Academic Journal of Modern Philology

ISSN 2299–7164
Vol. 18 (2022)
Special Issue
s. 7–22

Robert Borges
Institute of Slavic Studies, Polish Academy of Sciences
ORCID: 0000-0002-7647-4048

Sourcing Data from Wikipedia for the Study of Language Contact: the csbwiki[1]

Abstract

Contact-induced language change is pervasive in contexts involving historically minoritized languages, where social contexts are not particularly conducive to equitable intergroup relations. Empirically driven studies involving these language contexts allow us to more thoroughly understand the social and cognitive processes that lead to language change. Paradoxically, empirical data on minoritized languages is relatively scarce and expensive to generate. But in the digital age we have the ability to look beyond the traditional data types used in language studies, like spoken data gathered under fieldwork conditions, literature, etc. In this paper, I will explore the potential utility of user-created wiki data in investigating Polish influence on the Kashubian language.

Keywords: contact-induced language change, wiki data, Kashubian, corpus linguistics, vowel alternation

1. Introduction

In the field of contact linguistics, it is pragmatic to utilize diverse types of empirical data to observe and understand the processes and results of language contact. Many studies in contact linguistics are based on relatively small samples of spoken data, gathered in interviews, "traditional" linguistic elicitation, or recorded during (participant) observation. Reliance on this data type is justified by the fact that it is perceived to more closely represent natural spoken language, and study results are routinely qualitative. Contrariwise studies that rely on large(r) corpora of written data tend to result in more quantitatively-

1 This paper is a result of research conducted under the auspices of the project *New Speakers of Minority Languages: Proficiency, Variation, and Change* 2021–2023, hosted at the Institute of Slavic Studies: Polish Academy of Sciences, funded by the Polish National Science Centre (NCN) under the "POLS" instrument financed by Norway Grants. Contract #2020/37/K/HS2/02779.

doi: 10.34616/ajmp.2022.18.1

oriented illustrations of the distributions or spread of features of interest. I would argue that integrating these types of data lead to a more holistic picture of the effects of language contact.

In the cases of historically minoritized languages and languages with a small number of speakers, those that lack extensive infrastructure like large corpora and computational / NLP tools, researchers must be creative in generating their own sources of diverse data. In this paper, I will present the methods and explore the potential utility of using data from Wikipedia as a supplementary data source in the study of language contact; specifically, I examine the distribution of feature variants attested in spoken data, characteristic of Polish influence, in the Kashubian-language Wikipedia. In the next section, I introduce the wider research context and motivate the inclusion of a written corpus, specifically Wikipedia data, in the analyses. In Section 3, I introduce Wikis and Wikipedia along with a brief overview of the Kashubian-language Wikipedia. Then in Section 4, I explore the case of a single case of feature variants on order to estimate the utility of using Wiki data as a supplement in this case. Finally, I conclude with a discussion on the utility of the data and methods employed both generally and for this particular case – the study of Kashubian in contact with Polish.

2. Research Context: the New Speakers of Minority Languages Project

Within the project *New Speakers of Minority Languages: Proficiency, Variation and Change*, we seek to explore the relationship between language acquisition and contact-induced language changes. Particularly, we are interested in empirical evidence that not only supports a link between acquisition and change but illustrates the processes synchronically. The idea that the language acquisition process is at the locus of diachronic change has been proposed in a number of publications (*e.g.* Winford 2003; Hróarsdóttir 2004; Lightfoot 2007; Matras 2009; Meisel 2011; Diessel 2012), but empirical evidence that illustrate these processes, however, has been elusive in part due to faulty underlying assumptions, that the acquisition process has an end point and that the target of acquisition is a static variety equated either with "native"-like patterns of language behavior or with a prescriptive notion of some standard language variety. As a theoretical point of departure, we recognize several relevant premises that moderate first and second language acquisition.

- Language acquisition is a lifelong process.
  Although there is an undisputed intensely dynamic period of language learning,[2] thanks to the cognitive and neural plasticity of the human brain (Willis, Schaie, and Martin 2009), communicative knowledge is continuously reorganized and adapted throughout an individual's life, a process referred to as linguistic entrenchment (Hopper 1987; Schmid 2016a). There is a growing and convincing body of evidence from a variety of linguistic sub-disciplines supporting the existence of entrenchment (see: Schmid 2016b for a detailed overview). Already noted in Langacker's early proposal (1987) on the subject, entrenchment occurs via frequent and regular repetition and rehearsal of linguistic input. For language acquisition,

2 This occurs in early childhood for L1 acquisition. In L2+ acquisition it occurs during the initial period of exposure; immersion in "naturalistic" acquisition, course work, etc.

entrenchment leads to the emergence and (re)organization of variable schemas, allowing for a high degree of automaticity in language processing. Such automaticity is characteristic of proficient users of a language, resulting from efficient memory consolidation and chunking of lexical and morphosyntactic units.

- Language acquisition is intergenerational *and* lateral.

  Due to the assumption that acquisition can be completed, studies often focus on input children / learners receive from primary care givers / teachers in that highly dynamic period of early language acquisition. But given that entrenchment is a lifelong process, all communicative input is received with *potential* to affect change in an individual's repertoire of idiolectal patterns. Thus we could say that at the level of the individual, idiolects are a directly shaped by the linguistic input they experience.[3]
- All users of a language contribute to the *feature pool*[4] on which other speakers model their behavior.

  In this framework, the concept of "language" refers to the set of features that most commonly overlap in the idiolects of its speakers. Features that do not fall within these norms are considered peripheral, and belong to "varieties" or "dialects" of the language when they overlap in regional or social cross sections of the speaker population. All features, whether peripheral, varietal, or normative, are potentially exposed to other speakers and may play a role in the entrenchment process, eventually becoming part of those speakers' repertoires.

These premises are most easily understood with the example of the spread of technology-related neologisms to segments of the population who were already well into their adult years by the time the relevant technology was invented. The fact that, today, elderly people know and use terms like *download*, *to google sth*, *blog* / *vlog*, or *selfie* illustrates the lifelong continuity of language acquisition. Novel terms that were introduced by relatively few speakers in highly restricted context spread with the increasing prevalence of technology, which widened the contexts in which the terms were used, in turn generating a higher frequency of repetitions, and these novel terms spread on the analogy of an epidemic (Jiang *et al.* 2021). The example clearly illustrates the progression of incipient linguistic features progress to minor patterns and then normative features. Heine and Kuteva (2005) describe a similar progression as a diachronic process, but here we also see it in operation within the minds of individuals.

Synchronic language use has immediate effects on mental representation, that is, the language used by an individual and the language experienced by that individual is reinforced in, or incorporated into, the linguistic knowledge that that individual has at his/her disposal when communicating (Backus 2021). This means that an individual's lexicon, as in the examples above, and grammatical structures are *emergent*, fluid across the lifespan, reflecting experience and partially improvised within contexts of social interaction (Hopper 1987, Bybee 2010). The sheer frequency if form-meaning pairs that we experience

3 Of course the probability that input features are reflected in an individual's idiolect are mitigated by a host of factors including frequency and salience of features in the input, language attitudes and ideologies etc.

4 The terminology is obviously reminiscent of Mufwene's (2001) work. While there are numerous merits to Mufwene's work, I want to distance this account from any analogy between a feature pool in language and a gene pool in evolutionary biology. Mufwene himself presents a caveat in this regard (2001: 30), but I feel it is worth reiterating in consideration of the responses to his work.

has a role to play in entrenchment of our linguistic knowledge. The relatively high string frequency of *old man* in an ADJ-N schema compared to *man* as a verb would cause many users of English to have to reread (1) before realizing the appropriate and fully grammatical meaning of the example (second gloss line). Similarly, due to relatively high occurrence of AGENT-VERB-PATIENT schemas in Polish, without context, (2) will almost always be interpreted in the meaning given in (2a), despite case syncretism, free word order and grammaticality of (2b). Further, the situation of experience within social contexts preempt activation of appropriate linguistic knowledge to some extent, *e.g.* (3) will almost always get the reading in (3b) when spoken by a cashier in a grocery store.

(1) *The old man the boat.*
*DET ADJ N DET N
DET N V DET N

(2) a. *Auto uderzył dziecko.*
car. NOM hit.3SG child.ACC
'The car hit the child.'

b. *Auto uderzył dziecko.*
car.ACC hit.3SG child.NOM
'The child hit the car."

(3) a. *Do you wanna box for those groceries?*
Q 2SG AUX V PREP DET N
Shall we punch each other until only one of us is left uninjured enough to leave with the groceries?'

b. *Do you wann-a box for those groceries?*
Q 2SG V-DET N PREP DET N
'Would you like a cube-shaped container in which you can carry your groceries away?'

In multilingual individuals, there is no principled division of linguistic knowledge. Multilinguals who need to interact in a monolingual context need to cognitively suppress co-activation of linguistic knowledge from outside the single language of interaction (*cf.* Bobb, Wodniecka, and Kroll 2013, Green and Abutalebi 2013). Lexicon and schemas from multiple languages are active in parallel in multilingual individuals (Kroll and Bialystok 2013) and priming across languages has been demonstrated both experimentally (Kootstra, Hell, and Dijkstra 2012; Hell and Tanner 2012; Kroll and Bialystok 2013) and in spontaneous corpus data (Fernández, Souza, and Carando 2016; Gries and Kootstra 2016). The link between individual entrenchment and speech-community conventionalization comes in the form of a feedback loop; entrenched linguistic structures are most easily activated in the individual, which are then in turn utilized by the individual and modeled to other speakers, whose parallel structures are in turn activated. In other words, entrenchment contributes to conventionalization; conventionalization contributes to entrenchment (Bybee 2010, Hopper 2013, Schmid 2020, Backus 2021).

In settings where a community's speech behavior is stable and there is neutral or high prestige associated with the language variety, linguistic change is usually slow to the point of being unnoticeable among living people. Children —even multilingual children— tend to acquire target language structures

with a high degree of accuracy. The observation that noticeable language changes, especially those involving language structure, occur more readily when a substantial part of the speech community consists of second-language speakers has been made by multiple scholars (*e.g.* Weinreich, Labov, Herzog 1968; Thomason, Kaufmann 1988; Paulston 1994; Barðal 2009; Cognola, Bidese 2016). But in light of the above paragraphs, I would argue that this is the result of conventionalization of emergent multilingual behavior rather than "incomplete" or "imperfect" acquisition, as is often claimed.

Contexts involving endangered and historically minoritized languages, where revitalization activities and / or language activism takes place, create ideal set of circumstances to study the relationship between entrenchment in the individual and conventionalization of features originating from the acquisition and cognitive processes of bi/mutlilingual individuals. Particularly we consider the role of New Speakers to be of central importance. A New Speaker, following O'Rourke, Pujolar, and Ramallo (2015), refers to an individual who has learned a language with little or no exposure in the home via educational programs outside the home after a community-level shift. They often made a conscious choice as teenagers or young adults to engage with the language and live life through it and for that reason New Speakers of minority languages tend to be situated in relatively prominent positions within minority language communities. They tend to be engaged and active around issues of language rights and minority language education, many serving as language teachers themselves. Thus they have an increased potential, compared to learners of majority languages, to model language behavior and influence norms at a community level.

One of the case studies addressed in the current project involves New Speakers of Kashubian.[5] The working hypotheses of the project are: (a) that linguistic entrenchment can be measured together with lexical and morphosyntactic variation in individual speakers, thereby allowing for an understanding of degrees of acquisition without relying on native-speaker patterns or prescriptive notions of language behavior; (b) that differential patterns of linguistic input between New Speakers and native speakers result in different patterns in lexical/structural entrenchment, which in turn result in differential usage patterns; and (c) frequently used lexical/structural patterns, whether produced by native or New Speakers, provide model input for continued adaptation and reorganization of communicative knowledge across the speech community.

We utilize a relatively elaborate set of methodological procedures to address (a) and (b), including a sociolinguistic questionnaire (individuals' background, language attitudes, and proficiency self-assessment), a video narration task (elicits spontaneous spoken data under semi-controlled circumstances), a Rapid Automatized Picture naming task (measures automaticity in lecical production checks productive proficiency, as in Borges 2019), and a receptive proficiency task. (Thanks to COVID-19, these are administered remotely; see: Borges 2022a.) The methodology has proven adequate in addressing (a) and (b), but it lacks any means of pinpointing whether features observed in New Speaker varieties appear "in the wild" or possibility to observe their spread to other segments of the population. In order to address this, in the case of Kashubian, we turn to the Kashubian-language Wikipedia page for a relatively large corpus of readily available data both generated by and, in theory, used by the target speech community.

5 Kashubian (ISO-code: csb) is a West-Slavic dialect continuum spoken in a discontiguous area with in the Polish Pomeranian Voivodeship. Kashubian speakers are considered and autochtonous ethnolinguistic minority group and the language was officially recognized as a regional minority language in Poland in 2005.

3. Wikis and Wikipedia

The term *wiki* refers to "a hypertext publication collaboratively edited and managed by its own audience directly."[6] Wikis can be public or private, hosted on the public internet or within a private network. Wikipedia, which is perhaps the most popular and well-known wiki is a freely available online multilingual encyclopedia. At the time of this writing, Wikipedia hosts encyclopedia wikis in 316 languages,[7] all of which consist of articles contributed by the wiki's users themselves. Article pages are typically created from the bottom up, that is, users contribute content they deem to be relevant; there is no centralized content management and pages are largely *not* translated from some other language.

While highly convenient for the trivia buffs and information connoisseurs, Wikipedia pages also have a role to play in science. Since its inception in 2001, the site has published 55.89 million pages if all language wikis are considered,[8] making it an enormous source of decentralized language data. In addition to free access to individual pages on Wikipedia's website, the foundation regularly "dumps"[9] content of its wikis into structured files with various levels of detail – from simple meta information on the currently published revision, containing a list of pages by title, date of latest revision, and contributor, to complete history of the wiki.[10] Dumps are downloadable in compressed XML format, a which lends itself well to scripted processing and analyses; indeed the majority of the analyses presented here were conducted with Python. Since the intent of this paper is mainly to address methodological issues, data, scripts, and output of scripts are included in an open-access supplement (Borges 2022b).[11]

There is some precedence for using Wikipedia as a source in a variety of linguistics research (and related disciplines). Even a cursory exploration of the literature reveals papers addressing questions within variationist / corpus linguistics (Hiltunen 2014; Margaretha, Lüngen 2014; Hiltunen, Tyrkkö 2019), training computational / NLP models (Yano, Kang 2008; Nelken, Yamangil 2011), and utilizing wiki data in ESL and "applied" purposes (King 2015; Shi 2015). Few address any issue related to minority language Wikipedia pages (Tomás *et al*. 2008; Arkhangelskiy, Medvedeva 2016), but none that I could find directed attention to variation or (contact-induced) change in a minority or low-resource language.

3.1. The Kashubian Wikipedia: csb.wikipedia.org

The oldest articles in the Kashubian Wikipedia date from April 2004. The dump referenced in this paper (2022-05-01) consists of 8,705 pages including 168,225 revisions.[12] The most recent revisions in the dump contain 1,529,977 words, or 140,169,706 words if text from all revisions is considered. Pages and revisions have been contributed to by 1,576 registered users; additionally contributions have been made

6 [At:] https://en.wikipedia.org [date of access: 22 June 2022].

7 [At:] https://meta.wikimedia.org/wiki/List_of_Wikipedias [date of access: 22 June 2022].

8 [At:] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia [date of access: 22 June 2022].

9 "Dump" here refers to the creation a single file consisting of a structured record of all the pages at a the time of the dump.

10 Visit: https://dumps.wikimedia.org/backup-index-bydb.html to explore wikis that have dumps available to download.

11 In order to preserve readability for readers who are not enticed by such a level of detail, references to individual supplement files are given in footnotes.

12 Note that this figure differs from the number of articles reported on the csbwiki's landing page at the time of the dump, which was 5,428.

by non-registered users referenced by 6,409 IP addresses.[13] To summarize, the Kashubian Wikipedia is a relatively large resource in which we can examine the whole history of contributions, edits and "discussion sections" by page and user.[14]

As we specifically set out to explore whether learner-associated variants found in our New Speaker data are modeled to the wider speech community population, the prominent warning on the Kashubian Wikipedia landing page (Figure 1)...

PROSBA: Szkólnégò, chtëren zadôwô pisanié artiklów dlô kaszëbsczi Wikipedie, sertno prosymë ò jednoczasné sprôwdzanié lëcznëch felów w tekstach ùczniów.
'REQUEST: We kindly ask the teacher who uses writing articles on Kashubian Wikipedia as exercise to check simultaneously for multiple errors in pupils' texts'

... suggests that the Kashubian Wikipedia (*csbwiki* henceforth) data provides a promising avenue for this type of research.

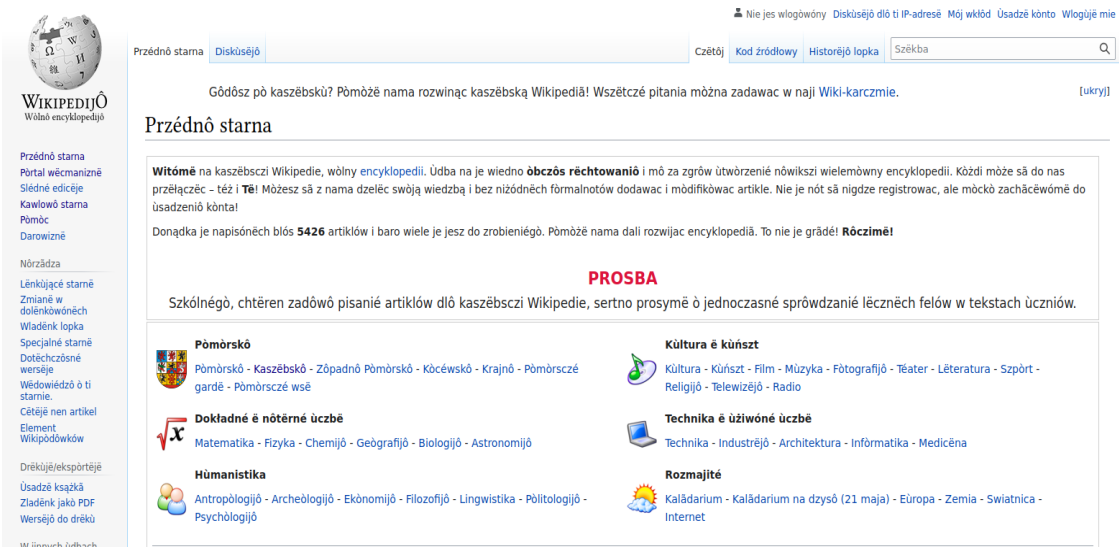

Figure 1. The landing page of Kashubian Wikipedia. Source: https://csb.wikipedia.com [date of access: 22 June 2022].

4. One instance of ô – a alternation

In this section, I will present a small case study which originates with Bandur's (2022) observation, made when working with spoken data gathered in the project, that Kashubian vowel distinctions which are not present in Polish, are often neutralized in New Speakers' Kashubian. A comparison of the vowel inventories

13 Theoretically, this could be a single individual or 6,409 individuals. As far as I can see there is no way to determine this.

14 Descriptive statistics presented in this paragraph were generated using the script `count-pages.py` and `count-words.py` in the supplement.

of Kashubian and Polish (Figure 2) reveals that Kashubian has an additional height distinction present across the horizontal space. Figure 3 illustrates the phenomenon with several instances of Kashubian vowels rendered with their closest Polish equivalents. This phenomenon has consequence not only for the phonology of Kashubian, but also for morphology, where morphemic forms are distinguished by these "additional" non-Polish vowels.

	Front	Central	Back
High	i	y	u
Mid	e ę		o ą
Low		a	

Polish vowels

	Front	Central	Back
High	i	ɨ	u
Mid-closed	e		o õ
Mid-open	ɛ	ə | ɜ	ɔ
Low		a ã	

Kashubian vowels

Figure 2: Polish vowels & Kashubian vowels. Source: Maturzysty 1992:11; Makùrôt 2016:18.

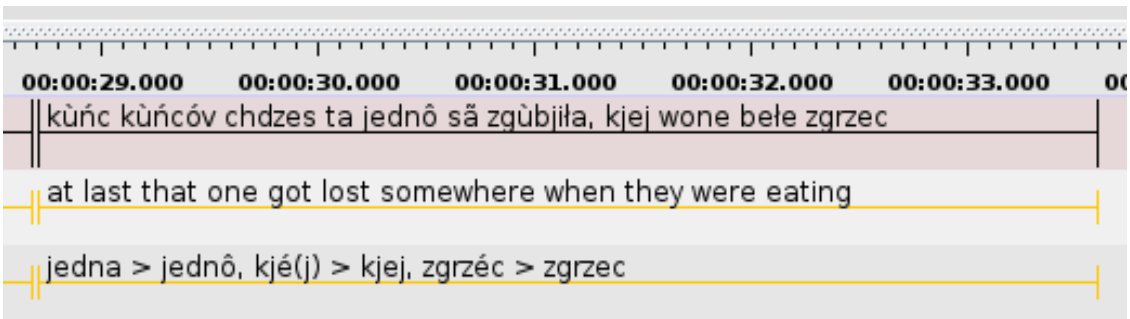

Figure 3: Printscreen of transcription and analyses of spoken Kashubian in Elan (Sloetjes and Wittenburg 2008) by a New Speaker, which exemplifies multiple instances of the type of vowel neutralization in question.

As a first attempt to evaluate the utility of the csbwiki data in relation to our research questions, I began manually searching the data set (ctrl-f with the xml document open in a text editor) for the individual instances that were already attested in our spoken data (such as the examples shown in Figure 3). The process was very tedious and did not result in many instances of interest. As a second attempt, I decided to take a broader, brute-force approach to sifting the data, but with a narrower focus in terms of the category of target variants – that is, precisely one set of variants. I first searched the data for all instances of word tokens containing <ô>, which represents the mid-central vowel /ɵ/ in orthography, and from this list of ô-words, I generated a second list of hypothetical a-words by replacing all instances of <ô> with <a>, which represents the low central vowel /a/ in orthography.[15] I then iterated over the data again and

15 This was done with the script `mk_oe-a_variants.py`, which dumps a list of all ô-words to the file `oe-words.txt`. A list of <ô> and <a> pairs is dumped to the file `oe-a_variants.json`. The strategy doesn't account well for words that have two or more instances of <ô>, *e.g.znônô*, 'known', which has three hypothetical variants: *znana*, *znanô* and

counted all instances of each word on the list of ô- and potential a-variants.[16] This resulted in 2,038 pairs of words where both the ô-variant and a-variant are attested in the data, as in (4), amounting to a total of total of 610,828 ô-variants and 2,520,170 a-variants from these pairs in all revisions.[17]

(4) a. *encyklopedijô* (N=1,364)
b. *encyklopedija* (N=191)

In this list of pairs there are lots of correct or potentially correct forms, place names for instance where both variants are accepted. Additionally, menu items, including category names, etc., are not particularly useful in determining whether learner behavior is modeled to the wide speech community. For this reason, I decided to further narrow the focus to a single pair of variants to explore how to work with the data; *mô* and *ma* were selected because there was a large, but manageable, number of examples (5), neither variant is a common place name or a likely menu item, and there are multiple functional meanings to each form in Kashubian, which partially overlap functionally with that of *ma* in Polish. Thus the particular case sits at the intersection of phonoloical and morphological systems in contact.

(5) a. *mô* (N=18,965)
b. *ma* (N=6,413)

Specifically, I examined these forms in the meaning of 3SG.have. The expected form in Kashubian in non-negated contexts is *mô* 's/he / it has', as in (6); no other senses of the form are apparent in the data. Following negation, though, 3SG.have takes the form *ma*, in the negative existential sense 'there isn't/aren't', but there are also other senses of the form as well.[18] This means that the number of examples presented in (5b) needed further filtering in order to understand the distributions of 3SG.have forms.

Collocations of *ni ma* in the negative existential sense, as in (7), were filtered from the pool of potential examples featuring the form *ma*; 1,169 such instances were identified.[19] A number of pages and discussion sections were started or otherwise appeared partially or entirely in Polish – *ma* examples were

znôna. There are 885 such words (of 11,693, ca. 7.6%) on the `oe-words.txt` list and 120 (of 2,038, ca. 5.9%) remain after populating a list of non-zero pairs `oe-a_counts.json`.

16 Counting of variants was done with the script `count_oe-a_variants.py`, which first dumps a list of all tokens and the N times they are attested in the data `raw_oe-a_counts.json`, and a second list where ô- and a-word pairs that both have a non-zero value are listed together, `oe-a_counts.json`.

17 It should be noted that not all lexemes in the a-variant list are necessarily versions of the ô-variant. Some a-variant forms are expected Kashubian forms, unrelated to the ô-variant.

18 Actual examples from the data were dumped into a file using the script `re-dump-variants.py`. In this file it is necessary to establish a regular expression pattern and a corresponding file name for the output file. Both *mô* and *ma* variant instances were dumped to the file `moe-ma_variants-dump.json`, *ma* instances to a file `ma_variants-dump.json`, etc. All patterns and dump file names used for this manuscript remain (commented out) in the script file.

19 Filtering data was done with the script `re-filter-dump.py`, which in this case first took the file `ma_variants-dump.json` as an input file and removed fragments matching an established regular expression pattern. Results are written to a file, which becomes the input to subsequent runs of the script with additional regular expression patterns. Like the previous script, patterns and output files used in this manuscript remain in the script file.

subject to another round of scripted filtering where an additional 1,213 instances of *ma* were discounted by searching Polish collocation *nie ma*. The remaining subset of *ma* instances was then manually filtered to remove additional irrelevant or otherwise unwanted examples; there were, for instance, still many wholly Polish fragments, instances of *ma* that are expected Kashubian in a sense other than 3SG.have, a small handful of typos, as well as quite a lot of other "junk" including links to pages in other languages, place names, URLs, and linked file names. During manual filtering, I also noticed instances of unexpected preverbal negation, with the forms *nié ma* (N=398) and *nië ma* (N=18), which were likewise filtered from the analysis by script.[20]

(6) a. *Elżbiéta Bùgajnô* [...] *mô* *stãpiéń doktora.*
Elżbieta Bugajna ... 3SG.have degree doctor
'Elizabeth Bugajna has a doctorate.'

b. *Zarno ówsa mô wãglowaodanë, strzód chtërnëch nôwiãcy je skrobi.*
grain oat 3SG.have carbohydrates among which most be starch
'Oat grains have carbohydrates, mostly starch.'

(7) a. *Òna żëje w Bôłce, ale terô ni ma ji wiele, a rëbôcë mają jiwer.*
she live in Baltic but now NEG 3SG.have her many and fishermen 3PL. have worry
'It lives in the Baltic, but now there are few and fishermen are worried.'

b. *W Pòlsce òd 1958 rokù tegò òbrzészkù ni ma ...*
in poland since 1958 year that obligation NEG 3SG.have
'There is no such obligation in Poland since 1958 …'

After all scripted and manual filtering, I was left with 553 instances of *ma* in 547 revisions on 22 pages if the entire revision history is considered.[21] In the current revision, 3SG.have *ma* appears on just 17 pages (18 instances), as in (8). Four instances of *ma* as 3SG.have were corrected in the revision history, three to *mô*. and one to *mdze*, 3SG.be. The other two instances were not corrected, per se, but all text was removed completely from the page on which these instances occurred, leaving only a redirect to another page (where there are no instances of *ma* as 3SG.have).

(8) a. *Rëmiô ma wiéchrzëzna 32,86 km².*
Rumia 3SG.have area 32,86 km²
'Rumia is 32,86 km².'

b. *Bôjka ma charakter uniwersalny...*
tale 3SG.have character universal
'The tale is of universal character.'

20 These forms of negation are unexpected because *nié* should have scope over the whole clause, and *nië* shouldn't exist at all by prescriptive standards. While potentially interesting examples, probably also having to do with multilingual processing and/or Polish interference, these were filtered because they are negated.

21 Manually filtered data is contained in the supplement file called `ma-dump_he-has.json`. Filtered results are counted with `count-filtered.py`.

c. *Tu je Kilimanjaro, nôwëszô wëszawa Africzi, co ma 5895 m.*
here 3SG.be Kilimanjaro the highest acclivity Africa which 3SG.have 5895 metres
'Here is Kilimanjaro, the highest mountain in Africa, which has 5895 meters'

d. *... na karkù ma czôrnô plamë.*
... on neck 3SG.have black spot
'... there is a black spot on the neck.'

e. *Miono Jóna ma dzysô tramwaj gduńszczi nr. 1014 ...*
name Jón 3SG.have today trolley Gdańsk.ADJ number 1014 ...
'Today the Gdańsk trolly number 1014 has the name Jón …'

Of the number of *mô* instances attested in all revisions listed in (5a), exactly 600 instances appear in the current revision.[22] Non-negated forms of *mô* account for 565 of these instances; assuming these are all well-formed instances in the meaning 3SG.have, this means that *ma* has a very low percentage of representation in all instances of forms with the positive 3SG.have sense — 3.19%. This result confirms that features associated with learners are in fact modeled to the wider population, albeit at a low ratio (ca. 1 in 31 instances) compared to the non-learner variant.

Interestingly, in looking carefully at *ma* and *mô* variation, we also find unexpected forms in contexts of pre-verbal negation; of the 30 instances of *ni mô*, which is the expected in cases of the sense 'X doesn't have' (as opposed to the existential sense 'there isn't / aren't'), three are unexpected in that their reading *is* existential. There are five additional instances of the collocation *nié mô*, which are unexpected, both because of the form of the negation (see: footnote 20) and the sense in some instances.[23]

The second question posed here, about who exactly is using which variants, is more difficult to answer. The instances of *ma* in the meaning 3SG.have were contributed by 15 users (including six users by IP address, see: footnote 13), so it is clear that the majority of these instances are not contributed by a single user. But this is approaching the extent of what we know for this particular set of variants. Wikipedia stores "Babel" data on its contributors, that is proficiency self-assessments on a seven-point Likert scale, where users are rated from completely no knowledge of a language to "Native"-like competence,[24] but users themselves are responsible for providing this data and most have not done so. In the case of Kashubian Wikipedia, only 64 users provided Babel data, listing 102 languages total.[25] The figure is reduced to 85 languages if those languages with a score of *0*,[26] are discounted.

22 These figures are generated with `count-moe.py`.

23 See: the file `ni-moe.txt` for these examples.

24 [At:] https://en.wikipedia.org/wiki/Wikipedia:Babel [date of access: 22 June 2022).

25 Babel data was sorted into "by language" and "by user" forms by the script `jsonify-babel.py`.

26 *0* indicates "you cannot understand the language at all". The Babel documentation recommends "Do not use [0] for every language that you don't know, but only when there is some reason why you might be expected to know it. For example, one may be of Italian descent, but does not speak the Italian language, or if one is Canadian but does not speak French. Similarly, one may usefully edit a project without speaking the language, such as adding links or images to Japanese Wikipedia without speaking Japanese." https://en.wikipedia.org/wiki/Wikipedia:Babel [date of access: 22.06.2022].

These data are of little use in constructing profiles of Kashubian Wikipedia contributors. Only 36 users list Kashubian in their profile, but 33 of them rated their knowledge as *0*, two as *1* "basic knowledge",[27] and one as *N*, "native speaker".[28] Only one user that contributed a *ma* as 3sg.habe example to the current revision has registered Babel information, with Kashubian at level 0, however, a look back into that page's revision history shows that this user did not originally contribute the instance of *ma*. In any case, it is clear that the Babel assessments are wholly insufficient to make any attempt at addressing the distribution of variant in relation to the proficiency of contributors. It may still be possible to evaluate users' by repeating procedures like outlined here with a number of other hypothesized learner – non-learner variants; where users consistently favor some variants over others, profiles could potentially be constructed, but these must originate in the language data itself. Other metrics may also factor into this type of measurement, such as the sheer amount of text contributed by a user, lexical density (balanced type-token ratios) of a user's contributions, etc., but this is beyond the scope of what can be done here.

5. Discussion and Conclusion

In this paper, I examined a single set of variants to explore whether the learner-associated variant, observed in New Speaker data, was modeled to the general population in the Kashubian Wikipedia data set. Indeed *ma* as 3sg.have *is* present in the data, however, with a very low relative frequency at just over 3% of all instances. This study provides no indication of when or how exposure to such variants become significant. Is there a tipping point in the ratio of variants present that might be indicative of how acceptable a particular variant is? Or, rather, do we need to examine how many members of the speech community actually "consume" variants; how often is the Kashubian Wikipedea read and by how many readers in this case? Assuming the accuracy of the theoretical framing outlined in Section 2, both the modeling and consumption of variants would be relevant to understanding the processes at hand.

This particular case study did not provide us with any information about the actual spread of apparent learner-modeled behavior to the wider speech community, since none of the relevant examples were contributed by individuals with a Babel profile on the site. Nevertheless, I would like to advovate the utility of the data and the methods presented here. Perhaps the most obvious advantage of utilizing Wikipedia data on low-resource, less-widely spkoken languages is that one can begin to work with a fairly large data set in a literal matter of minutes. In comparison to the expense of curating data by more traditional means (travel, field work, transcription, etc, not to mention recent health-related ethical concerns and environmental impact), ready-made, structured, accessible data sets are both convenient and welcome among many linguists. Data is largely written from the bottom up, which should greatly reduce translation effects that are usually apparent in translated parallel corpora. Another consideration is that data contributed to Wikipedia is not done so specifically as a contribution to linguistics research, which largely eliminates the well-known Observer's Paradox, meaning that studying Wiki data allows us to observe more naturalistic language behavior. However, the anonymity of data contributions makes

27 *1* indicates "basic ability – enough to understand written material or simple questions in this language." [At:] https://en.wikipedia.org/wiki/Wikipedia:Babel.

28 *I.e.*, "native-born speakers who use a language every day and have a thorough grasp of it, including colloquialisms and idioms." [At:] https://en.wikipedia.org/wiki/Wikipedia:Babel [date of access: 22.06.2022].

understanding the distributions of variants more difficult. The little background information available on contributors is both suspect in most cases and impossible to verrify; any information about contributors must be derived from the language data itself.

Another advantage to utilizing Wikipedia data is that it represents linguistic output, in that text is produced by individual writers, as well as linguistic input, that is, Wikipedia is also consumed by members of the Kashubian speech community. While also true of other media types, this is not necessarily the case with other types of data studied in Linguistics; those features produced by informants under duress of linguistic elicitation are not necessarily the ones that are commonly used or understood in everyday language. Therefore, data from Wikipedia or other forms of media has an important role to play in triangulating elicited or experimental observations "in the wild". This aspect has been foregrounded in this manuscript, as it is particularly important within the project's framework to have a relatively balanced picture of both production and reception of linguistic variation. A missing piece of the puzzle in this story, though, is that we actually have no idea how Kashubian Wikipedia is consumed. There appears to be no information available in a given Wikipedia dump about how often or by whom a particular page was accessed or interacted with.

Revision history is also an asset to working with Wikipedia data. When a particular feature is corrected in the revision history, it tells us that that particular feature is salient enough to be considered outside of the norms of the language and / or socially marked (*e.g.* as a dialect feature, or "not good Language X"). While in the case study presented here, only four variants of interest were corrected in the revision history, the structure of Wikipedia data allows to study the rise and conventionalization of variants by looking at the ratio of frequencies at which they are contributed and corrected over time.

I want to emphasize once again the importance of triangulating research findings across data types. In this case, observations made in semi-experimental spoken data were corroborated in the written wiki data. Further, I hope this paper has made the case, illustrating some of the benefits of studying variation in user-contributed, user-edited data such as the type found on Wikipedia. With a single case study of a single set of variants, one part of a two-part question has been successfully addressed, suggesting that further exploration of variation will provide more insight into the distribution, attitudes toward, and possible spread of such variants.

References

Arkhangelskiy, Timofey, Maria Medvedeva (2016) "Developing Morphologically Annotated Corpora for Minority Languages of Russia." [In:] Sandra Kübler, Markus Dickinson (eds.) *Proceedings of Corpus Linguistics Fest 2016, Bloomington, Indiana, USA, June 6–10, 2016* 1607 CEUR Workshop Proceedings. CEUR-WS.org. [At:] http://ceur-ws.org/Vol-1607/arckhangelskiy.pdf [date of access: 22.06.2022]; 1–6.

Backus, Albert (2021) "Usage-Based Approaches." [In:] Evangelia Adamou, Yaron Matras (eds.) *The Routledge Handbook of Language Contact.* London: Routledge; 110–126.

Bandur, Maciej (2022) "Vowel Alternations in New Speakers' Kashubian in Diachronic Perspective." A talk given at the conference *Regional Languages in Education: from Literature towards Literacy* [on:] 2022-06-16 in Rēzekne, Latvia. [At:] https://ispan.waw.pl/nespomila/static/assets/Bandur_vowel-alternations_Rezekne-20220617.pdf [date of access: 22 June 2022].

Barðal, Jóhanna (2009) "The Development of Case in Germanic." [In:] Jóhanna Barðal, S.L. Chelliah (eds.) *The Role of Semantic, Pragmatic, and Discourse Factors in the Development of Case*. Amsterdam: John Benjamins; 123–59.

Bobb, Susan C., Zofia Wodniecka, Judith Kroll (2013) "What Bilinguals Tell Us About Cognitive Control: Overview to the Special Issue." [In:] *Journal of Cognitive Psychology* 25 (5); 493–96.

Borges, Robert (2019) "Rapid Automatized Picture Naming as a Proficiency Assessment for Endangered Language Contexts: Results from Wilamowice." [In:] *Journal of Communication and Cultural Trends* 1 (1); 1–25.

Borges, Robert (2022a) "MoReDaT – a Modular Remote Data Collection Toolkit for Linguistics +." A conference paper from: *Digital Humanities in the Nordic and Baltic Countries (DHNB 2022): Digital Humanities in Action* conference which was held at Uppsala University in March 2022. The date of talk 2022-03-17. [At:] https://ispan.waw.pl/nespomila/static/assets/Borges_20220317-DHNB_MOREDAT.pdf [date of access: 22.06.2022].

Borges, Robert (2022b) Supplement the manuscript "Sourcing Data from Wikipedia for the Study of Language Contact: the csbwiki". Zenodo. [At:] https://doi.org/10.5281/zenodo.7442107 [date of access: 22.06.2022].

Bybee, Joan (2010) *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Cognola, Federica, Ermenegildo Bidese (2016) "On Language Acquisition and Language Change: Is Transmission Failure Favoured in Multilingual Heritage Contexts?" [In:] Ermenegildo Bidese, Federica Cognola, Manuela Caterina Moroni (eds.) *Linguistik Aktuell/Linguistics Today* 234. Amsterdam: John Benjamins; 337–370.

Diessel, Holder (2012) "New Perspectives, Theories and Methods: Diachronic Change and Language Acquisition." [In:] Alexander Bergs, Laurel Brinton (eds.) *Historical Linguistics of English*. Berlin: De Gruyter; 1599–1613.

Fernández, Eva M., Ricardo Augusto de Souza, Agustina Carando (2016) "Bilingual Innovations: Experimental Evidence Offers Clues Regarding the Psycholinguistics of Language Change." [In:] *Bilingualism: Language and Cognition* 20 (2); 251–68.

Green, David W., Jubin Abutalebi (2013) "Language Control in Bilinguals: The Adaptive Control Hypothesis." [In:] *Journal of Cognitive Psychology* 25 (5); 515–30.

Gries, Stefan Th., Gerrit Jan Kootstra (2016) "Structural Priming Within and Across Languages: A Corpus-Based Perspective." [In:] *Bilingualism: Language and Cognition* 20 (2); 235–50.

Heine, Bernd, Tania Kuteva (2005) *Language Contact and Grammatical Change*. Cambridge University Press.

Hell, Janet G. van, Darren Tanner (2012) "Second Language Proficiency and Cross-Language Lexical Activation." [In:] *Language Learning* 62 (3); 148–71.

Hiltunen, Turo (2014) "Choice of National Variety in the English-Language Wikipedia." [In:] Jukka Tyrkkö, Sirpa Leppänen (eds.) *Studies in Variation, Contacts and Change in English* online journal special issue entitled: *Texts and Discourses of New Media*, Volume 15. [At:] https://varieng.helsinki.fi/series/volumes/15/hiltunen/ [date of access: 22 June 2022].

Hiltunen, Turo, Jukka Tyrkkö (2019) "Academic Vocabulary in Wikipedia Articles: Frequency and Dispersion in Uneven Datasets." [In:] *From Data to Evidence in English Language Research. Language and Computers 83*. Helsinki: University of Helsinki; 282–306.

Hopper, Paul (1987) "Emergent Grammar." [In:] Jon Aske, Natasha Beery, Laura Michaelis (eds.) *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Grammar and Cognition*; 139–57.

Hopper, Paul J. (2013) "Emergent Grammar." [In:] James Paul Gee, Michael Handford (eds.) *The Routledge Handbook of Discourse Analysis*. London: Routledge; 301–314.

Hróarsdóttir, Thorbjörg (2004) "Language Change and Language Acquisition." [In:] *Nordlyd* 31 (1) online journal special issue: Anne Dahl, Kristine Bentzen Peter Svenonius (eds.) *Proceedings of the 19th Scandinavian Conference of Linguistics*; 117–131.

Jiang, Menghan, Xiang Ying Shen, Kathleen Ahrens, Chu-Ren Huang (2021) "Neologisms Are Epidemic: Modeling the Life Cycle of Neologisms in China 2008–2016." [In:] Joshua Snell (ed.) *PLOS ONE* 16 (2): e0245984.

King, Brian (2015) "Wikipedia Writing as Praxis: Computer-Mediated Socialization of Second-Language Writers." [In:] *Language, Learning and Technology* 19 (3); 106–123.

Kootstra, Gerrit Jan, Janet G. van Hell, Ton Dijkstra (2012) "Priming of Code-Switches in Sentences: The Role of Lexical Repetition, Cognates, and Language Proficiency." [In:] *Bilingualism: Language and Cognition* 15 (4); 797–819.

Kroll, Judith F., Ellen Bialystok (2013) "Understanding the Consequences of Bilingualism for Language Processing and Cognition." [In:] *Journal of Cognitive Psychology* 25 (5); 497–514.

Langacker, Ronald W. (1987) *Foundations of Cognitive Grammar: Theoretical Prerequisites. Foundations of Cognitive Grammar*, Volume 1. Stanford, USA: Stanford University Press.

Lightfoot, David (2007) "Language Acquisition and Language Change: Inter-Relationships." [In:] *Language and Linguistics Compass* 1 (5); 396–415.

Makùrôt, Hanna (2016) *Gramatika kaszëbsczégò jãzëka*. Gdańsk: Zrzeszenie Kaszubsko-Pomorskie.

Margaretha, Eliza, Harald Lüngen (2014) "Building Linguistic Corpora from Wikipedia Articles and Discussions." [In:] *JLCL* 29 (2); 59–82.

Matras, Yaron (2009) *Language Contact*. Cambridge: Cambridge University Press.

Maturzysty, Vademecum (1992) *Język Polski*. Warszawa: Wydawnictwo "Oświata".

Meisel, Jürgen M. (2011) *First and Second Language Acquisition: Parallels and Differences. Cambridge Textbooks in Linguistics*. Cambridge: Cambridge University Press.

Mufwene, Salikoko S. (2001) *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Nelken, Rani, Elif Yamangil (2011) "Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms." [In:] *AAAI*; 31–36.

O'Rourke, Bernadette, Joan Pujolar, Fernando Ramallo (2015) "New Speakers of Minority Languages: The Challenging Opportunity – Foreword." [In:] *International Journal of the Sociology of Language* 2015 (231); 1–20.

Paulston, Cheistina (1994) *Linguistic Minorities in Multilingual Settings: Implications for Language Policies*. Amsterdam: John Benjamins.

Schmid, Hans-Jörg (2016a.) "A Framework for Understanding Linguistic Entrenchment and Its Psychological Foundations." [In:] Hans-Jörg Schmid (ed.) *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Berlin: De Gruyter Mouton; 9–36.

Schmid, Hans-Jörg (ed.) (2016b) *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Berlin: De Gruyter Mouton.

Schmid, Hans-Jörg (2020) *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Shi, Jing (2015) "An Analysis of the Application of Wikipedia Corpus on the Lexical Learning in the Second Language Acquisition." [In:] *English Language Teaching* 8 (8); 171–80.

Sloetjes, Han, Peter Wittenburg (2008) "Annotation by Category: ELAN and ISO DCR." [In:] *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakesh: European Language Resources Association (ELRA).

Thomason, Sarah, Terrance Kaufmann (1988) *Language Contact, Creolization and Genetic Linguistics*. Berkely: University of California Press.

Tomás, Jesús, Jordi Bataller, Francisco Casacuberta, Jaime Lloret (2008) "Mining Wikipedia as a Parallel and Comparable Corpus." [In:] *In Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling-2008)*; 1–8.

Weinreich, Uriel, William Labov, Marvin Herzog (1968) *Empirical Foundations for a Theory of Language Change*. Austin: University of Texas Press.

Willis, Sherry L., Klaus W. Schaie, Mike Martin (2009) "Cognitive Plasticity." [In:] Vern L. Bengtson, Richard A. Settersten, Jr., Brian K. Kennedy, Nancy Morrow-Howell, Jacqui Smith (eds.) *Handbook of Theories of Aging*. New York: Springer; 295–322.

Winford, Donald (2003) *An Introduction to Contact Linguistics*. Oxford: Blackwell.

Yano, Tae, Moonyoung Kang (2008) "Taking Advantage of Wikipedia in Natural Language Processing." [At:] https://www.cs.cmu.edu/~taey/pub/wiki.pdf [date of access: 22.06.2022].

Received:
24.06.2022
Reviewed:
28.08.2022
Accepted:
17.10.2022