Павел Малецкій
АҐГ Науково-Технічный Університет/AGH University of Krakow
ORCID: 0000-0001-5050-2111

Магдалена Піотровска
АҐГ Науково-Технічный Університет/AGH University of Krakow
ORCID: 0000-0002-3974-9559

Аналіза і клясифікация русиньской бесіды языковым модельом штучной інтеліґенциі OpenIA Whisper

Abstrakt
Analiza i klasyfikacja języka rusińskiego przy użyciu modelu sztucznej sieci neuronowej ASR OpenIA Whisper
Artykuł przedstawia analizę lingwistyczną języka rusińskiego, koncentrując się na jego złożonych i zmieniających się aspektach, takich jak wymowa oraz różnice indywidualne, regionalne i historyczne. Do przeprowadzenia badania wykorzystano sztuczną sieć neuronową opartą na modelu OpenAI Whisper. Model ten, choć szkolony na danych z większości państwowych języków urzędowych, nie był bezpośrednio trenowany na bazach próbek języka rusińskiego ze względu na jego lokalny i mniejszościowy/etniczny charakter. Stąd próbki mowy tego języka klasyfikowane były przy użyciu najbardziej zbliżonych dostępnych etykiet, co pozwoliło na wyznaczenie podobieństwa języka rusińskiego do innych słowiańskich języków. Badanie objęło użytkowników zróżnicowanych pod względem płci, wieku i lokalizacji (Polska, Ukraina, Słowacja, Serbia), wykazując znaczące podobieństwa do języków dominujących w tych krajach oraz zależności między wyznaczonym podobieństwem językowym a wiekiem mówców.

Słowa kluczowe: język rusiński, fonetyka, klasyfikacja, asymilacja, IA, sztuczne sieci neuronowe, ASR

Abstract
Analysis and Classification of the Rusyn Language Using the OpenAI Whisper ASR Model
The paper presents a linguistic analysis of the Rusyn language, focusing on its complex and dynamic aspects, such as pronunciation and individual, regional, and historical variations. The study employed a neural network based on the OpenAI Whisper automatic speech recognition (ASR) model. While trained on data from a majority of official state languages, the model lacked direct training on Rusyn language samples due to its localized and minority/ethnic nature. Consequently, speech samples were classified using the closest available language labels, enabling the identification of similarities between Rusyn and other Slavic languages. The study encompassed a diverse range of speakers across gender, age, and location (Poland, Ukraine, Slovakia, Serbia), revealing significant similarities to the dominant languages in these respective countries. Furthermore, the research highlights correlations between the identified linguistic similarities and the age of the speakers.

Keywords: Rusyn language, Phonetics, Classification, Assimilation, IA, ANN, ASR

Ключовы слова: русиньскій язык, фонетыка, класифікация, асиміляция, ШІ, штучны невроновы сіти, ASR

1. Вступ

Русиньскій язык то східньославяньскій язык хоснуваный через русиньску етнічну групу, головні в карпатскым регіоні Серединовой Європы, котрый обнимат част Украіны, Словациі, Польщы, Мадяр і Румуніі. Має місцевы одміны, што оддзеркаляют зложены історичны і культуровы влияня поєдных регіонів. Класифікация русиньского языка в шыршым контексті східньославяньскых і інчых славяньскых языків є зложеном і дискусийном темом з огляду на його унікальне положыня і історичны влияня. Русиньскій язык має спільны приметы зо східньославяньскыма языками (як росийскій, украіньскій і білорускій), а тіж выказує влияня західньославяньскых языків (польскій, словацькый) і полудньовославяньскых (сербскій, хорватскій). Тота вельоаспектова мішанина оддзеркалят його положыня на скрижуваню тых языковых груп (Kushko 2007).

Історічні русиньскій язык был залежный од языковой політыкы пануючых держав ци цисарств, медже інчыма Австро-Мадярской Імпериі ци Совітского Союзу. Языковый проблем є кісно повязаный з достоменністю, а дискусиі на тему того ци русиньскій то окремый язык, ци диялект украіньского, тырвают іщы аж і днес. Прихыльниками той другой теориі сут мало не выключні дослідникы і політыкы, якы ідентыфікуют Руснів як Украінців. В Украіні Русины не сут офіційно узнаны як окрема етнічна група, што іщы веце комплікує охорону языка (Nikitin і ін. 2009). Русиньскій язык был скодифікованый в XX ст., што оддзеркаляло одроджыня народовой і языковой достоменности. Кодифікацийны діяня довели до выокремліня регіональных літературных стандардів опертых на місцевых приметах – в Словациі, Польщы, Украіні і в Сербіі (Plišková 2008).

Русиньскій єст переважні клясифікуваный як східньославянскій язык з огляду на його історичне і языкове коріня. Його клясифікацию комплікує єднак значуче влияня західньо- і полудньовославяньскых языків (Moser 2016).

Зміннінст языка, што обнимат дияхронны, диястратычны, дияфазычны і диятопічны аспекты, то вызваня для обробкы натурального языка (NLP, en. Natural Language Processing). Диялектычны ріжниці медже одмінами того самого языка влияют на ефектывніст аплікаций, такых як машынове тлумачыня ци розпознаваня бесіды (Zampieri і ін. 2020). Прото тіж росне заінтерсуваня досліджынями над обробком

споріднених языків, языковых одмін і диялектів, на што доказом сут чысленны публикациі і науковы подіі, такы як варштаты VarDial (Zampieri і ін. 2020).

В примірі языків з низкыма ресурсами, як нп. русиньскій, придбати одповідні языковы корпусы єст особливі тяжко (Scherrer, Rabus 2019). Часто мож тото розвязати транскрибуючы бесіду, як в примірі корпусу ArchiMob для німецкых диялектів або хоснуючы тлумачыня, як в приміри корпусу MADAR для арабскых диялектів (Bouamor і ін. 2019). Ахім Rabus і Ів Scherrer описують вызваня звязаны з твориньом ресурсів NLP для русиньского языка і пропонують індукцию морфосинтактычного лексикону зо схоснуваньом ресурсів славяньскых языків (Rabus, Scherrer 2017). Досліджают они морфосинтактычне етыкєтуваня (анґ. tagging) для русиньского языка, хоснуючы трансфер вчыня зо спорідненых языків. Вказаны выбраны однесіня сут доказом того, што тема є інтенсивні досліджана і експлоатувана. Істніют специялістычны проґрамы і аплікациі до досліджаня языковых примет, але схоснуваня моделю так барз зложеного і чутливого як в тым дописі єст інновацийным підходом.

Цілю працы є дослідити ступін розпознавальности і клясифікациі русиньского языка через модель автоматычного розпознаваня бесіды Open IA Whisper і оцінити влияня домінуючых языків держав, де жытют Русины на результат клясифікациі. Обсяг працы обнимат аналізу звуковых даных зобраных з русиньскоязычных радийовых авдиций зо штырьох держав: Польщы, Украіны, Словациі і Сербіі, а тіж другы (додатковы) аналізы з огляду на віковы групы бесідників. В статі поміщено основны інформациі на тему моделю OpenIA Whisper, в тым його архітектуру і приметы, што чынят можливом вельоязычну транскрипцию і дрібницьовый опис дослідничой методолоґіі, што обнимат зміст базы звуковых даных, альґоритмів сеґментациі і аналізы звуку та параметрів схоснуваных в модели.

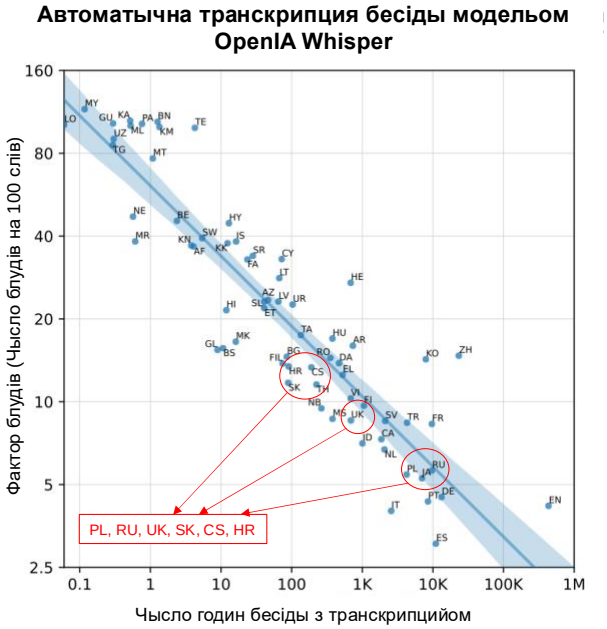
Як результаты представлено клясифікацию русиньского языка при увзглядніню держав, одкале походят бесідникы і ріжниц медже віковыма групами. Результаты омовлено в контексті влияня домінуючых языків на язык Русинів і проходячых асиміляцийных процесів. Представлено тіж квестиі звязаны з ограничынями моделю і специфікы языковой клясифікациі. Працу завершат підсумуваня, котре вказує на основны высновкы і можливы напрямы дальшых досліджынь в обшыри автоматычного розпознаваня бесіды для меншыньовых языків.

2. Языковый модель OpenIA Whisper

OpenIA Whisper то заавансуваный модель автоматычного розпознаваня бесіды ASR (анґ. Automatic Speech Recognition) (Radford і ін. 2022). Модель і проґрамовый код сут одкрыто доступны на ліценциі тыпу *open source*, што уможливят динамічне розвитя аплікациі і повязаны досліджыня в обшыри заавансуваной переробкы бесіды. Сесый модель вытренувано на основі веце як 680 тисячи годин вельоязычных награń з транскрипцийом. Обшырна і ріжнорідна база даных дозволила осягнути велику одпорніст системы на зріжницюваны акценты, галас на фоні ци специялістычну термінолоґію, спомогла при тым транскрипцию в вельох языках.

Архітектура OpenIA Whisper єст приміром реализациі системы ASR зо схоснуваньом невроновой сіти тыпу *Transformer* в конфіґурациі енкодер-декодер (анґ. encoder-decoder). Звуковы даны на вході моделю ділены сут на 30-секундовы часткы, пак переформлюваны на *log-Mel spectogram* (є то діґітальна, ґрафічна репрезентация звуку, яка бере до увагы приметы перцепциі чловека), а дале перерабляны через енкодер. Декодер передвидує зміст, што одповідат звукам і вводит додатковы інформацийны токены, такы як ідентыфікация языка бесідника і часовы значникы з докладністю до рівня фраз бесіды.

Во вчыню моделю OpenIA Whisper схоснувано барз велику і зріжницювану базу даных, што ілюструє Рыс. 1, на котрым вказано залежніст процента невластиві розпознаных слів од кількости даных до тренуваня. В базі была менше-веце третя част награń в інчым як анґлицкій языку. І хоц в базі єст о вельо менше тестовых награń в інчых языках, то ефектывніст розпознаваня языка, при базі о велькости більше як сто годин, єст на рівни 80 проц. і высшым. До того зараховуют ся вшыткы значучы з перспектывы досліджынь славяньскы языкы, што тіж зазначено на Рыс. 1.



Рыс. 1. Залежніст ефектывности моделю ASR, в залежности од велькости даных до тренуваня (Radford і ін. 2022). На выкресі зазначено выбраны славяньскы языкы (польскій, украіньскій, словацкій і росийскій).

Модель OpenIA Whisper розпознає язык выповідичерез вельоетаповый процес, котрый єст зінтеґруваный з його архітектуром енкодер-декодер. Лоґарытмічны Мелы спектроґрамы сут передаваны до енкодера, што выокремлят зложены звуковы приметы. Енкодер выокремлят взірці в звуковых даных на барз высокым рівни зложености, якы характеризуют ріжны языкы. Проґноза основана єст на высокоплощыновых приметах, што обнимяют фонетычны і фонолоґічны взірці особливы для каждого языка. В час вчыня модель был експонуваный на звуковы даны в барз вельох языках, вєдно з одповідніма для них текстовима етикєтами, в котрых сут токены до ідентыфікациі языка. В процесі декодуваня звуку декодер хоснує сесы вывчены токены, жебы окрислити язык на основі введеных даных/звуків. Завдякы так обшырному вчыню модель аналізує сиґналы в комплексовый спосіб в ріжных языках і акцентах, што поправлят його здібніст до прецизийного розпознаваня языка. OpenIA Whisper хоснує тіж здібности *zero-shot learning*, што означат, же на основі даных з вчыня може робити узагальніня, жебы розпознати і транскрибувати языкы, на якых не был конкретні вченый. Осігат тото хоснуючы вельоязычный характер даных до тренуваня і заавансуваны можливости выокремляня примет через енкодер.

Схоснуваня моделю OpenIA Whisper до клясифікациі русиньского языка, хоц модель тот не был вченый на русиньскій базі, мож пояснити парома чынниками. Модель OpenIA Whisper был вытренуваный на великій, вельоязычній базі даных, в якій были ріжны славяньскы языкы (такы як польскій, украіньскій, словацкій, сербскій, хорватскій ци росийскій), што сут зближены до русиньского языка під оглядом ґраматычной структуры, фонетыкы ци лексикы. Завдякы тому модель годен розпознати деякы взірці і приметы характерны для ґрупы славяньскых языків, што уможливят му зближену клясифікацию русиньского языка, хоц не было безпосереднього вчыня на його базі. В примірі русиньского языка, котрый не єст релятывні масовый, схоснуваня моделю вытренуваного на шырокым спектрі языків дозволят на його приближену клясифікацию на основі примет спільных з інчыма языками.

3. Методологія

a. База даных

До языковой аналізы схоснувано архівальны награня лемківской радийовой стациі Lem.fm. В базі звуковых даных є барз вельо зріжницюваных примірів русиньского бесідуваного языка, што уможливило комплексове досліджыня з шырокой перспектывы. З каждой авдициі усунено фраґменты музыкы, джінглі і реклямовы споты, а з інтервю лишено голос лем єдной особы.

Так приготовлена база складала ся з 470 ріжных авдиций. Час тырваня цілой базы то даде 121,8 годин. Поділ походжыня бесідників з кількістю авдиций і часом тырваня досліджаных пробок представлено в Табели 1.

Скорочыня	**Держава**	**Кількіст авдиций**	**Час тырваня [годин]**	**Чысло лекторів**
PL	Польща	209	45,4	веце як 100
UK	Украіна	99	37,5	веце як 40
SK	Словация	131	23,2	веце як 50
CS	Сербія	31	11,5	10

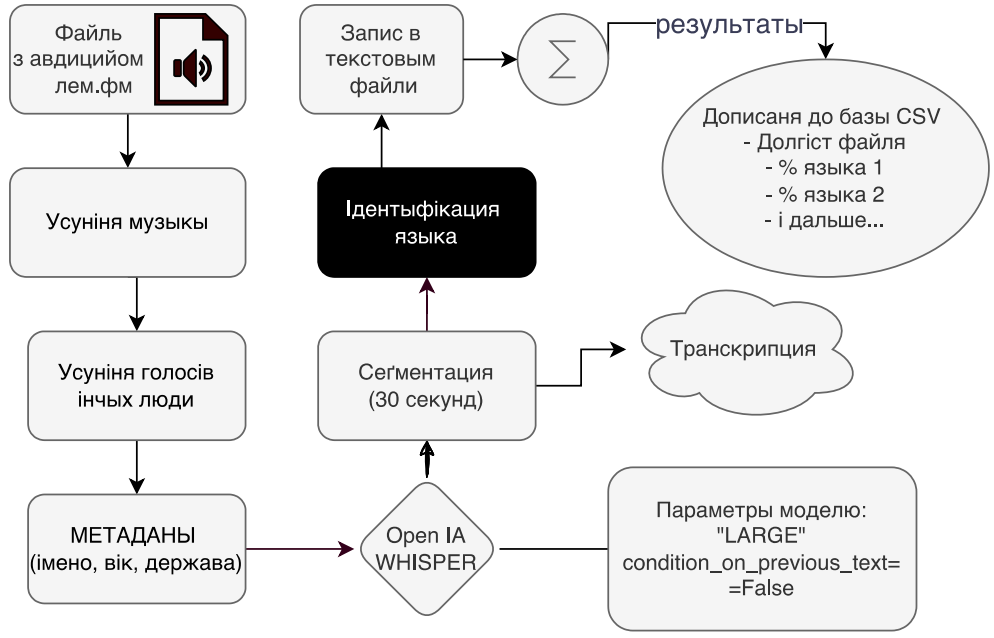
Таб. 1.

Додатковы метаданы приписаны до каждой авдициі:

- Вікова група: дві категориі – бесідникы што мают веце або менше як 70 років (лем для бесідників з Польщы).
- Ідентыфікация бесідника: мено або псевдонім.
- Назва файлів выгенеруваных в час аналізы.
- Полный час награня (в секундах) каждой звуковой пробкы.

б. Розпознаваня языка бесідника зо схоснуваньом моделю OpenIA Whisper

До аналізы і категоризациі даных опрацувано скрипт в языку *Python*, котрый уможливят автоматычный выбір і переробку звуковых файлів, языкову аналізу і регістрацию результатів.



Рыс. 2. Схема процесу ідентыфікациі языка радийовой авдициі.

Діяня скрипту схемово представлено на Рыс. 2, а поєдны етапы його діяня сут слідуючы:

- выбір каталоґу: выбір каталоґу з авдіо файлями при схоснувано ґрафічного інтерфейсу (GUI),
- сеґментация звуку: каждый авдіо файль єст діленый на часткы долгости 30 секунд,
- діяня моделю: до транскрипциі бесіды і вызначыня языка схоснувано модель Whisper в вариянті «*large*». Архітектура моделю запроєктувана єст з 32 верств сіти шырокости 1280 невронів і 20 голов (анґ. heads), што разом дає 1,55 милиярда параметрів. Єст то найвекшый модель в родині Whisper, што перекладат ся на высоку прецизию транскрипциі і розпознаваня языка завдякы його барз зложеній структурі. Модель робит транскрипцию бесіды для каждого сеґменту авдіо, а на выході моделю сут додатковы інформациі, в тым вызначеный язык. Параметры моделю сконфіґуруваны сут на вартости: *no_speech_threshold=0.8* і *condition_on_previous_text=False*, завдякы чому кажда наступна 30-секундова пробка аналізувана єст незалежні од попередньой. Шырокост сіти дефініює чысло невронів в каждій верстві моделю, значыт розмір вектора репрезентациі, што є перетворювана через модель на даній верстві, а чысло голов односит ся до чысла «голов увагы» (анґ. attention heads) в каждій верстві механізму увагы моделю. Тот механізм дозвалят зосередити ся модельови на ріжных частях секвенциі на вході в час переробкы. Кажда головка аналізує даны під інчым кутом, завдякы чому модель може ліпше выімати зложены залежности в языковых секвенциях,
- реґіструваня результатів: результаты транскрипциі і вызначеный язык для каждого сеґменту сут записуваны запоряд до текстовых файлів,
- експорт до CSV: результаты сут записуваны в файли CSV, в котрым сут первістны метаданы і інформациі о розпознаных языках і іх процентова участ в полным часі награня.

Скрипт додатково хоснує бібліотеку *librosa* до ладуваня звуку і *pydub* до сеґментациі даных.

4. Результаты

Даны переаналізувано в тот спосіб, же в каждій авдициі окреслено шор розпознаных языків, зачынаючы од найчастійше розпознаваного (#1), вєдно з ідентыфікатором державы бесідника. Пак порахувано процентову участ каждого з языків в роли языка #1,

#2 ітд. Результаты ґрафічні передставлено на выкресах (Рыс. 3а-г), вказуючы найчастійше розпознаваны языкы в бесіді Русинів, што жыют в Польщы, Словациі, Україні і Сербіі.

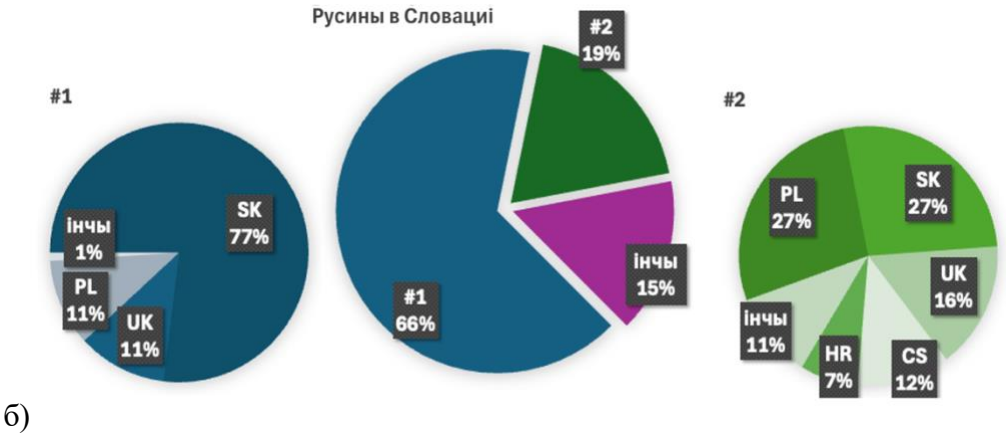
В аналізуваных авдициях з Польщы найчастійше розпознаваный язык (#1) становил 84 проц. полной долгости авдициі, а з того аж 92 проц. становил польскій язык, а 7 проц. украіньскій язык. Другій найчастійше розпознаваный язык (#2) становил 12 проц., з чого 63 проц. приписано украіньскому языкови, а 21 проц. польскому (Рыс. 3а).

В авдициях бесідників з Украіны (Рыс 3б), язык розпознаваный як першый (#1) становил 91 проц. цілой базы, з чого аж 99 проц. тых сеґментів приписано украіньскому языкови .

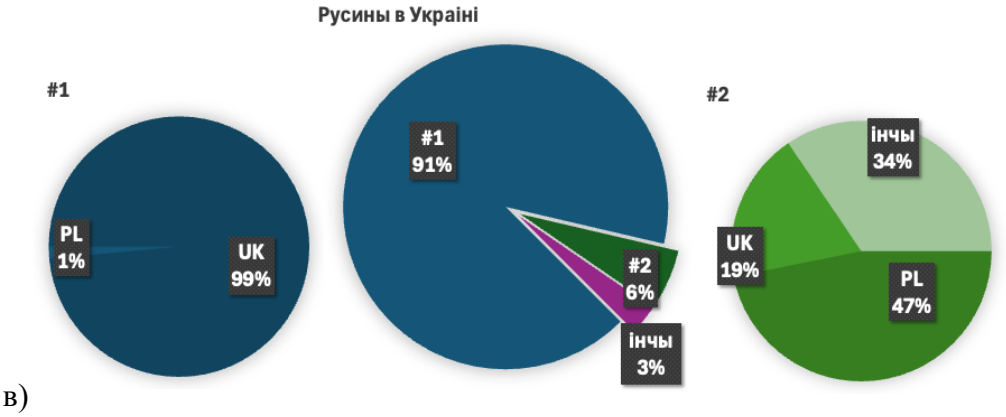
Для Русинів на Словациі найчастійше розпознацаный язык становил 66 проц., з чого 77 проц. приписано словацкому языковы. Другій найчастійше розпознаваный язык (#2) становил 19 проц., де польскій і словацкій становили 27 проц., а украіньскій язык 16 проц. (Рыс. 3в).

В приміри авдиций з Сербіі (Рыс. 3г), першый найчастійше розпознаваный язык (#1) становил 71 проц. награнь, з чого 87 проц. тых авдиций приписано хорватскому языкови. Другій найчастійше розпознаваный язык (#2) становил 22 проц. часу, з чого 78 проц. становил словеньскій язык.

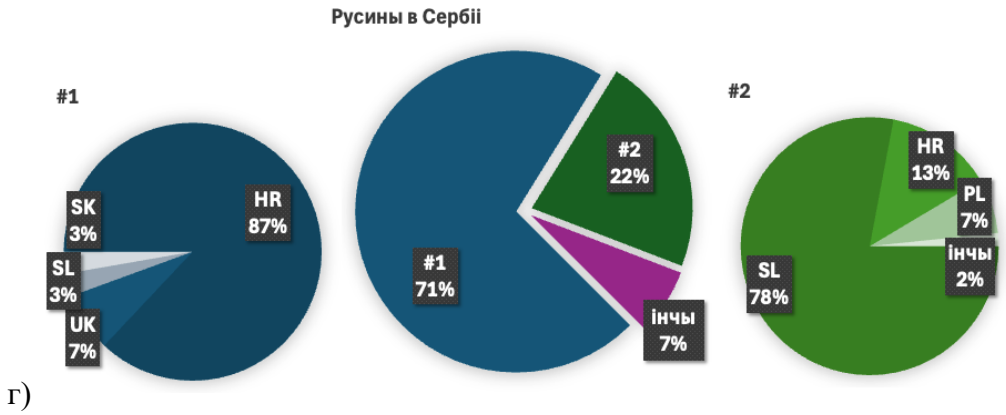
а)



б)



в)



г)

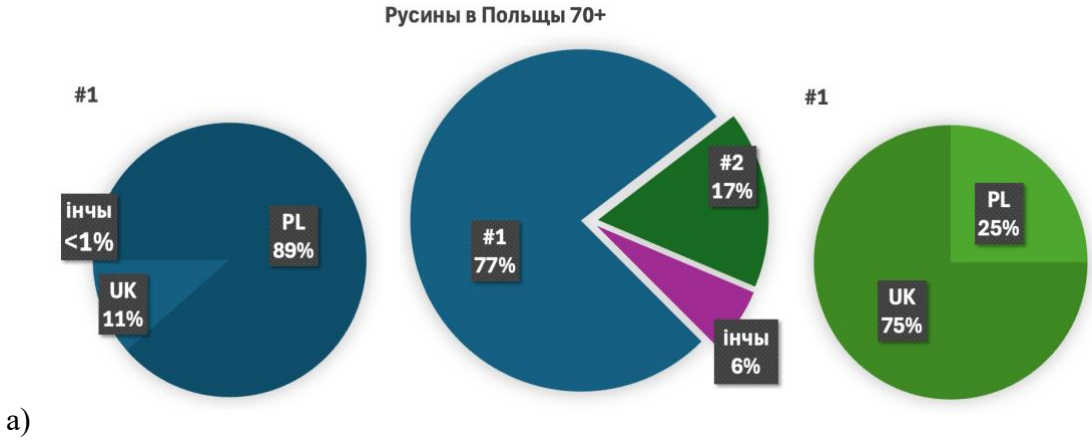
Рыс. 3 а–д. Процентове сопоставліня розпознаных языків Русинів з ріжных держав.

Додатковы аналізы зроблено для Русинів з Польщы. Выкресы (Рыс. 4 а, б) передставляют аналізу розпознаных авдиций Русинів з Польщы, котрых поділено на віковы групы: особы старшы як 70 років «70+» і особы молодшы як 70 років «70-». Треба додати, што база награń лекторів «70+» становит понад 60% лекторів з Польщы.

Ограничыня досліджыня з поділом на віковы групы выключні для лекторів з Польщы, выникат з двох причын. По перше, кількіст дальшых груп была за мала, жебы отримати одповідньо великы пробы. По другє, проблемом было рішыти вік лекторів з інчых держав, бо част редакторів, што робили награнá, не робит уж в радию.

В групі «70+» (Рыс. 4а) найчастійше розпознаваным языком (#1) был польскій, котрый становил 89 проц. авдиций, обнимаючы 77 проц. цілковитого аналізуваного часу. Як другій найчастійше розпознаваный язык (#2) выкрыто україньскій (75 проц. в тій катеґориі), што становит 17 проц. вшыткых аналізуваных авдиций. Сут гев тіж присутны інчы языкы, якы появляют ся в шестьох процентах розпознань.

В молодшій віковій групі «70-» найчастійше розпознаваным языком (#1) был польксій, обнимаючы веце як 99 проц. часу в тій катеґориі, з 98 проц. вшыткых аналізуваных авдиций (Рыс. 4б). Інчы языкы незначні фіґуруют в тій групі.



а)

б)

Рыс. 4 а, б. Процентове сопоставліня розпознаных языків Русинів з Польщы з поділом на віковы групы (веце або менше як 70 років).

5. Дискусия

На основі осягненых результатів явні видно влияня домінуючого языка (урядового) даной державы на клясифікацию русиньского языка через языковый модель. З вынятком

Русинів з Войводіны, бесіда Русинів з Украіны, Польщы і Словациі была клясифікувана через альґоритм як найбарже зближена до домінуючого в даній державі языка. Хоц флексийны, лексикальны і синтактычны ріжниці медже вариянтами русиньского языка сут з перспектывы його хосновательи або языкознавців рішучо меншы як ріжниці медже русиньскым а домінуючым языком даной державы, альґорытм, што комплексово аналізує языкову подібніст, выказує значучо векше зближыня до домінуючых языків. Особливо україньскій язык розпознаваный єст релятывні часто серед Русинів з Польщы і Словациі, але в Сербіі є інакше. В примірі Русинів з Войводіны найвекшу подібніст выказують хорватскій і словеньскій, при незначній участи сербского языка. Треба єднак зазначыти, же база пробок з Сербіі, на якій оперто аналізу, была найменше чысленна з увагы на ограниченый доступ до материялів.

Не сут знаны докладны механізмы і критериі, на основі якых модель рішат о клясифікациі языка. Діяня сіти тыпу *Transformer* єст барз зложеным процесом статистычной аналізы, опертым на ідентыфікациі взірців в даных языках, што дозволят вызначати подібніст без фактычного розумліня змісту.

Традицийны моделі ASR, основаны на вкрытых моделях Маркова (HMM, анґ. Hidden Markov Models), характеризували ся деяком міром детермінізму, што злегшало аналізу влияня акустычных примет на результат транскрипциі/клясифікациі языка. Сучасны моделі ASR, основаны на глубокых невроновых сітях, в тым рекуренцийных невроновых сітях (RRN, анґ. Recurrent Neural Networks), конволюцийных невроновых сітях (CNN, анґ. Convolutional Neural Networks), і архітектурі тыпу *Transformer*, сут неявны і недетерміністычны. Означат то, же тоты самы входовы даны (сиґнал бесіды) можут в ріжных примірах вести до кус інчых транскрипций. Тота недетерміністычніст выникат з великой кількосты параметрів моделю і зложеных, неленійовых реляций медже нима. Утруднят то безпосередню аналізу і выокремліня особливых примет акустычного сиґналу, якы рішают о конкретным результаті транскрипциі. Інакше бесідуючы, тяжко є докладні окреслити, котры фраґменты сиґналу бесіды і якым способом мали влияня на клясифікациі поєдных фонемів ци слів. Істніют єднак методы, што дозваляют на досліджаня і інтерпретатицю моделів ASR, хоц сут зложены і вымагают заавансуваного інформатычного знаня. Належат до них м.ін. (Rahate і ін. 2022):

- аналіза верств увагы (attention mechanism) дозволят візуалізувати звязок медже фраґментами сиґналу бесіды а елементами транскрипциі, вказуючы, на якы части звуку модель «зверат увагу» в час розпознаваня,

- методы основаны на ґрадієнтовым пошырюваню, дозваляют ідентыфікувати фраґменты сиґналу на вході, якы мают найвекше влияня на актывацию окремых невронів і в тот спосіб на кінцьовый результат транскрипциі,
- методы пертурбациі входу основаны сут на тым, же вводит ся невеликы зміны в сиґналі бесіды і обсервує, як влияют на результат транскрипциі, што дозвалят ідентыфікувати основны акустычны приметы.

Результаты вказуют праві зерове вызначаня росийского языка, што дакус заскакує, беручы до увагы його сутьове влияня і хоснуваня в Украіні, особливо на сході державы, одкале походила част бесідників.

Аналіза лемківского языка выказала тіж, што результаты ріжнят ся в залежности од віку бесідників. В звязку з тым проведено другу аналізу. Поділено пробкы на дві віковы групы: бесідників маючых менше і веце як 70 років. Подібніст лемківского языка до польского серед старшых осіб была значучо менша як серед молодшых осіб, што суґерує поступуючу языкову асиміляцию, особливо в обшыри фонетыкы. В лексикальным аспекті старшы бесідникы хоснували подібне, аж і векше чысло польонізмів або запожычынь, што може вказувати на зложены механізмы языковой асиміляциі медже поколінями.

6. Підсумуваня

В статі проанаіизувано здібніст моделю OpenIA Whisper до розпознаваня і клясифікаці русиньского языка, хоц модель не был безпосередньо вченый на русиньскых материялах. Проведено досліджыня на основі радийовых авдиций Русинів з Польщы, Украіны, Словациі і Сербіі, при увзглядніню ріжниц медже віковыма групами. Результаты всказуют, же модель OpenIA Whisper клясифікує русиньскій язык головні як урядовый язык даной державы, што насуват высновок о сутьовым влияню домінуючых языків на автоматычну клясифікацию меншынового языка. Достережено тіж значучу языкову асиміляцию в молодшій віковій групі, што свідчыт о поступуючым влияню польского языка в Польщы і інчых урядовых языків в інчых державах.

В будучых досліджынях плянує ся змодифікувати модель OpenIA Whisper так, жебы выеліминувати спомаганя для домінуючого языка в механізмі розпознаваня, што дозволит ліпше зрозуміти подібніст русинського языка до інчых славяньскых языків без сильного влияня офіцийного языка даной державы. Додатково плянує ся проаналізувати русиньскы пісні і співанкы а тіж інчы славяньскы пісні. В співі выступуют інчы

механізмы емісиі звуку як в бесіді, што може вказати шыршый контекст. Сесы досліджыня можут причынити ся до барже прецизийной і актуальной клясифікациі меншыновых языків і ліпшого зрозумліня механізмів асиміляциі і языковой ґлобалізациі.

Вшыткы первістны даны і дрібницьовы результаты вызначены проведеным експериментом сут доступны для заінтересуваных через майльовый контакт з автором.

Бібліоґрафія

Bouamor, Houda, Hassan, Sabit, Habash, Nizar. 2019. «The MADAR Shared Task on Arabic Fine-Grained Dialect Identification». В: *Proceedings of the Fourth Arabic Natural Language Processing Workshop.* Ред. Wassim El-Hajj, Lamia Hadrich Belguith, Fethi Bougares, Walid Magdy, Imed Zitouni, Nadi Tomeh, Mahmoud El-Haj, Wajdi Zaghouani, 199–207. Florence: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4622.

Kushko, Nadiya. 2007. «Literary Standards of the Rusyn Language: The Historical Context and Contemporary Situation». *The Slavic and East European Journal* 51, ч. 1: 111–132.

Moser, Michael. 2016. «Rusyn: A New-Old Language In-between Nations and States». В: *The Palgrave Handbook of Slavic Languages, Identities and Borders*. Ред. Tomasz Kamusella, Motoki Nomachi, Catherine Gibson, 124–139. London: Palgrave Macmillan. https://doi.org/10.1007/978-1-137-34839-5_7.

Nikitin, Alexey G., Kochkin, Igor T., June, Cynthia M., Willis, Catherine M., Mcbain, Ian, Videiko, Mykhailo Y. 2009. «Mitochondrial DNA Sequence Variation in the Boyko, Hutsul, and Lemko Populations of the Carpathian Highlands». *Human Biology* 81, ч. 1: 43–58. https://doi.org/10.3378/027.081.0104.

Plišková, Anna. 2008. «Practical Spheres of the Rusyn Language in Slovakia». *Studia Slavica Academiae Scientiarum Hungaricae* 53, ч. 1: 95–115. https://doi.org/10.1556/SSlav.53.2008.1.6.

Rabus, Achim, Scherrer, Yves. 2017. «Lexicon Induction for Spoken Rusyn – Challenges and Results». В: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Ред. Tomaž Erjavec, Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef

Steinberger, Roman Yangarber, 27–32. Valencia: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1405.

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, Sutskever, Ilya. 2023. «Robust Speech Recognition via Large-Scale Weak Supervision». В: *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. Ред. Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, Jonathan, 1–28 (28492–28518). Honolulu: JMLR.org.

Rahate, Anil, Walambe, Rahee, Ramanna, Sheela, Kotecha, Ketan. 2022. «Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions». *Information Fusion* 81: 203–239. https://doi.org/10.1016/j.inffus.2021.12.003.

Scherrer, Yves, Rabus, Achim. 2019. «Neural Morphosyntactic Tagging for Rusyn». *Natural Language Engineering* 25, ч. 5: 633–650. https://doi.org/10.1017/S1351324919000287.

Zampieri, Marcos, Nakov, Preslav, Scherrer, Yves. 2020. «Natural Language Processing for Similar Languages, Varieties, and Dialects: A Survey». *Natural Language Engineering* 26, ч. 6: 595–612. https://doi.org/10.1017/S1351324920000492.