Winning Hearts & Tongues: A Polish to Lemko Case Study

Petro Orynycz
Orynycz.com
Washington, D.C.
p@orynycz.com

Tom Dobry
Antech Systems, Inc.
Lexington Park, Maryland
tom.dobry@antechsystems.com

ABSTRACT

When minority and local languages are lost, national security suffers: not only are significant increases in suicidality, depression, diabetes, assault, and substance abuse often documented, a void is created that has historically been exploited by adversaries. For example, millions from minority language communities ahistorically assume the Russian language and/or identity as their own in Ukraine, Belarus, NATO allies, and even the United States. If native language communication gaps remain in the hands of adversaries only, using their long experience with these languages, NATO remains at a major disadvantage attempting to engage these communities. In Europe, psychic wounds inflicted in part by language loss have not been closed by assimilation. Instead, cities experience bursts of isolating tensions in the West and eastern populations are convinced by adversarial powers that those powers are their true allies, who understand and respect them. Nor is education in the official language a panacea: in the case of Ukraine (and even Spain), non-trivial differences between local *lects* and the official language create openings for adversaries to fan the flames of separatism.

Using machine translation engines to empower NATO and its partners in training recruits or acting on the ground in the language closest to their hearts and minds can win immediate 'us'-ness and showcase NATO's embraced polycultural vision. Artificial intelligence and rule-based engines were assembled to translate between the official language of Poland and that of its indigenous Lemko minority, which has long been targeted by foreign powers. Engines were scored translating from Lemko to Polish using metrics developed with support from DARPA, producing a bilingual evaluation understudy (BLEU) score of 31.13 and translation edit rate (TER) of 54.10. Meanwhile, in the other direction, the engines scored TER 53.73 and BLEU 29.49, a score 6.5 times better than that of Google Translate's Polish-Ukrainian service.

ABOUT THE AUTHORS

Petro Orynycz (Петро Оринич) (https://orcid.org/0000-0003-3094-9156) is a scientist, software engineer at Antech Systems, Inc., computational linguist, and natural language engineer working to empower local communities and strengthen national security, public health and safety, education, and civil society by developing, deploying, and delivering cutting-edge solutions that leverage the latest breakthroughs in artificial intelligence and other technologies. His work currently focuses on machine learning, neural machine translation, and hybrid systems for endangered, minority, and Indigenous language revitalization. He graduated from the Institute of East Slavic Philology of Jagiellonian University in Cracow (Poland), where he was on Google Translate's Russian-English team amid its 2016 neural artificial intelligence breakthrough. He has been writing for *Springer Nature* publications, with his most recent works being *Say It Right: AI Neural Machine Translation Empowers New Speakers to Revitalize Lemko* and *BLEU Skies for Endangered Language Revitalization: Lemko Rusyn and Ukrainian Neural AI Translation Accuracy Soars.* His engines were also mentioned in the Cambridge University Press journal *Natural Language Engineering*. He has nearly two decades of transatlantic experience as a project manager, localization engineer and computational linguist specializing in Russian, Polish, Ukrainian, Rusyn and Lemko for top language service providers (LSPs), national defense, heavy industry, Raytheon, Siemens, BMW, Mercedes-Benz, investigators, philanthropists, and scientists.

Please cite as: Orynycz, P., & Dobry, T. (2023). Winning Hearts & Tongues: A Polish to Lemko Case Study. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Winning Hearts & Tongues: A Polish to Lemko Case Study

Petro Orynycz
Orynycz.com
Washington, D.C.
p@orynycz.com

Tom Dobry
Antech Systems, Inc.
Lexington Park, Maryland
tom.dobry@antechsystems.com

INTRODUCTION

Training outcomes stand to benefit from the use of machine translation for Indigenous and minority languages and dialects, whose usage is increasingly and significantly ($p \leq 0.05$) associated in scientific literature with sharper minds, more resilient psyches, and hardier health, not to mention sixfold lower suicide rates (Hallett et al., 2007, p. 398). Heritage language use may steel against foreign adversary influence, and in the North Atlantic theater, may prevent targeted populations from falling into Russian or other ahistorical ethnolinguistic identities when coping with the devastating aftermath of language loss. While the localization of materials into local dialects and languages may have previously been beyond the means of war-torn communities and governments, thanks to recent breakthroughs in artificial intelligence and computational linguistics, it is now possible to contemplate affordable devices that are cheaper, faster, and better than humans at translating into low-resource Indigenous and minority languages.

The problem of language loss is not limited to Europe. While the global language endangerment situation may not be as dire as available data had suggested in the early nineties, available statistics still paint a grim picture. In an oft-cited work dubbed "the great linguistic call to arms" by Simmons and Lewis (2013), Krauss had warned in 1992 that from half to 90% of the world's languages were set to become extinct this century. In addition, he had posited a "documented rate of destruction" of 90% of Indigenous languages in the Anglosphere, where English predominates, and an estimated 50% moribundity rate for the entire Soviet Union, where Russian was dominant (Krauss, 1992, p. 5). Twenty years later, Simmons and Lewis (2013) used updated data to estimate that 1,360 of 7,103 living languages (19%) are not being transmitted to the next generation (p. 12), a figure that rises to 30% in Eastern Europe (p. 13).

Neuroscience and Learning Outcomes

The latest research indicates that using a native language may mean more mental bandwidth is available for learning, and that test scores significantly improve. An investigation at the McGovern Institute for Brain Research headed by Massachusetts Institute of Technology (MIT) researchers earlier this year observed a relatively low brain response to native language stimuli when measured using the functional magnetic resonance imaging (fMRI) technique (Malik-Moraleda et al., 2023). As an explanation, the researchers suggested that expertise reduces the amount of brainpower required for a task (Mesa, 2023). In a recent study for the World Bank, Soh, Del Carpio and Wang (2021) found that using a non-native language of instruction may be detrimental, and to males especially. In the study, math and science test scores among students in Malaysia dropped significantly after the language of instruction was switched from Malay to English (Soh et al., 2021, pp. 4, 17, 18–19).

National Security

According to North Atlantic Treaty Organization (NATO) Special Operations School faculty members White and Overdeer, Russia may exploit ethnic cleavages in targeted societies as a lever of hybrid warfare in an attempt to achieve foreign policy objectives (2020, pp. 31–33), with ethnolinguistic differences being "readily available and easy to exacerbate" (p. 40). Below, the instigation and exploitation of ethnolinguistic strife in both western and eastern Europe is explored.

Spain: Catalonia

The public use of Catalan, a minority language spoken in Northeastern Spain, was prohibited by the Franco government until 1975 (Miller & Miller, 1996, p. 113). Rather than resolve strife, that policy may have caused it to fester. In a story for *The New York Times*, Schwirtz and Bautista (2021) cited a June 2020 European intelligence report asserting that the Russian Federation military intelligence system's elite *Unit 29155* had been on the ground in

Catalonia around the time of a 2017 independence referendum when the "secretive protest group" *Tsunami Democràtic* occupied the Barcelona airport and cut off the main highway linking Spain to its northern neighbors. Three days later, a colonel in Russia's Federal Protective Service and a close relative of a top presidential adviser deeply involved in Russia's efforts to support separatists in Ukraine flew in from Moscow for a strategy session to discuss the Catalan independence movement (Schwirtz & Bautista, 2021).

Russian Federation support for the Catalan independence movement reportedly even included an offer of 10,000 troops and 500 billion United States dollars in the event of independence (Baquero et al., 2022; see also Brunet, 2022, p. 74). Louise I. Shelley of the Terrorism, Transnational Crime and Corruption Center at George Mason University in Virginia called Russia reaching out to separatist leaders in Spain consistent with past behavior, and explained, "The linkages between the Catalonians and the Russians go back to the Soviet era. Before the collapse of the USSR, high-level meetings were held in Barcelona with distinguished Russians" (Baquero et al., 2022).

Western Ukraine

In Ukraine, non-trivial differences between local *lects* and the literary standard taught in schools create openings for adversaries to stoke the flames of separatism. According to a 2012 report by Rating, only 54% of ethnic Ukrainians used their heritage language, with 29% using Russian and 17% a mix of the two (p. 9). That year, nine books were printed in Russian for every one in Ukrainian, and only 13% of print media copies were written in Ukrainian (Moser, 2016a, p. 604).

Two decades ago, the United States Department of State's annual Country Reports on Human Rights Practices for 2002 reported as follows:

> Some pro-Russian organizations in the eastern part of the country complained about the increased use of Ukrainian in schools and in the media. They claimed that their children were disadvantaged when taking academic entrance examinations, since all applicants were required to take a Ukrainian language test. (Department of State, 2003, p. 1758)
>
> Rusyns (Ruthenians) continued to call for status as an official ethnic group in the country. Representatives of the Rusyn community have called for Rusyn-language schools, a Rusyn-language department at Uzhhorod University, and for Rusyn to be included as one of the country's ethnic groups in the 2001 census. According to Rusyn leaders, more than 700,000 Rusyns live in the country. (Department of State, 2003, p. 1759)

As a starting point for the wider issues mentioned by the Department of State, which are outside the scope of this paper, former Harvard Ukrainian Research Institute fellow Michael Moser explained:

> Rusyns can probably be best described as those remainders of Ruthenians/Rusyns who have not been willing to join the modern Ukrainian national and linguistic movement… initially this reluctance was not based on any Rusyn identity in the modern sense, but resulted from Russophile views that Ruthenians/Rusyns/Little Russians belong to one indivisible Russian people and there was no place for a Ukrainian nation and a Ukrainian language. (Moser, 2016b, p.127)

In June 2007, the "Russian World Foundation" was founded in Moscow by presidential decree, and started funding "compatriots" in Ukraine, bestowing over 1,200,000 United States dollars by March 2011 (Moser, 2016a, p. 607).

A gathering took place at the Russian Drama Theater in the far-western city of Mukachevo, Ukraine, on October 25, 2008 (Wiktorek, 2010, p. 100). There were even reports of a hundred-odd out-of-towner armed individuals outside (Ukrajins′ke nacional′ne objednannja, 2009; see also Wiktorek, 2010, p. 100). Whatever happened there, at 8:30pm that night, a proclamation of "restoration of Rusyn statehood" appeared in Russian on the online platform `rusin.forum24.ru`. It mentions among its grievances "the replacement of the Rusyn state language with Galician Ukrainian, the language of Polish Galicia, Rusyns' northern neighbor." (2-nd Europan [*sic*] Congress Subcarpathion [*sic*] Rusyns, 2008).

In the run-up to ordering his army to overtly invade Ukraine to conduct a widescale "special military operation," the president of the Russian Federation had devoted a full paragraph to the "fate of Subcarpathian Rus'" in his essay *On the Historical Unity of Russians and Ukrainians*:

> I will separately discuss the fate of Subcarpathian Rus', which ended up in Czechoslovakia after the collapse of Austria-Hungary. A considerable portion of the local inhabitants comprised Rusyns. Although it is now rarely remembered, after the liberation of Transcarpathia by Soviet troops, a congress of the Orthodox population of the territory declared support for inclusion of Subcarpathian Rus' into the Russian Soviet Federative Socialist Republic or directly into the Soviet Union as a separate, Carpatho-Russian republic. (Putin, 2021)

In another incident in the region, two members of the Polish far-right organization *Falanga*, whose members had been on the ground among Russian separatists in Eastern Ukraine, set fire to a cultural center of the Hungarian indigenous ethnolinguistic minority in the regional capital of Uzhhorod in 2018 by dousing it with gasoline and throwing in a Molotov cocktail (Górzyński, 2018).

Health and Safety

Suicidality

Sixfold higher suicide rates have been observed in communities where fewer than half report conversational knowledge of their heritage language (Hallett et al., 2007, p. 398). On a positive note, youth suicide rates dropped to zero in all cases but one where a majority reported ability to hold a conversation in their heritage language (p. 397). In a 2022 study by Pezzia and Hernandez, those who did not speak a heritage language fluently, but whose parents did (p. 95), were most likely to have suicidal thoughts (p. 98). As an explanation for the tie between language loss and suicidal ideation, Pezzia and Hernandez suggest "acculturative stress or social exclusion" resultant from acceptance as a full member of one's ethnic group being prevented by lack of fluency in its language (p. 100).

Depression

After controlling for age, gender, education, financial situation, and ethnic group membership, researchers found that concealment of identity by avoiding use of a heritage language in public (termed *language avoidance*) is a statistically significant ($p = 0.006$) predictor of being categorizable as "depressed" owing to production of a score of 5 or higher on Kroenke and Spitzer's Patient Health Questionnaire 9 (Olko et al., 2023, pp. 5–6). As a theorized mechanism, the researchers mentioned ethnic discrimination inducing chronic stress, leading to persistent hyperactivity of the hypothalamic-pituitary-adrenal axis and resultant heightened levels of corticotropin-releasing factor and cortisol, pointing to the work of Willner (2017), as well as Slavich and Irwin (2014).

Diabetes

After adjustment for socio-economic factors, diabetes mellitus was significantly ($p = 0.005$) less prevalent in communities with Indigenous language knowledge (Oster et al., 2014, p. 9).

Tobacco use

Being more English-language acculturated has been significantly associated with smoking among older Asian American adolescents in New York City (Rosario-Sim & O'Connell, 2009). In another study, use of English at home was associated with higher smoking prevalence rates among Asian American youth ($p = 0.021$), as was high English proficiency ($p = 0.040$) (Chen et al., 1999, p. 325). Among Hispanic girls, those who spoke English with their parents smoked more than those who spoke both English and Spanish with their parents ($p < 0.0001$), as well as girls who spoke Spanish with their parents ($p < 0.01$) (Epstein et al., 1998, p. 586).

Substance use and assault

According to the Australian Bureau of Statistics (2011/2012), Aboriginal youth between the ages of fifteen and twenty-four years who spoke an Indigenous language were less likely to have used illicit substances (16% vs 26%), less likely to report binge drinking in the previous two weeks (18% vs. 34%), and less likely to have been a victim of physical or threated violence in the previous year (25 vs 37%).

SOLUTIONS SO FAR

Neural Artificial Intelligence

The neural machine translation breakthrough by an international team with Defense Advanced Research Projects Agency (DARPA) funding under the Broad Operational Language Translation (BOLT) project (Cho et al., 2014) as well as Google (Sutskever et al., 2014) gave rise to engines capable of achieving quality scores on par with those of humans. However, training neural engines requires more data than is generally available for low-resource languages.

Rule-Based Machine Translation

Rule-based translation engines of the past were generally considered to have been wastes of money (Hajič et al., 2000, p. 7) with the notable exception of the Prague-based RUSLAN system funded by the Soviet-founded Council for Mutual Economic Assistance (COMECON), which produced Czech to Russian translations of mainframe computer operating system documentation (p. 7), with translations of two in five sentences being correct, another two in five only containing minor errors, and only one in five requiring substantial editing or retranslation (p. 8).

The main reasons given for the apparent disappointment in Prague over the results of Czech to Russian rule-based systems was that the task itself was too complex, and that Czech and Russian are not closely related enough to make such an approach viable. Unrealistic expectations and lack of objective evaluation metrics might be added to the list. Meanwhile, results translating from Czech into Slovak and Polish, all more closely related West Slavic languages, were quite encouraging (Hajič et al., 2000, p. 12).

Hybrid Neural/Rule-Based Machine Translation

In results presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), a rule-based Lemko to Polish engine was combined with a Polish to English rule-based engine to produce the world's first published results for machine translations from Lemko to English (Orynycz et al., 2021). The next year, translations in the opposite direction were produced by modifying the system and running it in reverse (Orynycz, 2022). Improvements made to that engine by overhauling it and increasing its vocabulary later led to a 35% improvement in translation quality (Orynycz, 2023).

NEW SOLUTIONS

Rule-Based Machine Translation Expert System

An inference engine was hand coded via test-driven development to reflect truths contained in a knowledge base assembled in consultation with the work of subject area experts. This approach also allows for manual elimination of foreign interference and purging of Russian and other loanwords. Dictionaries consulted included Horoszczak's bidirectional Polish-Lemko dictionary (2004), Pyrtej's Lemko-Ukrainian dictionary (2004), Duda's Ukrainian-Lemko dictionary (2011), and Rieger's Lemko-Polish glossary (1995), as well as his Lemko-Polish glossary based on recordings from the village of Bartne (2016). The grammars of Fontański and Chomiak (2000) as well as Pyrtej (2013) were consulted in coding rules to inflect words by grammatical categories such as number, case, and gender.

Transformer Artificial Intelligence

The neural machine translation breakthrough was followed closely by the introduction by scientists at Google Brain and Google Research of the *Transformer* architecture, which is based solely on attention mechanisms and dispenses with recurrence and convolutions entirely (Vaswani et al., 2017). For this experiment, we trained transformer based artificial intelligence models to translate from Polish into Lemko, and as far as we are aware, are first to publish results.

MATERIALS AND METHODS

Material

Data

Artificial intelligence models were created using a corpus comprising 1,611,352 source words (as counted by Microsoft Word 365) across 112,507 lines penned by Polish-born native speakers of Lemko, together with their translations into Polish by the Google Cloud Platform Translation Application Programming Interface (API) configured to translate as if from Standard Ukrainian using neural machine translation.

Lemko (also known as *Lemko Rusyn*) genetically belongs to the southwestern Ukrainian dialect system, within which it is differentiated by fixed stress on the penultimate (next-to-last) syllable (Danylenko, 2020). Such dialects are indigenous to territories now under the governance of Poland and, since 1993, the Slovak Republic.

In interwar Poland, the government fostered separate Lemko, Hutsul, and Boiko identities in an effort to counteract the Ukrainian movement, whose teachers had been dismissed (Moser, 2016b, p. 128). In 1935, Russophile teachers were replaced with Poles, and Lemko was finally removed from schools in 1937 (p. 128). About two-thirds of Lemko speakers in Poland were deported to Ukraine between 1945 and 1947, with the remaining 40,000 to 50,000 resettled primarily to newly annexed, formerly German territories of Communist Poland (p. 131). According to preliminary results for Poland's 2021 census, 12,700 listed "Lemko" as an ethnicity (Główny Urząd Statystyczny, 2023, p. 3).

Methods

Preprocessing

First, all text was lowercased. Next, a space was added before and after all non-alphanumeric characters. Initial and final whitespace was also stripped from each line. Then, the above corpus was processed using Moslem's script (2023a) for cleaning and filtering parallel datasets (commit `db6f441`), leaving 33,612 lines comprising 610,990 source words as tallied by Microsoft Word 365.

Subword tokenization

Unigram subwording models were trained using Moslem's script (2021a) (commit `fbf2488`). Next, those models were employed to tokenize both the source and target text using subwording script number two of the same commit (Moslem, 2021b).

Data splitting

2,000 lines from the above corpus were split off for evaluation using Moslem's script (2023b) for that purpose (commit `e6decb7`).

Training artificial intelligence models

Artificial intelligence models were trained using the TensorFlow version of the OpenNMT toolkit for neural machine translation, which is the successor to Harvard's *seq2seq-attn* sequence-to-sequence model with attention (Klein et al., 2017, p. 68). The command for starting the training and evaluation loop was launched with automatic configuration for the *Transformer* model. Automatic evaluation was also enabled, and set to run every 5,000 steps using the bilingual evaluation understudy (BLEU) metric and export a model when a new high score was achieved. Training was conducted on the *Google Colabatory* platform utilizing NVIDIA A100 graphical-processing units and a high random-access memory runtime state. Training was permitted to run overnight.

Inference engine

A translation inference engine was crafted on the basis of Klein's Python serving client script (commit `2b196ff`) (2021), which was modified to accommodate source and target subword tokenization models, as well as optimize spacing and capitalization to better conform to the expectations of artificial intelligence models and end users. Translation predictions were saved to file for subsequent quality evaluation.

Quality evaluation
The quality of translations was evaluated using metrics whose development was funded by DARPA: both BLEU (Papineni et al., 2002) and the Translation Edit Rate (TER) (Snover et al., 2006). The scores themselves were calculated using the industry-standard methods developed at Amazon Research by Post (2018).

RESULTS

Translation Quality Scores

The experimental rule-based expert system outperformed all others by every metric when translating from Polish to Lemko and vice versa.

Polish to Lemko Translation Quality
When translating from Polish to Lemko, the experimental expert rule-based system achieved a bilingual evaluation understudy quality score of BLEU 29.49, which is 6.50 times better than Google Translate's Ukrainian service. Meanwhile, the experimental artificial intelligence Transformer neural machine translation system achieved a score of BLEU 15.90 after 30,000 training steps, which was 3.50 times better than Google Translate's Ukrainian. When measured using the alternative TER metric, the experimental expert, rule-based system scored TER 53.73, which is 61% better than Google Translate's Ukrainian service.

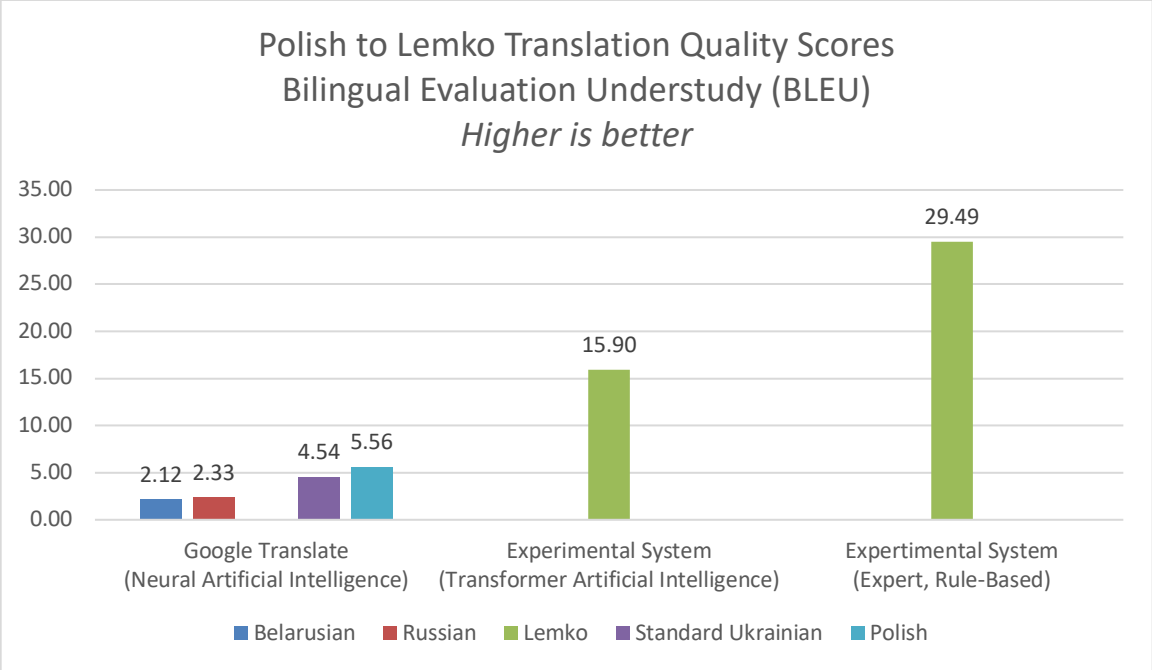

Figure 1. Polish to Lemko Translation Quality: BLEU Scores

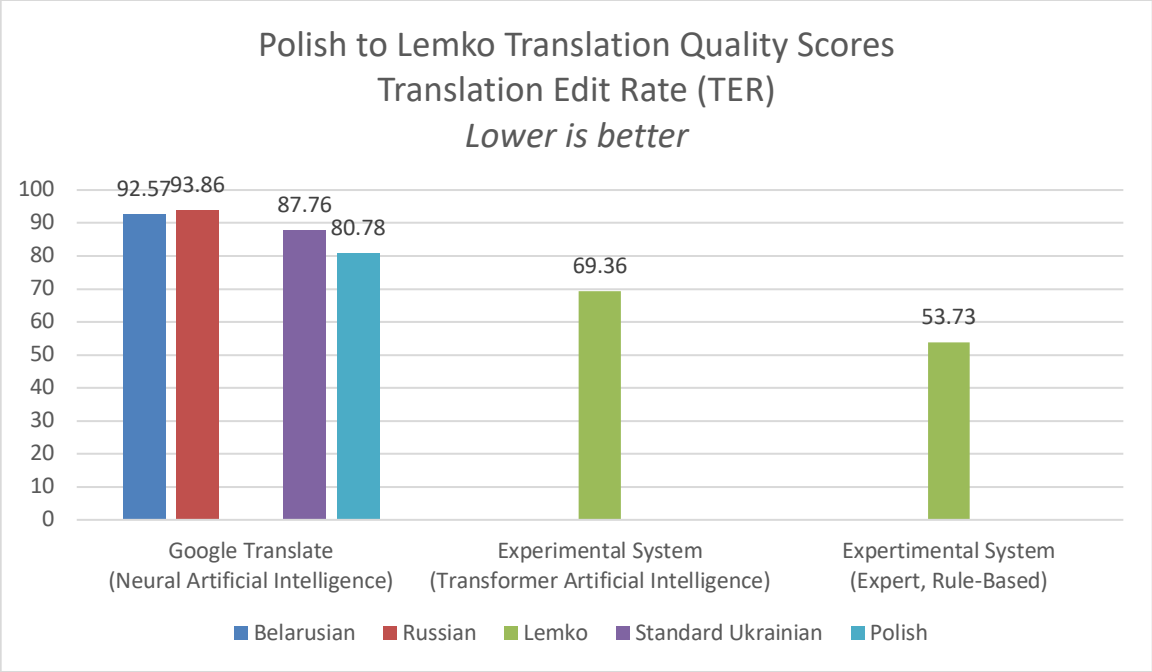

Figure 2. Polish to Lemko Translation Quality: TER Scores

Lemko to Polish Translation Quality

The experimental, rule-based expert system outperformed all others by every metric when translating from Lemko to Polish, achieving a bilingual evaluation understudy quality score of BLEU 31.13, which was 1.4 times better than the performance of Google Translate's Ukrainian service at BLEU 22.16.

Samples

Table 1. Example Polish to Lemko Translations

English meaning (human translator)	***In texts for example, and I mainly study texts, I have this source, they wrote: the Austrians were murdering us, so what will those awful Muscovites they're trying to scare us with do to us?***				
Polish (human translator)	Na przykład oni w tekstach, a ja głównie badam teksty, mam takie źródło, pisali: Austriacy nas mordowali, to co zrobią ci straszni Moskale, którymi nas straszą?				
Truth: Lemko reference (native speaker)	І они наприклад в текстах, а я ґоловні досліджам тексты, то значыт мам такє джерело, писали: но Австриякы нас мордували, то што зроблят тоты страшны Москалі, котрыма нас страшат?		I ony napryklad v tekstach, a ja holovni dosljidžam tekstŷ, to značŷt mam takie džerelo, pysaly: no Avstryjakŷ nas morduvaly, to što zrobljat totŷ strašnŷ Moskalji, kotrŷma nas strašat?		
System		**Translation Hypotheses**		**Quality Scores**	
		Cyrillic	**Transliteration**	**BLEU**	**TER**
Experimental	**Expert System (Rule-Based)**	Наприклад они в текстах, а я ґоловні бадам текстий, мам такы джерело, писали: Австриякы нас мордували, то што зроблят тоты страшны москале, котрыма нас страшом?	Napryklad ony v tekstach, a ja holovni badam tekstyj, mam takŷ džerelo, pysaly: Avstryjakŷ nas morduvaly, to što zrobljat totŷ strašnŷ moskale, kotrŷma nas strašom?	46.32	34.48
	Artificial Intelligence (Transformer)	Примірово, в текстах, а я ґоловні в заміріню тексту, маме джерело, писали: австриякы австриякы мордували, же то што зроблят стабілизацию тому, котрыма нас прострашыли?	Prymirovo, v tekstax, a ja holovni v zamirinju tekstu, mame džerelo, pysaly: avstryjakŷ avstryjakŷ morduvaly, že to što zrobljat stabilyzacyju tomu, kotrŷma nas prestrašŷly?	27.65	55.17
Google Translate	**Polish**	На прзиклад оні в текстах, а я ґлувнє бадам тексти, мам такє зьрудло, пісалі: Аустряци нас мордовалі, то цо зробьон ці страшні Москалє, ктурімі нас страшон?	Na przyklad oni v tekstach, a ja gluvnje badam teksty, mam takje źrudlo, pisalji: Austriacy nas mordovalji, to co zrobjon ci strašni Moskalje, kturymi nas strašon?	14.21	68.97
	Ukrainian	Наприклад, у своїх текстах, а я в основному досліджую тексти, у мене є таке джерело, вони писали: Австрійці нас повбивали, що будуть робити ті страшні москалі, якими вони нам погрожують?	Napryklad, u svojix tekstax, a ja v osnovnomu doslidžuju teksty, u mene je take džerelo, vony pysaly: Avstrijci nas povbyvaly, ščo budut' robyty ti strašni moskali, jakymy vony nam pohrožujut'?	9.43	82.76
	Russian	Например, в их текстах, а я в основном исследую тексты, у меня есть такой источник, они писали: Нас убили австрийцы, что будут делать те страшные москвичи, которыми они нам угрожают?	Naprimer, v ix tekstax, a ja v osnovnom issleduju teksty, u menja est' takoj istočnik, oni pisali: Nas ubili avstrijcy, čto budut delat' te strašnye moskviči, kotorymi oni nam ugrožajut?	9.43	86.21
	Belarusian	Напрыклад, у сваіх тэкстах, а я ў асноўным тэксты дасьледую, у мяне ёсьць такая крыніца, яны пісалі: Аўстрыйцы нас забілі, што будуць рабіць тыя страшныя маскалі, якімі яны нам пагражаюць?	Napryklad, u svaix tèkstax, a ja ŭ asnoŭnym tèksty das'leduju, u mjane ës'c' takaja krynica, jany pisali: Aŭstryjcy nas zabili, što buduc' rabic' tyja strašnyja maskali, jakimi jany nam pahražajuc'?	4.99	96.55

DISCUSSION

Policy Implications

Learning, public health, and security outcomes may improve if educational, training, community outreach, and other materials are localized into regional dialects and languages in addition to national standard ones. To avoid straining human resource capacities, linguists could be tasked with post-editing the output of expert and artificial intelligence machine translation systems, as opposed to translating by hand. More affordable access to translated materials could bring improvements to social services in underserved areas. Stonewall et al. list being multilingual, and thus inclusive, high on their list of best practices for engaging underserved populations (2017). The European Union has been funding research suggesting machine translation can be used to facilitate civic participation, as well as strengthen public health and safety among underserved communities (Nurminen & Koponen, 2020).

Technological Implications

Things are on track for commercially viable machine translation into Lemko at the press of a button to become a reality. Continued test-driven development of expert, rule-based systems seems poised to offer the quickest path to superhuman translation quality scores. Transformer-based artificial intelligence systems may win out in the long term.

Some tweaks to the artificial intelligence training procedure merit experimentation. The corpus filtering script may have been overzealous for this task and overly shrunk the corpus size, hindering performance. The script might be omitted in a future experiment. Overfitting may be hampering scores, and perhaps the evaluation interval of 5,000 steps should be shortened. Using the expert rule-based system to translate corpora into Polish from Lemko as opposed to the Google Cloud Platform service might result in better results. Incorporating automatic spelling correction modules might also improve scores globally.

Russian and other foreign linguistic interference might be countered programmatically by purging loanwords using *find-replace* algorithms. National language academies and other authorities might find such capabilities useful. It is possible that translation quality has already reached superhuman levels, a hypothesis that could be tested in future experiments.

ACKNOWLEDGEMENTS

We would like to thank our advisor, James Ohlman of CAE Inc. (United States), for his guidance.

DECLARATION OF COMPETING INTERESTS

The primary author serves as a quality control specialist for the Google Translate project out of San Francisco.

REFERENCES

2-nd Europan [*sic*] Congress Subcarpathion [*sic*] Rusyns [rusin]. (2008, October 25). MEMORANDUM 2-go Evropejskogo Kongressa Podkarpatskix Rusinov o prinjatii AKTA PROVOZGLAŠENIJA vosstanovlenija rusinskoj gosudarstvennosti [Memorandum of the Second European Congress of Subcarpathian Rusyns on the Adoption of a Proclamation of Restoration of Rusyn Statehood] [Online forum post]. *Informacionnoe Agenstvo Podkarpatskoj Rusi. IAPR. Forum podkarpatskix rusinov*. http://rusin.forum24.ru/?1-9-0-00000005-000-0-0-1224955832

Australian Bureau of Statistics, (2012). Culture, Heritage and Leisure: Speaking Aboriginal and Torres Strait Islander Languages. *Aboriginal and Torres Strait Islander Wellbeing: A focus on children and youth*. (Original work published 2011) Retrieved May 1, 2023 from https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1E6BE19175C1F8C3CA257A0600229ADC

Baquero, A., Hall, K.G., Tsogoeva, A., Albalat, J.G., Grozev, C., Bagnoli, L., IStories, & Vergine, S. (2022, May 8). *Fueling Secession, Promising Bitcoins: How a Russian Operator Urged Catalonian Leaders to Break With Madrid*. Organized Crime and Corruption Reporting Project (OCCRP). https://www.occrp.org/en/investigations/fueling-secession-promising-bitcoins-how-a-russian-operator-urged-catalonian-leaders-to-break-with-madrid

Brunet, F. (2022). *The Economics of Catalan Separatism*. Cham: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-14451-6

Chen, X., Unger, J.B., Cruz, T.B., & Johnson, C.A. (1999). Smoking patterns of Asian-American youth in California and their relationship with acculturation. Journal of Adolescent Health, 24(5), 321-328. https://doi.org/10.1016/S1054-139X(98)00118-9

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 http://dx.doi.org/10.3115/v1/D14-1179

Danylenko, A. (2020). "Carpatho-Rusyn", in: *Encyclopedia of Slavic Languages and Linguistics Online*, Editor-in-Chief Marc L. Greenberg. Consulted online on 13 June 2023 http://dx.doi.org/10.1163/2589-6229_ESLO_COM_031960

Department of State (2003). *S.Prt. 108-30, Volume I - COUNTRY REPORTS ON HUMAN RIGHTS PRACTICES FOR 2002 VOLUME I*. Washington, D.C: U.S. Government Publishing Office. https://www.govinfo.gov/app/details/CPRT-108JPRT86917/CPRT-108JPRT86917

Duda, I. (2011). *Lemkivs'kyj slovnyk* [A Lemko Dictionary]. Ternopil: Aston.

Epstein, J. A., Botvin, G.J., & Diaz, T. (1998). Linguistic acculturation and gender effects on smoking among Hispanic youth. Preventive medicine, 27(4), 583–589. https://doi.org/10.1006/pmed.1998.0329

Fontański, H., & Chomiak, M. (2000). *Gramatyka języka łemkowskiego* [A Grammar of the Lemko Language]. Katowice: „Śląsk" Sp. z o.o. Wydawnictwo Naukowe.

Główny Urząd Statystyczny (2023). *Wstępne wyniki NSP 2021 w zakresie struktury narodowo-etnicznej oraz języka kontaktów domowych* [Preliminary Results of the 2021 Census in Terms of National and Ethnic Structure and Language Used at Home]. Retrieved June 11, 2023 from https://stat.gov.pl/spisy-powszechne/nsp-2021/nsp-2021-wyniki-wstepne/wstepne-wyniki-narodowego-spisu-powszechnego-ludnosci-i-mieszkan-2021-w-zakresie-struktury-narodowo-etnicznej-oraz-jezyka-kontaktow-domowych,10,1.html

Górzyński, O. (2018, March 3). *Russia's Covert Campaign to Inflame East Europe*. The Daily Beast. https://www.thedailybeast.com/russias-covert-campaign-inflaming-east-europe

Hajič, J., Hric, J., & Kuboň, V. (2000, April). Machine translation of very close languages. In *Sixth Applied Natural Language Processing Conference* (pp. 7–12). http://dx.doi.org/10.3115/974147.974149

Hallett, D., Chandler, M.J., & Lalonde C.E. (2007): Aboriginal language knowledge and youth suicide. *Cognitive Development*. 22(3), 392–399. https://doi.org/10.1016/j.cogdev.2007.02.001

Horoszczak, J. (2004). *Słownik łemkowsko-polski, polsko-łemkowski* [Lemko-Polish and Polish-Lemko Dictionary], Warszawa: Rutenika.

Klein, G. (2021). *Inference with TensorFlow Serving*. Retrieved June 5, 2023, from https://github.com/OpenNMT/OpenNMT-tf/blob/master/examples/serving/tensorflow_serving/ende_client.py

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A.M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pp. 67–72. https://doi.org/10.18653/v1/P17-4012

Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1), 4–11. https://doi.org/10.1353/lan.1992.0075

Malik-Moraleda, S., Jouravlev, O., Mineroff, Z., Cucu, T., Taliaferro, M., Mahowald, K., Blank, I., & Fedorenko, E. Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI. Cold Spring Harbor Laboratory. Advance online publication. https://doi.org/10.1101/2023.01.19.524657

Mesa, N. (2023, February 3). Your native tongue holds a special place in your brain, even if you speak 10 languages. Science, https://doi.org/10.1126/science.adh0055

Miller, H., & Miller, K. (1996). *Language Policy and Identity: the case of Catalonia*. International Studies in Sociology of Education, 6(1). https://doi.org/10.1080/0962021960060106

Moser, M. (2016a). Language Politics in Contemporary Ukraine (25 February 2010–25 February 2011). In *New Contributions to the History of the Ukrainian Language* (pp. 601–619). Canadian Institute of Ukrainian Studies Press. https://www.ciuspress.com/product/new-contributions-to-the-history-of-the-ukrainian-language/

Moser, M. (2016b). *Rusyn: A New–Old Language In-between Nations and States*. In: Tomasz Kamusella, Motoki Nomachi, Catherine Gibson (Eds.), The Palgrave Handbook of Slavic Languages, Identities and Borders, 124–139. https://doi.org/10.1007/978-1-137-34839-5_7

Moslem, Y. (2021a). *Training SentencePiece models for the source and target*. Retrieved June 4, 2023, from https://github.com/ymoslem/MT-Preparation/blob/main/subwording/1-train_unigram.py

Moslem, Y. (2021b). *Subwording the source and target files*. Retrieved June 4, 2023, from https://github.com/ymoslem/MT-Preparation/blob/main/subwording/2-subword.py

Moslem, Y. (2023a). *Filtering/Cleaning parallel datasets for Machine Translation*. Retrieved June 4, 2023, from https://github.com/ymoslem/MT-Preparation/blob/main/filtering/filter.py

Moslem, Y. (2023b). *Splitting the parallel dataset into train, development and test datasets for Machine Translation*. Retrieved June 4, 2023, from https://github.com/ymoslem/MT-Preparation/blob/main/train_dev_split/train_dev_test_split.py

Nurminen, M., & Koponen, M. (2020). Machine translation and fair access to information. *Translation Spaces*, 9(1), 150–169. https://doi.org/10.1075/ts.00025.nur

Olko, J., Galbarczyk, A., Maryniak, J., Krzych-Miłkowska, K., Iglesias Tepec, H, de la Cruz, E., Dexter-Sobkowiak, E., & Jasienska, G. (2023): The spiral of disadvantage: Ethnolinguistic discrimination, acculturative stress and health in Nahua indigenous communities in Mexico. *American Journal of Biological Anthropology*, 1–15. https://doi.org/10.1002/ajpa.24745

Orynycz, P. (2022, May). Say It Right: AI Neural Machine Translation Empowers New Speakers to Revitalize Lemko. In *Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings* (pp. 567–580). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-05643-7_37

Orynycz, P. (2023, July). BLEU Skies for Endangered Language Revitalization: Lemko Rusyn and Ukrainian Neural AI Translation Accuracy Soars. In *International Conference on Human-Computer Interaction* (pp. 135–149). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35894-4_10

Orynycz, P., Dobry, T., Jackson, A., & Litzenberg, K. (2021). Yes I Speak… AI neural machine translation in multi-lingual training. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*. https://www.xcdsystem.com/iitsec/proceedings/index.cfm?Year=2021&AbID=96953&CID=862

Oster, R.T., Grier, A., Lightning, R., Mayan, M.J., & Toth, E.L. (2014). Cultural continuity, traditional Indigenous language, and diabetes in Alberta First Nations: a mixed methods study. *International Journal for Equity in Health*, 13(92), 1–11. https://doi.org/10.1186/s12939-014-0092-4

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318). https://doi.org/10.3115/1073083.1073135

Pezzia, C., & Hernandez, L.M. (2022). Suicidal ideation in an ethnically mixed, highland Guatemalan community. *Transcultural Psychiatry*. 59(1), 93–105. https://doi.org/10.1177/1363461520976930

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers,* pp. 186–191. Brussels: Association for Computational Linguistics http://dx.doi.org/10.18653/v1/W18-6319

Putin, V. *Ob istoričeskom edinstve russkix i ukraincev* [On the Historical Unity of Russians and Ukrainians]. Retrieved May 15, 2023 from http://kremlin.ru/events/president/news/66181

Pyrtej, P. (2004). *Korotkyj slovnyk lemkivs'kyx hovirok* [A Brief Dictionary of Lemko Dialects]. Ivano-Frankivs'k: Siversija MB.

Pyrtej, P. (2013). *Lemkivs'ki hovirky. Fonetyka i morfolohija* [The Lemko Dialects. Phonetics and Morphology]. Gorlice: Zjednoczenie Łemków.

Rating, (2012). *Pytannja movy: rezul'taty ostannix doslidžen' 2012 roku* [The Language Question: Results of the Latest Research in 2012]. Retrieved August 26, 2023 from https://ratinggroup.ua/files/ratinggroup/reg_files/rg_mova_dynamika_052012.pdf

Rieger, J. (1995). *Słownictwo i nazewnictwo łemkowskie* [Lemko Vocabulary and Nomenclature]. Warszawa: Wydawnictwo Naukowe Semper.

Rieger, J. (2016). *Mały słownik łemkowkiej wsi Bartne* [A Small Dictionary of the Lemko Village of Bartne]. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Rosario-Sim, M.G., & O'Connell K.A. (2009). Depression and Language Acculturation Correlate With Smoking Among Older Asian American Adolescents in New York City. *Public Health Nursing* 26(6), 532–542. https://doi.org/10.1111/j.1525-1446.2009.00811.x

Schwirtz, M., & Bautista, J. (2023, September 23) Married Kremlin Spies, a Shadowy Mission to Moscow and Unrest in Catalonia. *The New York Times*. Retrieved May 16, 2023 from https://www.nytimes.com/2021/09/03/world/europe/spain-catalonia-russia.html

Simmons, G.F., & Lewis, M.P. (2013). The world's languages in crisis: a 20-year update. In E. Mihas, B. Perley, G. Rei-Doval & K. Wheatley (Eds.), *Responses to Language Endangerment: In honor of Mickey Noonan. New directions in language documentation and language revitalization* (pp. 3–20). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.142.01sim

Slavich, G.M., & Irwin, M.R. (2014). From stress to inflammation and major depressive disorder: a social signal transduction theory of depression. *Psychological Bulletin*, 140(3), 774–815. https://doi.org/10.1037/a0035302

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, (pp. 223–231). https://aclanthology.org/2006.amta-papers.25

Soh, Y.C., Del Carpio, X.V., & Wang, L.C. (2021). The Impact of Language of Instruction in Schools on Student Achievement: Evidence from Malaysia Using the Synthetic Control Method. *World Bank Group Policy Research Working Paper 9517*. http://hdl.handle.net/10986/35031

Stonewall, J., Fjelstad, K., Dorneich, M., Shenk, L., Krejci, C., & Passe, U. (2017, September). Best practices for engaging underserved populations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 130–134). Sage CA: Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/1541931213601516

Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. https://proceedings.neurips.cc/paper_files/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Ukrajins′ke nacional′ne objednannja (2009). *Zakarpats'ke UNO obicjaje vlasnymy sylamy protydijaty separatystam* [Transcarpathian Ukrainian National Organization Promises to Counter Separatists on May 1st with its Own Forces] Retrieved June 10, 2023, from https://zaxid.net/zakarpatske_uno_obitsyaye_vlasnimi_silami_protidiyati_separatistam_1_travnya_n1076607

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

White, D.J., & Overdeer, D. (2020). Exploiting Ethnicity in Russian Hybrid Threats. *Strategos: Scientific journal of the Croatian Defence Academy* 4(1), 31–49. https://hrcak.srce.hr/242087

Wiktorek, A.C. (2010). *Rusyns of the Carpathians: Competing agendas of identity*. Washington, D.C.: Georgetown University. https://repository.library.georgetown.edu/handle/10822/552816

Willner, P. (2017). The chronic mild stress (CMS) model of depression: History, evaluation and usage. *Neurobiology of Stress*, 6, 78–93. https://doi.org/10.1016/j.ynstr.2016.08.002