2023. *Proc Ling Soc Amer* 8(1). 5467. https://doi.org/10.3765/plsa.v8i1.5467.

On "historical unity" of Russian and Ukrainian: A linguistic perspective on language conflict and change

Anyssa Murphy, Lex Whalen, Stanley Dubinsky, Michael Gavin, John F. Bailyn & Jackson Ginn*

Abstract. This paper focuses on Putin's (2021) misguided claim regarding "historical [linguistic] unity" of Russian and Ukrainian. Their being two distinct languages is not in question, as opposed (for example) to Serbian and Croatian. However, it is important to substantiate the objective reality of those differences, taking a strong stand against unjustified claims about linguistic [unity] where there are no grounds for them. Implementing a Python-coded algorithm, like those described in Nerbonne & Kretzschmar 2013, we calculate Levenshtein distance between frequency-based word lists, in a manner sensitive to both organic and contact-induced change, to fully reveal Ukrainian's complex relationship with both Russian and Polish.

Keywords: language conflict; Ukrainian; Russian; Polish; lexical similarity

1. Introduction. Vladimir Putin (2021) writes that "Russians, Ukrainians, and Belarusians are all descendants of Ancient Rus … Slavic and other tribes across the vast territory … [who] were bound together by *one language* …, economic ties, … [and] faith" (emphasis added). He claims "people both in the western [i.e., Ukrainian and Belarusian] and eastern [i.e., Muscovite] Russian lands spoke the same language [n.b., *spoke*, not *speak*]." This claim, used to justify Russian sovereignty over former Soviet territories, should be approached with a healthy amount of skepticism.

As such, this paper focuses on the misguided claim regarding "historical [linguistic] unity" of Russian and Ukrainian. Among speakers of both Russian and Ukrainian in Ukraine, there is no question about the reality of their being two distinct languages, as opposed (for example) to Serbian and Croatian. To substantiate the objective reality of those differences, we employ lexical distance measures which assess actual linguistic distance, in a manner sensitive to both organic and contact-induced change, to fully reveal Ukrainian's complex relationship with both Russian and Polish.

Considering how the shifting of borders of Ukraine have led it to become distinct from Russian, we observe the effects of political borders upon the linguistic ideology/identity of the peoples on either side of them. The interplay of language ideology and language change reveals that the first is subject to shifting political climates while the second is immutable, and in this way, Ukraine is a case study for ways in which linguistic and political borders intersect, and, in turn, how dialect continua are affected by political conflict. While border changes can instigate

*Authors: Anyssa Murphy, University of South Carolina (ajmurphy@email.sc.edu), Lex Whalen, University of South Carolina (lawhalen@email.sc.edu), Stanley Dubinsky, University of South Carolina (dubinsk@mailbox.sc.edu), Michael Gavin, University of South Carolina (mgavin@mailbox.sc.edu), John F. Bailyn, Stonybrook University (john.bailyn@stonybrook.edu), Jackson Ginn, University of South Carolina (jrginn@email.sc.edu).

immediate changes to language ideology, ideologies alone do not determine linguistic form, and while language ideologies may change quickly, languages do not.

2. Background. International borders have moved across Ukraine throughout its history. For most of the last six centuries, it was part of other kingdoms and empires – first as part of the Polish-Lithuanian Commonwealth (PLC) and later within the Russian Empire. Ukrainian[1] has, during these times, been mostly recognized as merely a regional variety of a larger East Slavic continuum.

From about 1400 CE until the 17th-18th centuries, Ukraine was ruled by the Kingdom of Poland and the Polish-Lithuanian Commonwealth (Figure 1) and Ukrainian had long-term contact with Polish and linguistic influence from its capital city, Kraków. This period saw an high degree of interaction between the Ukrainian-speaking and Polish-speaking people. This is particularly true of those with the cultural and economic prestige to enable their participation in scholarly and literary activities (Łesiów et al. 1998, Sovtys 2020). As discussed in Sovtys 2020, in extending its rule over surrounding peoples, "Poland assimilated ... [the cultures of] Ukrainians, Belarusians, and Lithuanians, making their traditions the source of its power and vitality," uniting these people "against the threat of Moscow on one hand and Germany on the other" (pp. 30).

Figure 1. Polish–Lithuanian Commonwealth (1619)[2]

Legend: Polish Crown Possessions; Duchy of Prussia; Grand Duchy of Lithuania; Duchy of Courland, Lithuanian fiefdom; Duchy of Livonia

The influence of international boundaries on both objective and subjective linguistic realities cannot be overstated. The shifting of an international border in a region where related

[1] The term *Ukrainian* is used to refer both to Modern Ukrainian and its predecessors, Old Ukrainian, Ruthenian, and Prosta Mova.

[2] https://commons.wikimedia.org/wiki/File:Rzeczpospolita2nar.png

languages are spoken can quickly change the linguistic affiliations of the affected population. In the Ukrainian and Russian case, the changes were not merely affiliative, but were substantial and observable. As Figure 2 shows, Russian and Ukrainian (along with Belarusian) are originally part of an East Slavic dialect continuum and subfamily of the Slavic languages. While the imposition of a Polish border between Ukrainian and Russian speaking regions did not lead to Ukrainian speakers considering themselves to be speakers of Polish, it did disrupt the relationship between Ukrainian and Russian and led to Ukrainian both being influenced by Polish and losing some of its similarity to Russian.

Figure 2. Partial Slavic language tree, showing Polish, Ukrainian, and Russian[3]

There was a high degree of bilingualism in the Ukraine region of the Polish-Lithuanian Commonwealth, as native speakers of both Polish and Ukrainian engaged in linguistic and ethnic 'mixing' (Łesiów et al. 1998). It is no surprise that during this time, the Ukrainian language was heavily influenced by Polish, sometimes borrowing Polish words' pronunciations and spellings wholesale, and sometimes "Ukrainianiz[ing]" them. Ukrainian tended, in particular, to take up "words … for the designation of concepts and relations in civil life" (Łesiów et al. 1998: 396-397; Table 1).

[3] Lynch, Jack. February 2014. From Jack.Lynch@rutgers.edu

English	Ukrainian		Polish	Russian	
thank you	дякую	/dʲakuju/	dziękuję	спасибо	/spasʲibo/
onion	цибуля	/cibulʲa/	cebula	лук	/luk/
January	січень	/sičenʲ/	styczeń	январь	/janvarʲ/
morning	ранок	/ranok/	rano	утро	/utro/
red	червоний	/červonij/	czerwony	красный	/krasnyj/
second	другий	/drug'ij/	drugi	второй	/vtoroj/

Table 1. Ukrainian/Polish lexical correspondence, contrasted with Russian

Centuries later, most of Ukraine (excepting Galicia and Transcarpathia) came to be part of the Russian Empire. In the 17th century, the rise of Cossack power against Poland along the lower Dnieper River (largely in response to the growing "Polonization" of the Ukrainian nobility) led to a Russo-Polish War (1654-1667) after which eastern Ukraine was transferred from Poland to Russia (Chynczewska-Hennel 1986). Cossack demands of Russia included recognition of the Ukrainian language, but this was denied under Tsarist rule By the start of the 19th century, Russia had banned all teaching of Ukrainian and by 1876, in response to the growth of Ukrainian literary language, prohibited the use of Ukrainian in publishing, public performances, and lectures (Pompino-Marschall et al. 2017). Thus, Ukrainian (an East Slavic language) developed for three centuries under the influence of Polish (a West Slavic language), followed by another three centuries mostly under East Slavic Russian influence, resulting in a language with a distinct identity and form.

Well before the 19th century, Ukrainian had become a language of a "national consciousness" (Chynczewska-Hennel 1986), and Russian imperial language policy through this time set the stage for later conflict after the Russian Revolution. The prohibitions on the use of Ukrainian at the end of the 19th c. in areas under Russian rule, as detailed in Pompino-Marschall et al. (2017), were in part a response to a growing Ukrainian literary movement. In areas not under Russian rule, especially in the Austrian Empire, Ukrainian was subject to other influences which helped to arouse language-centered nationalism among the population. Darden (2009) describes the language situation in western Ukraine under Austrian and Hungarian rule between 1867 and 1914. Austrian efforts to de-Russify Galicia (Figure 3) involved applying pro-Ukrainian language policies to promote Ukrainian ethnic nationalism, feeding a desire for independence after World War 1 and helping to spark the 1917-1921 Ukrainian War of Independence following the Russian Revolution.

Figure 3. Austro-Hungarian Empire, ca. end of 19th c.[4]

In Hungarian-ruled Transcarpathia, an ethnically identical Ukrainian region directly south of Galicia, such language policies were not applied. The results of Austria's pro-nationalist Ukrainian language policies are seen to persist into the 20th century when Ukrainians in former Galicia (but not those in former Transcarpathia) pursued armed rebellion against the Soviet Empire from 1946 to 1952. The battles of the Ukrainian Insurgent Army (Ukrainska Povstanska Armiia, or UPA) constituted the largest and most protracted armed resistance to Soviet rule after the Second World War, and it is significant that the UPA soldiers were recruited almost exclusively from the very region that was most influenced 50-70 years earlier by Austrian nationalistic language policies.

3. Methods in service of objective measures of linguistic difference

> *Although one should always respect a group's right to call their language whatever they want, [we] … take a strong stand against unjustified claims about linguistic [unity]* ***where there isn't any ground for them****.*
>
> [Bailyn 2020:26, emphasis added]

Moving away from subjective perceptions to objective measures of linguistic reality, we implement a R-coded algorithm, similar to those described in Nerbonne & Kretzschmar 2013, that calculates Levenshtein distance between frequency-based lists of words drawn from natural corpora. Reported in Dubinsky et al. 2022, this method reliably measures English, French, Spanish, and German, and appears to be sensitive to both phylogenetic similarity (PhS) and contact similarity (CS). Applying these measures to Polish/Ukrainian/Russian and Serbian/Bosnian/Croatian, we expected to find Ukrainian PhS with Russian and CS with Polish (in much the same way that English has PhS with German and CS with French).

These hypotheses were assessed using similarity measures between the lexicons of language pairs (with corpora that were more representative of speakers' working lexicons than are language dictionaries). The following languages were compared: Bosnian (Bs), Croatian (Hr),

[4] Adapted from https://freepages.rootsweb.com/~menzak/genealogy/maps/map-ah.jpg

English (En), French (Fr), German (De), Polish (Pl), Russian (Ru), Serbian (Sr), Spanish (Es), Ukrainian (Uk).

3.1. PROCEDURES:

(1) Word lists were generated from a subtitle corpus (OpenSubtitles.org) using open-source code (Hermit Dave 2019) to compile the most frequent 50K words for each language. Source word lists were automatically translated into each target language, using R Studios and Google Translate, yielding 4.5M data points across 45 language pairs.
(2) Grammatical stop words (e.g., articles and auxiliaries) were removed.
(3) Digraphs, non-Latin characters, and other opaque orthography were converted into regular, stable phonemic symbols, and then into R-readable code.
(4) Vowels were deleted from the character strings, as they are least likely to preserve cognate or borrowed forms.
(5) Levenshtein distance was calculated by comparing source consonant strings to those of a target. Perfect matches scored 1, partial phonetic similarity scored 0.66 or 0.33, and no-matches scored 0. L-distance was a sum of scores divided by source word character length.

The current data set retrieves lexical frequency lists from OpenSubtitles.org, a collection of film subtitles from hundreds of popular movies. As such, this dataset ideally represents a more naturalistic lexicon for the speakers of each language, representing the sorts of words a contemporary speaker may be expected to interact with on a day-to-day basis. This can be taken in contrast, for example, with lexical lists sourced from dictionaries, which may not actually represent a modern speaker's natural lexicon.[5]

Table 2 illustrates, with English source and Spanish target examples, how this method works. The English word *father* is translated into Spanish *padre*. Vowels are removed and digraphs replaced to yield *fθr* and *pdr*, respectively (note that "θ" replaces "th" irrespective of voicing, as the algorithm does not score on this basis). In comparing the two strings, f/p and θ/d are both partial matches and r/r is a match, yielding a score of 0.553 for the word pair. In the second case, English *fast* and Spanish *rápido* yield *fst* and *rpd*, respectively, with no match for either f/r or s/p and a partial match for t/d. The word pair scores 0.110. In the last case, English *rapid* translates to Spanish *rápido*, yielding *rpd* in both cases and a score of 1.000.

En Source	**S. output**	**Es Target**	**T. output**	**Score**
father	fθr	padre	pdr	0.553
fast	fst	rápido	rpd	0.110
rapid	rpd	rápido	rpd	1.000

Table 2. English > Spanish consonantal distance scores

[5] A few pertinent notes about our corpus: It is apparent that this corpus contains a certain bias to English, as many of the films included were first filmed in English and were subsequently translated into the target languages. The corpus also contained an apparent bias toward Russian, as we found the presence of some Russian words in the Ukrainian wordlist, as well as the presence of Russian characters that do not exist in Ukrainian (particularly, ы).

Table 3 and 4, here below, illustrate the similarity of Ukrainian-Russian and Ukrainian-Polish pairs, respectively. For these language pairings, Cyrillic characters were replaced with Latin letters in order to make the comparisons, using transliteration practices that have been standardized for each language.

En gloss	**Uk Source**	**S. output**	**Ru Target**	**T. output**	**Score**
how many	скільки (skilʹky)	sklʲk	сколько (skolʹko)	sklʲk	1.000
citizenship	громадянство (hromadyanstvo)	hrmdʲnstw	гражданство (grazhdanstvo)	grzdnstv	0.665
Thank you	дякую (dyakuyu)	dʲk	спасибо (spasibo)	spsʲb	0.165

Table 3. Ukrainian > Russian consonantal distance scores

En gloss	**Uk Source**	**S. output**	**Pl Target**	**T. output**	**Score**
together	разом (razom)	rzm	razem	rzm	1.000
citizenship	громадянство (hromadyanstvo)	hrmdʲnstw	obywatelstwo	byvtlstv	0.456
thank you	дякую (dyakuyu)	dʲk	dziękuję ci	d͡zkyt͡ɕ	0.665

Table 4. Ukrainian > Polish consonantal distance scores

In the first row of each table, we see a perfect match word pair. Table 3 has Ukrainian *скільки* ‘how many’ translated into Russian *сколько*, yielding the string *sklʲk* in both cases. Table 4 has Ukrainian *разом* ‘together’ translated into Polish *razem*, yielding the string *rzm* in both cases. Rows 2 and 3 illustrate the Ukrainian words *громадянство* ‘citizenship’ and *дякую* ‘thank you’ translated into Russian in Table 3 and into Polish in Table 4. There we find that the Russian translation of *громадянство* ‘citizenship’ yields a slightly higher similarity score than does the Polish translation of the word, and that the Polish translation of *дякую* ‘thank you’ yields a much higher similarity score than does the Russian translation.

4. Results. We focus here on the results of comparing two sets of three languages from among the 45 pairings compiled: Ukrainian-Russian-Polish (Uk-Ru-Pl) and Bosnian-Croatian-Serbian (Bs-Hr-Sr). The results of the six pairings of Uk-Ru-Pl and six pairings of Bs-Hr-Sr are provided in Table 5 and shown highlighted on the scatterplot in Figure 4. In both, the average lexical similarity (ALS) score (as a percentage) represents the average score of all source-target word pairs, where an average similarity score of 0.665 would be presented as 67%. The perfect match percentage (PMP) is the percentage of word pairs out of the total number of words paired that had a lexical similarity score of 1.000 (as illustrated by English *fast*-Spanish *rápido* in Table 2, Ukrainian *скільки* ‘how many’/Russian *сколько* in Table 3, and Ukrainian *разом* ‘together’/Polish *razem* in Table 4).

In Table 5, we see that the ALS score for Bs-Hr-Sr is uniformly quite high, ranging from 66% to 70%. We also note that Croatian and Serbian (in both directions) yield the highest PMPs (43-44%) Turning to the focus of this paper, Uk-Ru-Pl, we find that Russian and Ukrainian (in either direction) yield both the highest ALS scores (57-63%) and the highest PMPs (9-12%). At the same time, it is noteworthy that the PMPs of Ukrainian and Polish are higher (7%) than those of Russian and Polish (3-4%).

	ALS	**PMP**
pl → ru	36%	3%
pl → uk	39%	7%
ru → pl	37%	4%
ru → uk	57%	9%
uk → ru	63%	12%
uk → pl	40%	7%
bs → hr	67%	31%
bs → sr	66%	30%
hr → bs	68%	33%
hr → sr	69%	44%
sr → hr	68%	43%
sr → bs	70%	32%

Table 5. Average lexical similarity (ALS) and perfect match percentage (PMP)

These results come into sharper focus when highlighted in Figure 4 which plots, for each language pair, ALS along the x-axis and PMPs on the y-axis. The solid red line is a regression analysis of the expected correlation of ALS and PMP, based on the 45 language pairings sampled. Here, we see that Bs-Hr-Sr have collectively the highest ALS scores with all six pairings having higher than expected PMPs. We also see that PMPs for Croatian-Serbian are quite higher than those of any Bosnian pairing. Turning again to Uk-Ru-Pl, we can note the following: (i) the PMPs for Ukrainian and Russian are far less than what is expected and (ii) the PMPs for Ukrainian and Polish are rather higher than those of Russian and Polish.

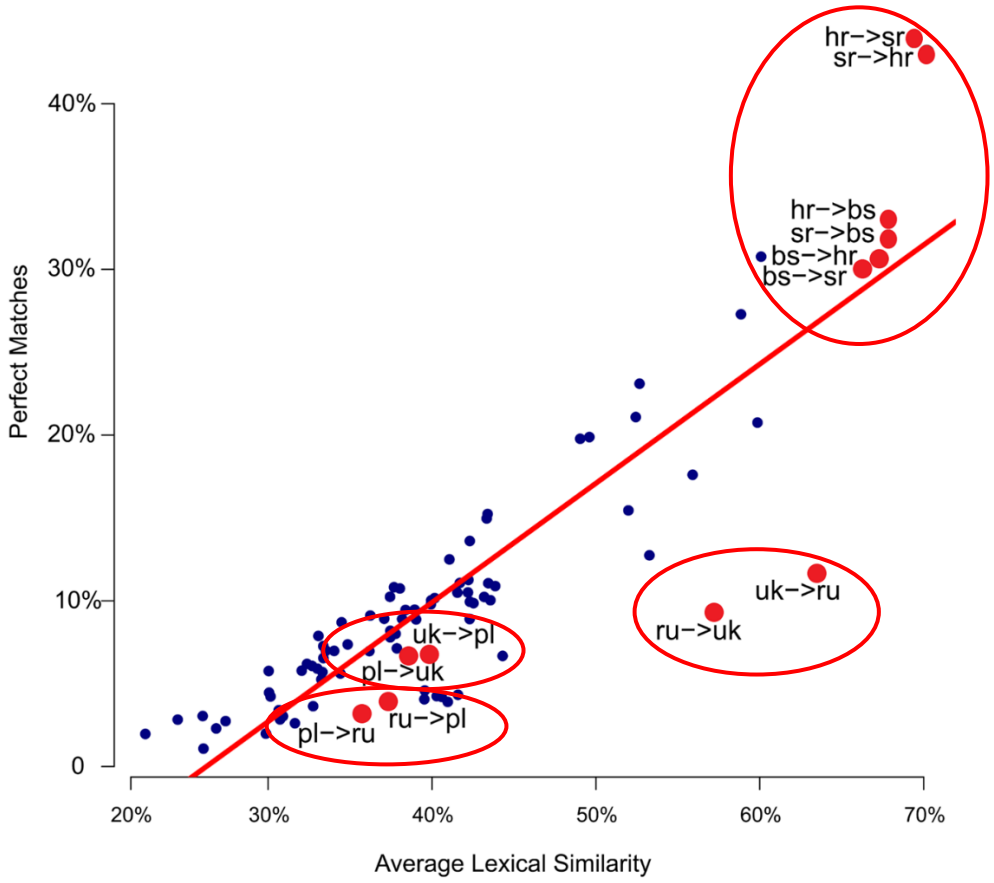

Figure 4. Scatterplot of average lexical similarity (x-axis) and perfect match percentage (y-axis). Bosnian (bs), Croatian (hr), Polish (pl), Russian (ru), Serbian (sr), Ukrainian (uk).

5. Discussion & conclusions. Considering what we've observed, we would say that Croatian and Serbian, on the extreme high end of ALS scores (68.5%) and PMPs (43.5%), display a level of lexical similarity and perfect matches one would expect from two varieties of the same language. On the other hand, it is clear that one cannot make similar claims for Ukrainian and Russian. Their combined ALS scores (60%) are unsurprising given their East Slavic phylogenetic relatedness. Yet, they register half the PMP that their similarity would predict (10.5% rather than ~20%). Note that the PMPs for Ukrainian and Polish, 7%, are twice those of Russian and Polish, 3.5%. Given that Ukrainian and Russian are both East Slavic languages and Polish is West Slavic, the higher-than-expected PMPs for Ukrainian and Polish, along with the unexpectedly low PMPs for Ukrainian and Russian, are suggestive of contact-induced similarity (credibly associated with Ukraine's long having been part of Poland). This supports the conclusion that Ukrainian, while maintaining its East Slavic phylogenetic relationship with Russian, has drifted away from it due to lexical borrowing from Polish, among other things.

5.1. LIMITATIONS. A key limitation in the present study concerns the corpus (OpenSubtitles.org) from which the word lists were drawn. This corpus tends to be derivative of scripts likely written in dominant languages (e.g., English, French, Russian), leading to an abundance of

vocabulary borrowed from English (e.g., *Superman* and *Hell Boy*). Similarly, Russian words were found inserted unchanged into Ukrainian subtitles. The overall results remain significant.

5.2. FUTURE DIRECTIONS. Contextualize current data by examining more languages. Use a wider range of corpora, correcting for OpenSubtitles bias and for the effect of genre on our measurements. Expand similarity measures beyond the lexicon.

References

Bailyn, John. 2020. Mistaken identity: Two case studies in the politicization of language. In Aleksandra Bednarowska, Beata Kołodziejczyk-Mróz & Piotr Majcher (eds.), *Slavic-German encounters in literature, culture and language II*, 13–40. Hamburg: Studien zur Germanistik. https://doi.org/10.1525/9780520337893-003.

Chynczewska-Hennel, Teresa. 1986. The national consciousness of Ukrainian nobles and Cossacks from the end of the sixteenth to the mid-seventeenth century. *Harvard Ukrainian Studies* 10(3/4). 377–392. https://www.jstor.org/stable/41036263.

Darden, Keith. 2009. Resisting occupation: Lessons from a natural experiment in Carpathian Ukraine. Manuscript. Yale University.

Dave, Hermit. 2019. Word list by frequency based on Open Subtitle corpus. Invoke IT Limited. Retrieved 18 December 2022. https://invokeit.wordpress.com/2019/02/15/word-list-by-frequency-based-on-open-subtitle-corpus-2018/.

Dubinsky, Stanley, Michael Gavin, Harvey Starr, Dawson Petersen, Gareth Rees-White, Kaitlyn Smith, Vivian D'Souza, Lex Whalen, Jackson Ginn & Ashley Bickham. 2022. The Language Conflict Project: Perspectives on 21st century ethnolinguistic conflict. Communication, Conflict and Peace. Online International Conference organized by the Archbishop Desmond Tutu Centre for War and Peace Studies, Liverpool Hope University.

Łesiów, Michał, Robert De Lossa & Roman Koropeckyj. 1998. The Polish and Ukrainian languages: A mutually beneficial relationship. *Harvard Ukrainian Studies* 22. 393–406. https://www.jstor.org/stable/41036749.

Nerbonne, John & William Kretzschmar, Jr. 2013. Dialectometry++. *LLC: Journal of Digital Scholarship in the Humanities* 28(1). 2–12. https://doi.org/10.1093/llc/fqs062.

Pompino-Marschall, Bernd, Elena Steriopolo & Marzena Żygis. 2017. Ukrainian. *Journal of the International Phonetic Association* 47(3). 349-357.

Putin, Vladimir. 2021. On the historical unity of Russians and Ukrainians. *President of Russia*, 12 July 2021. Presidential Executive Office. Accessed 6 July 2022: http://en.kremlin.ru/events/president/news/66181.

RStudio Team. 2018. RStudio: Integrated development for R. RStudio Inc., Boston, MA. http://www.rstudio.com/.

Sovtys, Natalia. 2020. The peculiarities of the Ukrainian-polish linguistic and cultural frontier. *Езиков свят-Orbis Linguarum* 18(2). 29–35. https://doi.org/10.37708/ezs.swu.bg.v18i2.4.