbian speakers used 13% Noun Borrowings, Burgenland Croatian speakers 22.4% Noun Borrowings, and Molise Slavic speakers 46.1% Noun Borrowings.

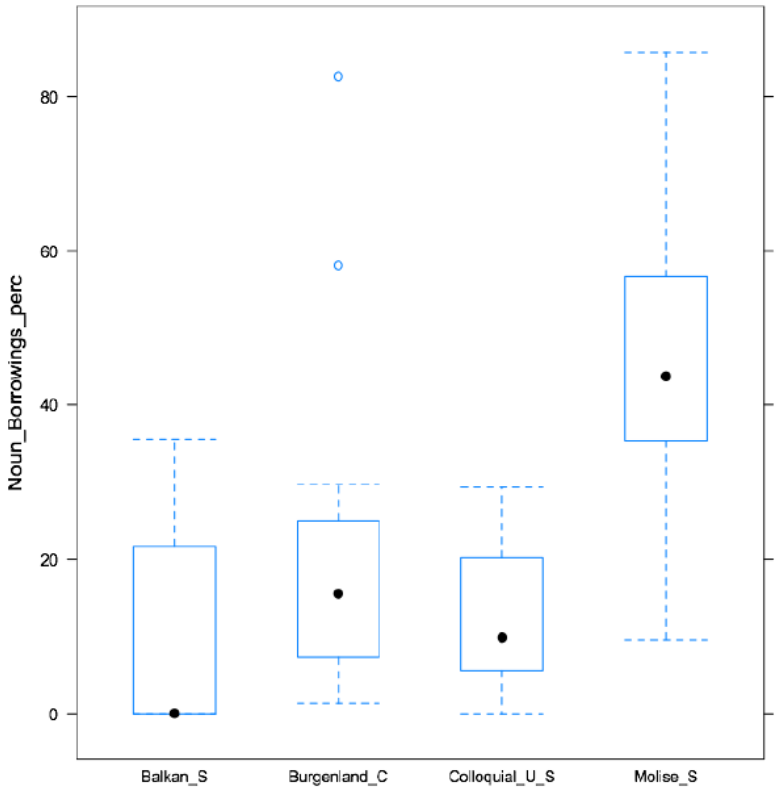



FIGURE 6. Noun Borrowings with respect to Language

The differences and similarities in the rates of Noun Borrowings depending on the Locality appear clearly in Figure 7. It appears that there is little variation in the production of Noun Borrowings for the three Molise Slavic communities, namely for Acquaviva, Montemitro, and San Felice. The graph also shows the difference in Balkan Slavic between the sub-corpus of the 2000s recorded in Liti and the sub-corpus of the 1970s recorded in Hrisa, the latter having practically no Greek nouns.

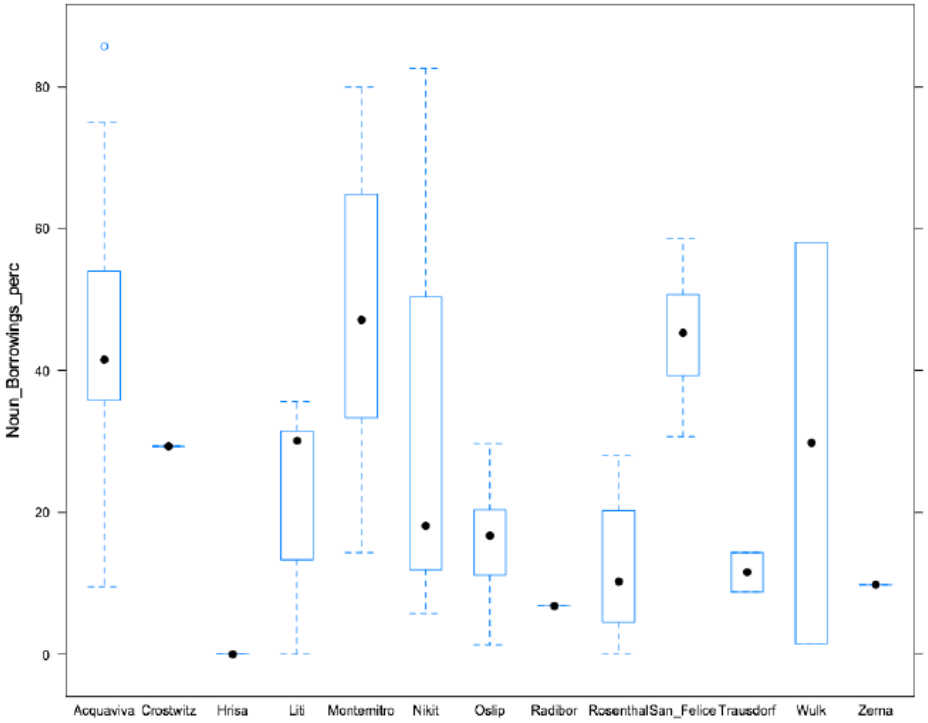



FIGURE 7. Noun Borrowings with respect to Locality

Figures 8a and 8b show the distribution of Noun Borrowings for each language group with respect to Sex and Age. Similar to the results for Borrowings, it can be seen that there is no clear-cut differentiation of Noun Borrowing production within each language community with respect to Sex and Age.

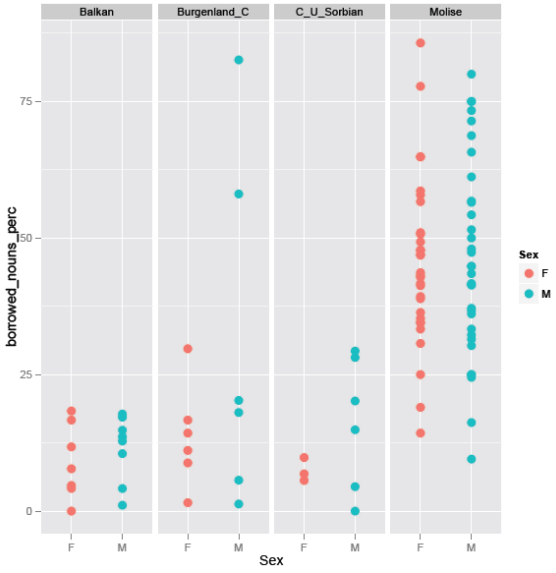



FIGURE 8a. Noun Borrowings with respect to Language and Sex

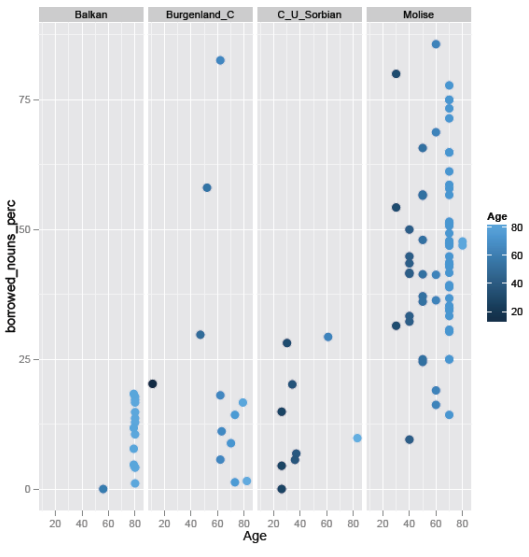



FIGURE 8b. Noun Borrowings with respect to Language and Age

In order to explore more detailed relationships between various factors and the production of Noun Borrowings, several (general) linear mixed models with binomial family

were constructed. The dependent variable is the log odds ratio of Noun Borrowings to Noun Non-borrowings. The fixed factors are Sex, Age and, Language (Balkan Slavic, Burgenland Croatian, Colloquial Upper Sorbian, and Molise Slavic), Type (nouns or non-nouns), and Text (spontaneous or elicited). The random factors are Speaker and Locality. They are treated as random categorical factors as they do not exhaust the population and are nonsystematic, idiosyncratic, and unpredictable in the current study.

Seven models were built to test the fixed and random effects. ANOVA showed that, similar to the results for Borrowings, Language alone increased the model's predictability significantly ($\chi^2(3) = 21.6$, $p < .001$). However, none of the other factors predicts the usage of Nouns Borrowing ($p$s > .1) and there are no interactions ($p$s > .3). Molise Slavic speakers produced the most Noun Borrowings (Zs > 5, $p$s < .001).

We then broke down the whole dataset into subgroups by languages to explore the results in detail. For Molise Slavic and Burgenland Slavic speakers, there are no significant interactions or main effects between the various subject groups ($\chi^2$s < 1).

For Colloquial Upper Sorbian speakers, there is a marginal difference for Sex ($\chi^2(2) = 5.9$, $p = .05$) with male speakers using more noun borrowings than female speakers (z=2.8, $p = .004$); see Table 6.

TABLE 6. The mean values and standard deviation of Noun Borrowings with respect to Sex in Colloquial Upper Sorbian

| | FEMALE | MALE |
|---|---|---|
| NOUNS | 4.3 (1.2) | 13.0 (13.5) |

For Balkan Slavic speakers, there is a significant effect for Sex ($\chi^2(1) = 11.6$, $p < .001$), with the male speaker having significantly more noun borrowings than the female speakers (z=2.4, p= .01); see Table 7. However, this result should be interpreted by taking in consideration the fact that the sub-corpus from Hrisa which was recorded in the 1970s does not include any male speakers.

TABLE 7. The mean and standard deviation of Noun Borrowings with respect to Sex in Balkan Slavic

| | FEMALE | MALE |
|---|---|---|
| NOUNS | 3.1 (4.9) | 8.9 (4.2) |

Last, data exploration by Random Forests (Breiman, 2001) was used to evaluate the relevance of various predictors for the entire dataset. Similar to Borrowings, the percentage of Noun Borrowings is the dependent variable, and the random predictors are Language (Balkan Slavic, Burgenland Croatian, Colloquial Upper Sorbian, and Molise Slavic), Locality (13 different locations of recording), Speaker, Sex, Age, Text (spontaneous or semi-spontaneous through elicitation), and Recording sessions. The graph in Figure 9 shows the dot plots for the rank of importance. Visual inspection shows that the most important predictor for Noun Borrowings is Language, similar to the results of the analysis of Borrowings. The model has a Concordance Index of 0.80.

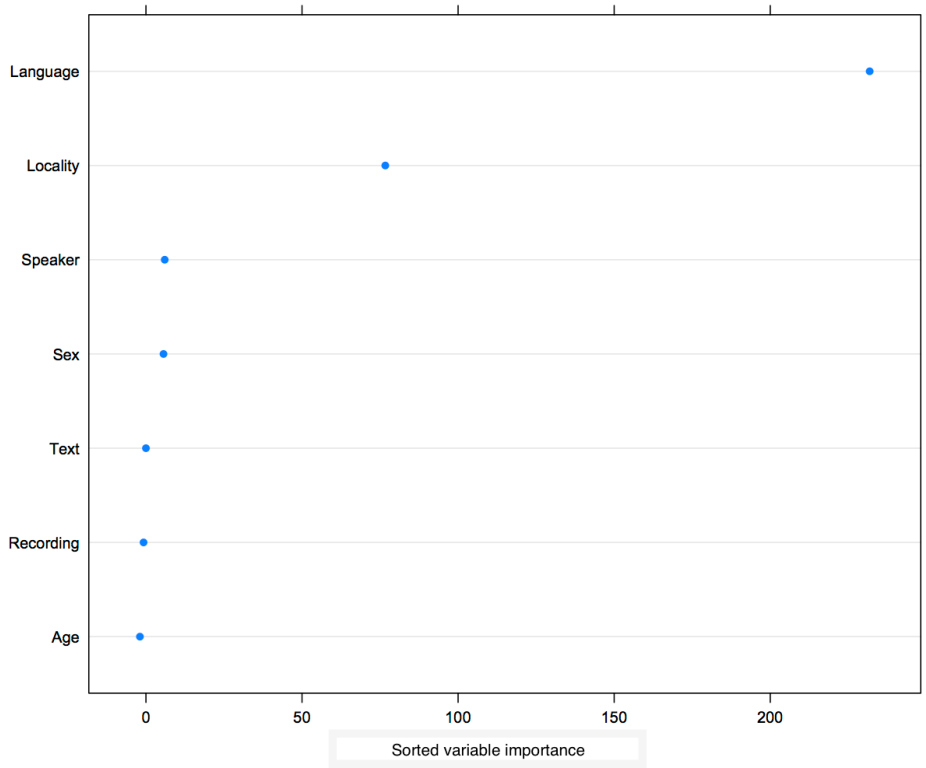



FIGURE 9. Importance ranking of variables (Random Forests) for Noun Borrowings

## 6. Discussion

The statistical analysis of the four Slavic spoken corpora shows that the best predictor for the ratio of Borrowings is Language, as Molise Slavic speakers show considerably different patterns of borrowing than speakers from Balkan Slavic, Burgenland Croatian, and Colloquial Upper Sorbian. The fact that individual speakers conform to the patterns prevailing in their bilingual communities is also discussed in other studies on language contact based on naturalistic data (see Poplack, 1985; Adamou and Granqvist, 2014; Travis and Torres Cacoullos, in press).

More generally, the analysis of the Slavic corpora shows that some languages have very low proportions of borrowings from the current-contact language, while others have significantly higher proportions of borrowings. More specifically, the quantitative analysis reveals that the Slavic-speaking communities from Greece, Germany, and Austria produce very few borrowings in their spontaneous or semi-spontaneous speech, i.e., less than 5%, as opposed to the Molise Slavic communities from Italy which use on average 22.6% Italian words. We suggest referring to the first group of languages as ‘low borrowers’ and to languages such as Molise Slavic as ‘high borrowers’ in accordance to the typology introduced in Tadmor (2009).

Notice that the higher or lower ratio of borrowings in the Slavic corpora under study does not in any way reflect the speakers’ ability in any of the languages in contact. In all cases, with one exception, the participants were fluent speakers of the Slavic and of the non-Slavic languages in contact.

It is also interesting to note that the degree of borrowing established in this study and the extent of convergence, which was discussed in other studies, do not always coincide. For

example, Molise Slavic communities from Italy are not only high borrowers, as our study shows, but their language shows in addition a great number of contact-induced changes on all levels of grammar, including the systems of gender and case, verbal aspect and tense, modality, the development of an indefinite article, and word order (Breu 2011). However, in the Colloquial Upper Sorbian corpus, despite the low proportion of borrowings that our study reveals, the influence of German is significant through covert strategies such as semantic extension, calques and convergence, i.e., personal pronouns have become obligatory, an article system has evolved with a regularly used definite and indefinite article, the dual has been reduced to a dependent form governed by the number 'two' and the pronoun 'both', the passive is formed with the auxiliary *hodwać* borrowed from German *werden* 'become', an expletive *to* has been introduced by calquing German *es* 'it', and a length opposition has developed in the vowel system (Scholze, 2008; Breu, 2012).

With respect to the social factors, we note that age and sex appear to be significant within some communities, but the weight of these factors in the Random Forests analysis is secondary.

In order to account for the differences in the above-mentioned borrowing patterns we now turn to qualitatively discuss three types of extralinguistic factors: the degree of everyday use, literary tradition and prescriptive attitudes, and institutional support for the Slavic languages under study. Table 8 summarizes the high or low proportion of borrowings and the degree of everyday use, the existence of a literary tradition and prescriptive attitudes, as well as institutional support in the present and the past (coded as 'high', 'low', and 'variable' when the factor depends on the locations).

TABLE 8. Extralinguistic factors with respect to ratio of Borrowings in the Slavic corpora

| | EVERYDAY USE OF THE TWO LANGUAGES IN CONTACT | | LITERARY TRADITION AND PRESCRIPTIVE ATTITUDES | | INSTITUTIONAL SUPPORT | | RATES OF BORROWINGS IN THE CORPORA |
|---|---|---|---|---|---|---|---|
| | *present* | *past* | *present* | *past* | *present* | *past* | |
| COLLOQUIAL UPPER SORBIAN | high | high | high | high | high | variable | low |
| MOLISE SLAVIC | variable | high | variable | low | low | low | high |
| BURGENLAND CROATIAN | variable | variable | variable | variable | variable | variable | low |
| BALKAN SLAVIC | variable | low | low | low | low | low | low |

The absence of everyday use of the Slavic language in the present setting may account for the low proportion of borrowings for the last speakers of Balkan Slavic Nashta recorded in Greece. Indeed, the last fluent speakers of Nashta no longer use the language in their daily interactions and prefer to discuss specific topics directly related to traditions and past-life events. However, the low proportion of Greek words in Balkan Slavic Nashta seems to reflect a pattern which was established when the language was still actively spoken. Indeed, the speakers of Balkan Slavic recorded in Hrisa in the 1970s produce practically no Greek borrowings, despite the fact that at the time the language was still actively spoken by the parental and grandparental generations. The small amount of Greek words probably indicates that Greek was not an everyday contact language during Ottoman times, prior to the

integration of the Slavic-speaking populations in the Greek state in the early-twentieth century.

However, everyday use of the two languages in contact, in the past or present, is not a good predictor for the rates of borrowings. We observe for example that active bilingual communities, such as the Colloquial Upper Sorbian and Burgenland Croatian communities, show very low rates of borrowings from German. In order to understand the relatively low proportion of German words in Colloquial Upper Sorbian despite the vitality of the bilingual community, one must take into consideration the strong linguistic tradition of Sorbian intellectuals who have promoted the use of Standard Upper Sorbian. Thus, the avoidance of German borrowings can be understood as the result of a long tradition led by prescriptive Sorbian intellectuals, which is still powerful in the contemporary setting. Similarly, the low proportions of borrowings in the Burgenland Croatian corpus may be due to the influence of Standard Burgenland Croatian and a long literary tradition.

In contrast, Molise Slavic speakers produce a great amount of borrowings from Italian. The high rates of Italian borrowings in Molise Slavic seem to have been established in the past and do not result from the present use of the two languages. Indeed, everyday use of Molise Slavic varies depending on the locality, but the rates of borrowings are similar in all three communities. Moreover, Molise Slavic speakers had no literary traditions in the past and the influence of prescriptive attitudes in the contemporary setting, namely through education, remains marginal.

To conclude, the everyday use of two languages at the level of the community, in the present or in the past, is not a good predictor for the low or high rates of borrowing found in a bilingual corpus. The results of the Slavic datasets analysed in this paper suggest that bilingual communities with everyday use of the two languages in contact do not necessarily show high rates of borrowings. The settings for such 'low borrowers' are characterised by century-long prescriptive attitudes promoted by local intellectuals or the authorities, as is the case of Colloquial Upper Sorbian and Burgenland Croatian. Indeed, other processes of lexical creation can be activated under certain circumstances, most particularly when there is institutional support for the minority languages combined with a long literary tradition. In contrast, the settings in which we find 'high-borrowers', such as the Molise Slavic communities, are characterized by lack of normative pressure combined with century-long, everyday use of the languages in contact.

## Abbreviations

Glosses follow the Leipzig Glossing Rules when applicable.
ACC accusative; AOR aorist; ART article; AUX auxiliary; COMP complementizer; GEN genitive; INDF indefinite; IPRF imperfect; LOC locative; LVM loan verb marker; M masculine; MID middle; NOM nominative; REFL reflexive; PFV perfective; PL plural; PRS present; PTCP participle; PTL particle.

## Acknowledgments

This research received funding from the French National Research Agency (ANR) and the Deutsche Forschungsgemeinschaft (DFG) project 'Electronic database of endangered Slavic varieties in non-Slavic speaking European countries' (ANR-09-FASHS-025 and DFG BR 1228/4-1), 2010-2012. Evangelia Adamou also received support for the analysis of the data from the French National Research Agency (ANR) for the programme 'Empirical Foundations of Linguistics' (ANR-10-LABX-0083). A special thank-you to the speakers of the languages under study for their collaboration in this project. Thanks are due to Séverine Guillaume for

technical assistance with the Pangloss database. We would also like to thank Harald Baayen for his precious advice on the statistical analyses, and the anonymous reviewers for their insights.

## References

Adamou, Evangelia. 2013. Nashta Corpus. In Evangelia Adamou, Walter Breu, Georges Drettas, and Lenka Scholze (eds.), Electronic database of endangered Slavic varieties in non-Slavic speaking European countries, ANR-DFG EuroSlav2010. http://lacito.vjf.cnrs.fr/pangloss/ (accessed 23 June 2014).

Adamou, Evangelia and Granqvist, Kimmo. 2014. Unevenly mixed Romani languages. *International Journal of Bilingualism*. Prepublished, March 2014. doi:10.1177/1367006914524645

Adamou, Evangelia. In press, 2016. *A corpus-driven approach to language contact: Endangered languages in a comparative perspective.* Berlin and New York: de Gruyter.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45: 5–32.

Breu, Joseph. 1970. *Die Kroatensiedlung im Burgenland und in den anschließenden Gebieten.* Wien: Franz Deuticke.

Breu, Walter. 2000. Der Verbalaspekt in der obersorbischen Umgangssprache im Rahmen des ILA-Modells, In Walter Breu (ed.), *Slavistische Linguistik* 1999, 37–76. München: Otto Sagner.

Breu, Walter. 2011. Language contact of minority languages in Central and Southern Europe: a comparative approach. In Bernd Kortmann and Jan van der Auwera (eds.), *The Languages and Linguistics of Europe. A Comprehensive* Guide, 429-451. Berlin/New York: De Gruyter.

Breu, Walter. 2012. Aspect forms and functions in Sorbian varieties. *Language typology and universals (STUF)* 65(3): 246–266.

Breu, Walter. 2013. Molise Slavic Corpus. In Evangelia Adamou, Walter Breu, Georges Drettas, and Lenka Scholze (eds.), Electronic database of endangered Slavic varieties in non-Slavic speaking European countries, ANR-DFG EuroSlav2010. http://lacito.vjf.cnrs.fr/pangloss/ (accessed 6 July 2014).

Breu, Walter, Jasmin Meinzer, Lenka Scholze, and Maria Utschitel. 2013. Burgenland Croatian Corpus, In Evangelia Adamou, Walter Breu, Georges Drettas, and Lenka Scholze (eds.), Electronic database of endangered Slavic varieties in non-Slavic speaking European countries, ANR-DFG EuroSlav2010. http://lacito.vjf.cnrs.fr/pangloss/ (accessed 6 July 2014).

Breu, Walter, Lenka Scholze, and Maria Utschitel. 2013. Colloquial Upper Sorbian Corpus. In Evangelia Adamou, Walter Breu, Georges Drettas, and Lenka Scholze (eds.), Electronic database of endangered Slavic varieties in non-Slavic speaking European countries, ANR-DFG EuroSlav2010. http://lacito.vjf.cnrs.fr/pangloss/ (accessed 6 July 2014).

Drettas, Georges. 2013. Bulgarian-Macedonian Corpus. In Evangelia Adamou, Walter Breu, Georges Drettas, and Lenka Scholze (eds.), Electronic database of endangered Slavic varieties in non-Slavic speaking European countries, ANR-DFG EuroSlav2010. http://lacito.vjf.cnrs.fr/pangloss/ (accessed 6 July 2014).

Faska, Helmut. 1998. Koncepcija zbliženja a zjednoćenja spisownych formow serbšćiny. In Helmut Faska (ed.), Serbšćina, 167–177. Opole: Universytet Opolski.

Gardner-Chloros, Penelope. 2009. *Code-switching*. Cambridge: Cambridge University Press.

Haspelmath, Martin and Uri Tadmor (eds.). 2009. *Loanwords in the World's Languages: A Comparative Handbook.* Berlin: Mouton de Gruyter.

Heine, Bernd and Tania Kuteva. 2005. *Language contact and grammatical change.* Cambridge: Cambridge University Press.
Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3): 651–674.
Krauss, Michael. 2006. Classification and terminology for degrees of languages endangerment. In Matthias Brenzinger (ed.), *Language Diversity Endangered*, 1–8. Berlin: Mouton de Gruyter.
Matras, Yaron. 2009. *Language contact.* Cambridge: Cambridge University Press.
Matras, Yaron and Jeanette Sakel (eds.). 2007. *Grammatical borrowing in cross-linguistic survey.* Berlin/New York: Mouton de Gruyter.
Muysken, Pieter. 2000. *Bilingual Speech: A Typology of Code-mixing.* Cambridge University Press: Cambridge.
Myers-Scotton, Carol. 1993. Duelling Languages: Grammatical Structure in Code-switching. Oxford: Clarendon press.
Myers-Scotton, Carol. 2002. *Contact linguistics, bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.
Neweklowsky, Gerhard. 1978. *Die kroatischen Dialekte des Burgenlandes und der angrenzenden Gebiete.* Wien: Verlag der Österreichischen Akademie der Wissenschaften.
Poplack, Shana. 1985. Contrasting patterns of code-switching in two communities. In Henry J. Warkentyne (ed.), *Methods V: papers from the V International Conference on Methods in Dialectology*, 363–385. Victoria, B.C.: University of Victoria.
Poplack, Shana. 1993. Variation theory and language contact. In Dennis Preston (ed.), *American dialect research: An anthology celebrating the 100th anniversary of the American Dialect Society*, 251–286. Amsterdam: Benjamins.
Poplack, Shana. 2001. Code-switching. In Neil Smelser and Paul Baltes (eds.), *International encyclopedia of the social and behavioral sciences*, 2062–2065. Elsevier Science Ltd.
Poplack, Shana and Dion, Nathalie. 2012. Myths and facts about loanword development. *Language Variation and Change* 24 (3): 279–315.
R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Scholze, Lenka. 2008. *Das grammatische System der obersorbischen Umgangssprache im Sprachkontakt. Mit Grammatiktafeln im Anhang*. Bautzen: Domowina.
Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothor. 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* 8 (25). URL http://www.biomedcentral.com/1471-2105/8/25.
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9 (307). http://www.biomedcentral.com/1471-2105/9/307
Szucsich, Luka. 2000. Das Burgenlandkroatische: Sprachwandel, Sprachverfall, Sprachverschiebung und Sprachassimilation. In Lew N. Zybatov (ed.), *Sprachwandel in der* Slavia, 853–875. Frankfurt a.M. et al.: Peter Lang.
Tadmor, Uri. 2009. Loanwords in the world's languages: findings and results. In Martin Haspelmath and Uri Tadmor (eds.). *Loanwords in the World's Languages: A Comparative Handbook. Loanwords in the World's Languages: A Comparative* Handbook, *55*–75. Berlin: Mouton de Gruyter.
Tagliamonte, Sali and Harald Baayen. 2012. Models, Forests, and Trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2): 135–178.

Thomason, Sarah and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics.* Berkeley/Los Angeles: University of California Press.

Travis, Catherine and Rena Torres Cacoullos. 2013. Making voices count: Corpus compilation in bilingual communities. *Australian Journal of Linguistics* 33(2): 170–194.

Travis, Catherine and Rena Torres Cacoullos. (in press). Gauging convergence on the ground: Codeswitching in the community. *International Journal of Bilingualism*.

Treffers-Daller, Jeanine. 1994. *Mixing two languages: French-Dutch contact in a comparative perspective.* Berlin: Mouton de Gruyter.

van Hout, Roeland and Pieter Muysken. 1994. Modelling lexical borrowability. *Language Variation and Change* 6(1): 39-62.

Winford, Donald. 2003. *An introduction to contact linguistics.* Oxford: Blackwell.