Automatically Distinguishing between Written Output Produced by Heritage and Non-Heritage Learners of Polish as a Foreign Language

Simon Zuberek

Abstract

This project investigates written output produced by learners of Polish as a foreign language. Written compositions by college-level heritage and non-heritage students are analyzed through the lens of Polish nominal morphology. Three case error types commonly committed by heritage learners are isolated as features. Augmented by per-character entropy values, these features are employed to develop a supervised machine learning classifier to automatically differentiate between written output produced by the two learner groups. The obtained results suggest that the four feature sets are effective in distinguishing between heritage and non-heritage written production. They also indicate that some features are more effective in facilitating that task than others.

1 Introduction

In recent years the foreign language classroom has been infiltrated by data-driven solutions ranging from intelligent tutoring to automatic grammatical error detection and correction. Even though many of these solutions claim to be language-agnostic, in practice they tend to be available for languages that are already well-resourced and commonly taught (e.g. Spanish, French, German, etc.), with their less-commonly taught counterparts largely overlooked. This project attempts to narrow this innovation gap in regards to Polish. Available at a handful of institutions in the U.S. and Canada, Polish is considered a less-commonly taught language (LCTL). The majority of Polish language programs are managed by a single language instructor, offer no more than a couple of sections per academic year, and feature comparatively low enrollments. Similarly to other LCTLs, Polish classes are largely attended by heritage learners, students who "have been raised with a strong cultural connection to a particular language through family interaction" (Van Deusen-Scholl, 2003). The presence of such students poses a unique set of challenges for curricular design. In order to develop effective teaching materials instructors need to know the aspects of the target language their heritage students are likely to struggle with, and assign those aspects appropriate curricular emphasis. The goal of this project is therefore twofold. On the one hand, it hopes to assist instructors in differentiating between heritage students and their non-heritage classmates. On the other hand, it aims to help instructors isolate the features that are challenging for heritage learners and rank them in the order of salience, such that the features that are most significant in distinguishing between the two groups of learners may be given priority over their less significant counterparts.

2 Research Questions

Studies have shown that there are marked differences between language output produced by the heritage and non-heritage learners of Polish. While a number of these differences relate to the learners' dominant language (e.g. heritage learners of Polish in Czech Republic are likely to make different mistakes than their counterparts in Austria), there are also features that characterize Polish heritage output across the board. In particular, the complex terrain of Polish nominal morphology seems to facilitate innovative case use exclusive to heritage language production. These error types are generally absent from non-heritage output, and observed in learners across different socio-linguistic contexts such as the U.S., Germany, and Sweden (Laskowski, 2014, p. 194). The more salient of these error types are enumerated by Wolski-Moskoff (2019) and summarized below:

- While the majority of Polish prepositions take a single case, a handful of them can take two or more cases. There is a group of prepositions that will take either the locative or the accusative case, depending on whether they appear in a static or non-static contexts. That nuance is largely lost to heritage learners, who tend to overgeneralize the locative case use in contexts that are non-static and therefore requiring an accusative object.

- Somewhat exceptionally, a handful of Polish verbs require that the direct object always be in the genitive case. Heritage learners tend to glance over that rule, applying the accusative wherever the genitive is required.

- The majority of direct objects in affirmative sentences are expressed in the accusative case. Verb negation requires that they be rendered in the genitive case. Heritage learners tend to ignore that rule, retaining the accusative case in negated declarative constructions.

Based on the above observations this project seeks to answer the following two research questions:

- Are the above features truly symptomatic of heritage learner output, and if so, are they sufficient to distinguish it from non-heritage production?

- Are the listed features equally important in making the above distinction, and if not, how effective is each feature in correctly classifying heritage learner output?

3 Previous Research

The project is motivated by Wolski-Moskoff's (2019) study of case use in heritage Polish. The dissertation offers a thorough examination of Polish nominal morphology in the heritage learner output. It catalogs the nominal morphological features of both heritage and non-heritage learner production and juxtaposes them against those characterizing native speakers. Of particular value for this project is the author's distillation of the nominal morphological features that are exclusive to heritage learners. The fact that these features are confirmed for the learners across a spectrum of dominant languages makes them particularly useful for the task at hand.

If Wolski-Moskoff (2019) provides a clear description of error categories, Koppel et al. (2005, August) explore how error types can be used to automatically classify written output. Their study provides useful suggestions on developing a machine learning (ML) classifier to automatically determine the dominant language of the author behind an English text. The work is instructive not only when it comes to feature extraction but also in combining and analyzing feature sets to maximize classification accuracy. Even though it explores errors made in English texts - whereby it does not engage in nominal morphology - its insights provide useful cues for the current project.

The work that is perhaps most relevant for this project is a thesis by Hoyos (2021), which specifically deals with the morphology of the Polish genitive case. The author focuses on the genitive masculine inanimate singular and develops a mechanism for its assessment in written texts. As the genitive case is notoriously challenging to master for the learners of Polish, Hoyos offers a solution that automatically detects its incorrect use and provides grammatical feedback to the learner. Whereas the current project lacks a learner feedback component, Hoyos' insights pertaining to automatic case detection provided useful cues for extracting the morphological features of heritage learner output. To that end, his implementation of the SpaCy Polish Language pipeline[1] for parsing and tagging was especially valuable.

[1] https://spacy.io/models/pl

4 Method

The solution developed to answer the research questions is a supervised ML classifier, available at `classify_final.py`. Based on the analysis provided by Wolski-Moskoff (2019), the program extracts three morphological feature sets, each of which represents a nominal error category characteristic of heritage learner output.

Starting with post-prepositional features, `LOC_post_prep()` scans the document in search of common prepositions taking the locative case in static and the accusative case in dynamic contexts. The list of these bivalent prepositions was sourced from Sadowska (2012). The function then counts the instances of post-prepositional locative use and produces a feature count vector thus capturing the intuition that heritage speakers are more likely to overgeneralize the post-prepositional application of the locative case in non-static contexts. Consequently, the counts of post-prepositional locative instances are expected to be significantly higher in the documents produced by heritage learners.

The second feature set is extracted with `genitive()`. The function iterates over the sentences in each document, looking for the lemmas of the verbs requiring genitive direct objects. A list of verb lemmas taking the genitive case has been adapted from Mędak (2011). It includes the infinitive forms of more commonly used Polish verbs, along with their lemmas, as specified by the SpaCy lemmatizer[2]. Whenever a genitive collocating verb is detected, the function scans for its nominal object string and checks its case. The correct usage of the genitive case is subsequently quantified and vectorized into a feature set. Here the expectation is that heritage output should show much fewer instances of the post-verbal genitive use than its non-heritage counterpart.

The third and final morphological feature set reflects the use of the genitive case in declarative negated constructions. `negation()` first scans each sentence for negated verbs. Upon successfully isolating a negated verbal construction, it iterates over the remainder of the sentence string in search of verbal objects. Once an object is found, the function looks up its case, only increasing the counter if the object is rendered in the genitive. The frequencies of genitive use for each document are then aggregated into a count vector, which reflects the assumption that documents produced by heritage learners feature significantly fewer instances of genitive objects post-negation, than their non-heritage counterparts.

The fourth extracted feature set contains the average per-character entropy for each document. The computation is done based on a WFST language model of the corpus of the heritage train data assembled in `heritage_train.txt`. The model was developed using OpenGrm-NGram[3] command-line tools and the Command Line script for computing it is included in `lm.sh`. The resulting language model `lm.fst` is a 6-gram, character-based representation with Witten-Bell smoothing. The choice of the 6-gram is motivated by the fact that the average Polish word is six characters long (Moździerz, 2020).

The use of the average per-character entropy values as a feature set is suggested by the observation that heritage language learners are more likely to employ non-standard vocabulary in circulation in their communities. These lexical items are absent from the standard vocabulary, and come into the dialect as borrowings from the host language (e.g. in Polish-speaking communities in North America “insurance” is expressed as “insiura”, as opposed to the standard “ubezpieczenie”). For an extended discussion on the lexical features of the Polish dialect spoken in the immigrant communities in North America turn to Moskoff-Wolski (2019). The per-character entropy values for such non-standard borrowings are guaranteed to diverge from the values typical for standard Polish, and may therefore help differentiate between the two categories of written output. The entropy feature set is extracted with the `entropy()` and `bits_per_char()` functions, based on the language model `lm.fst`.

Each of the four feature sets was extracted from the train data and used to train the classifier. The following five classifying algorithms were implemented: Multinomial Naive Bayes, Complement Naive Bayes, Decision Trees, Random Forest, and Support Vector Classifier. Their performance was first evaluated on a test data set, where the classification

[2] https://spacy.io/usage/linguistic-features

[3] https://www.opengrm.org/twiki/bin/view/GRM/NGramLibrary

accuracy was computed for each feature set individually, as well as for all four sets combined. Subsequently, the same classification algorithms were evaluated for accuracy on a 10-fold cross-validated train set, first for each feature set in isolation, followed by all feature sets combined. Thus obtained accuracy scores were also compared with the baseline.

5 Baseline

The written samples produced by heritage learners ought to be meaningfully different from those produced by their non-heritage classmates. In order to be able to assess this difference a metric is required to score each written sample. The metric initially selected for the baseline is a model of the language used by heritage learners, computed with OpenGrm-NGram command-line tools, and based on a corpus of heritage output `concatenated_baseline.txt`. The computed language model `baseline_lm.fst` is subsequently passed into `score.py` to evaluate students' written compositions. The script iterates over the heritage and non-heritage test output contained in `heritage_baseline.txt` and `non_heritage_baseline.txt` respectively whereby it scores each sentence against the language model. The scores appended in front of each line express the average values of bits per character for that line, and can be previewed in `heritage_baseline_scored.txt` and `non_heritage_baseline_scored.txt` for heritage and non-heritage test data respectively. The script `score.py` also generated summary statistics files `heritage_baseline_scores.txt` and `non_heritage_baseline_scores.txt` listing the relevant metrics for each population, including a list of per-character entropy scores, sample size, arithmetic mean, and variance. The metrics are displayed in the Table 1.

	Size in Sentences	**Mean Score**	**Score Variance**
Heritage	96	2.363	0.588
Non-heritage	170	2.505	0.707

Table 1: Summary statistics for the heritage and non-heritage baselines.

The lower the number of bits per character, the more likely our language model judges the string to be. The fact that the mean score for heritage test samples is lower than for non-heritage test samples indicates that the written output produced by heritage speakers is more similar to the heritage training data than the written samples produced by non-heritage learners – an encouraging result.

The final step in developing the baseline is to ascertain if the written samples produced by the two populations were meaningfully different. Let the null hypothesis stipulate that the two collections of written samples were produced by the same population and are therefore not statistically different. The difference in sample sizes and variances requires that the independent t-test be used to either prove or disprove the null. The script for the t-test file is included in `ttest.py`. The test statistic t = -1.364 validates the lower scores obtained by the non-heritage speakers, however at the p-value of $p = 0.174$, where $p > 0.05$, that difference is statistically insignificant. Therefore we cannot reject the null hypothesis that the heritage written samples significantly differ from non-heritage samples as per their per-character entropy scores.

The above result may indicate that a 6-gram character-based language model is perhaps not the best baseline to compare the two datasets. As an alternative baseline, `classify_final.py` includes a dummy classifier imported from the `sklearn` package[4]. Considering that there are two labels (one for heritage and one for non-heritage samples) and that each label applies to the same number of test samples (nine heritage and nine non-heritage essays), the dummy baseline accuracy should be equivalent to a coin toss, or 0.5. Running the dummy classifier confirms that expectation by yielding the accuracy score of 0.5, which will serve as the baseline for further evaluation.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

6 Data

The data used to develop this project comes from a variety of sources. For the purpose of training the classifier, heritage and non-heritage output was required. The majority of the heritage output was sourced from the corpus of Heritage Language Variation and Change (HLVC)[5] (Nagy, 2011) developed and maintained at the University of Toronto. The HLVC corpus includes a collection of recordings of interviews conducted with speakers of Polish, as well as and their transcripts. All interview records are divided into conversations conducted with heritage speakers and those administered to native speakers. For the purpose of this project only heritage transcripts were extracted from the corpus. Upon extraction, the records were processed with the ELAN [6] platform to isolate only the parts produced by the heritage interviewees and leave out all non-heritage output. The interviewee contributions were saved to 35 text files. Each file was first manually spell checked and later normalized with `normalize.py`. The files are further processed with `normalizeML.py`.

The above processing steps were also applied to a small collection of six transcripts of interviews conducted with heritage speakers of Polish residing in the greater Chicago area.

The resulting collection of combined documents was concatenated with `concatenate.py` into one training file `heritage_train.txt`. The file contains 41 lines, corresponding to 41 heritage documents, where each new sentence is indicated by a capital letter and concluded with a period.

The non-heritage training data were sourced from PoLKo, the Polish Learner Corpus [7] (Zasina et al., 2020), which collects writings in Polish as a foreign language at various proficiency levels. A total of 36 documents were extracted from the corpus, normalized with `polko_normalize.py` and concatenated into `nonheritage_train.txt`. Similarly to the heritage training data the resulting file contains 36 lines, corresponding to 36 non-heritage compositions, with sentence boundaries indicated by sentence-initial capitalization and sentence-final punctuation.

Lastly, the data for testing consists of short essays drafted by college-level learners of Polish at the University of Illinois at Chicago. 18 essays were collected in total, nine of them authored by heritage and the other nine by non-heritage learners of Polish. All texts were normalized with `essay_normalize.py`, and concatenated into `heritage_test.txt` and `nonheritage_test.txt` respectively. Each of the resulting files corresponds in its formatting to the training data.

7 Results and Evaluation

The four feature sets extracted from the training files `heritage_train.txt` and `nonheritage_train.txt` were used to train the classifier. The program's performance was measured with the metric of accuracy and first evaluated on the test data sets `heritage_test.txt` and `nonheritage_test.txt`. In an effort to ascertain if some features are better predictors than others, each feature set was tested independently. This was followed by testing all feature sets combined, in order to determine if they are sufficient in differentiating between the two written output categories. Finally, the project employed five different classification algorithms: Multinomial Naive Bayes, Complement Naive Bayes, Decision Trees, Random Forest, and Support Vector Classifier. The test data accuracy values obtained by these five models, across all feature sets are summarized in Table 2.

When run on the test data set, the classifier did not perform much better than the baseline. The accuracy is generally within the coin toss range, which does not provide enough evidence to confirm that the extracted feature sets are enough to differentiate between heritage and non-heritage output. The only exceptions here are the two tree-based classification algorithms: Decision Trees and Random Forest. For the feature set reflecting the use of locative after bivalent prepositions, these two models did worse than chance, yielding accuracy scores of 0.389. This rather poor performance was offset by the feature set

[5] https://ngn.artsci.utoronto.ca/HLVC/0_0_home.php

[6] https://archive.mpi.nl/tla/elan

[7] http://utkl.ff.cuni.cz/teitok/polko/index.php?action=home

Model	LOC after Prepositions	GEN after Verbs	GEN of negation	Per-Character Entropy	All features
Multinomial NB	0.5	0.5	0.5	0.5	0.5
Complement NB	0.5	0.5	0.5	0.5	0.5
Decision Trees	0.389	0.556	0.5	0.5	0.5
Random Forest	0.389	0.556	0.5	0.5	0.5
SVC	0.5	0.5	0.5	0.5	0.5

Table 2: Test data accuracy scores.

capturing the presence of genitive direct objects following certain verbs, where the two models did better than chance, generating 0.556 accuracy. The discrepancy in accuracy scores between the three morphological features suggests that some of these features are better predictors than others, with post-verbal genitive object use carrying the most salience, followed by the presence of the genitive in negated declarative sentences, and capped by the use of the locative case following certain bivalent prepositions. Based on the above results, it seems that when it comes to nominal morphology, it is the use of the genitive case that emerges as the best predictor of heritage background.

In order to further explore this finding, the classifier was also run on the 10-fold, cross-validated training data set. Here, the results were much more promising, as displayed in Table 3.

For individual features only the Naive Bayes models did not exceed the baseline accuracy. This was no longer the case for all features combined, where both Naive Bayes algorithms did better than the baseline, yielding accuracy of 0.668. The SVC classifier further improved on that result, correctly labeling 0.864 of the data. The most impressive performance was again displayed by tree-based models, which correctly classified all the data, given the combined feature matrix. Moreover, the obtained accuracy scores ranging from 0.668 to 1 on the combined features seem to suggest that the extracted feature sets may be sufficient in differentiating between heritage and non-heritage output.

When it comes to ranking individual features, it is only possible to do so in the context of tree-based and SVM classification models. In comparison to the test data, where the features with the most salience were related to the use of the genitive case, the feature yielding highest accuracy on cross-validated train data reflects the use of the locative in prepositional phrases. Even though this finding differs from the outcomes obtained with the test data, it serves to corroborate the claim that different morphological features carry different effectiveness when it comes to classification. However, unequivocally ranking them in the order of importance requires further investigation.

8 Conclusion and Future Work

The current project set out to examine if various case error types were sufficient in distinguishing between written output produced by heritage and non-heritage learners of Polish, and if so, whether some of these error types were better predictors of the author’s heritage background than others. Although the accuracy rates obtained on the test data do not provide a conclusive answer to the first research question, the same classifier run on cross-validated train data was able to generate accuracy not only far exceeding the baseline, but in some cases reaching 100%. Between the two sets of test data, the results cautiously suggest that the discussed case error types may be sufficient to differentiate between heritage and non-heritage learners. The results yielded by the cross-validated train data, and to a lesser extent by the test data, also demonstrate that some error types

Model	LOC after Prepositions	GEN after Verbs	GEN of negation	Per-Character Entropy	All features
Multinomial NB	0.489	0.489	0.489	0.489	0.668
Complement NB	0.5	0.5	0.5	0.5	0.668
Decision Trees	0.853	0.794	0.839	1	1
Random Forest	0.912	0.744	0.	1	1
SVC	0.5	0.5	0.5	1	0.864

Table 3: Cross-validated training data accuracy scores.

are more effective in making that distinction than others. Generally speaking, the higher the accuracy score obtained by an error category, the greater its predictive effectiveness.

This relationship is not without importance for the foreign language classroom. Although instructors tend to have little difficulty with differentiating between their heritage and non-heritage students, what seems to be more challenging is tailoring the curriculum to adequately meet the diverse needs of the two populations. Whereas offering pedagogical advice is outside of the scope of this paper, the project provides an insight into the notoriously difficult topic of nominal morphology. In doing so it confirms Wolski-Moskoff's (2019) finding that certain case errors are characteristic of heritage production, and by quantifying their salience, it suggests to instructors a possible pedagogical sequence that may be used to address them.

Moving forward, the obvious discrepancy in performance between the test data and the cross-validated train data needs to be addressed. Part of the reason for this disparity is the fact that the heritage data used for training was sourced from oral interview transcripts, rather than student written output. As such the heritage training data employed in this project represents a different genre, and features a different sentence structure, and lexical variety. It also required a very different approach to cleaning and normalization than written essays.

The decision to employ this type of data was motivated by difficulties with obtaining student compositions. Despite the author's reaching out to the members of the North American Association of Teachers of Polish, to date only two instructors contributed heritage essays to this project.

Overall, data scarcity proved the biggest challenge when working on this project. The classifier was trained on the collection of 41 heritage interview transcripts and 36 authentic non-heritage essays, for a total of 77 documents. The trained model was then tested on nine heritage and nine non-heritage compositions. In total only 95 documents were used in the development of this project. Obtaining more data should improve the classifier, result in a higher accuracy, and provide a more definitive ranking of feature set salience.

Apart from using exclusively written compositions and obtaining more data, the model could benefit from adding additional features. One of the prime candidates is the presence of lexical items borrowed or adopted from the host idiom. Even though the implemented per-character entropy scoring attempts to account for that feature, a different metric may yet prove more effective. Another heritage error type discussed by Wolski-Moskoff (2019) is the application of redundant prepositions to verbs and participles that do not require them. Determining which prepositions are redundant and extracting their counts is left to future researchers.

Lastly, the project may stand to benefit from implementing different classification models or fine-tuning the ones currently employed. Should such adjustments yield little improvement in accuracy, it may be worthwhile to turn to unsupervised learning. However, for that to be effective, future researchers would need access to the amounts of training data far exceeding what the author was able to secure.

Acknowledgments

I would like to thank Dr. Izolda Wolski-Moskoff of the University of Illinois at Chicago for helping me obtain the test data, as well as a portion of the training data. Her kind assistance in broadcasting the call for data to Polish language instructors in North America was also appreciated.

I would also like to thank Dr. Agnieszka Jeżyk of the University of Toronto for sharing a number of her students' compositions. Unfortunately, due to the timing, I was unable to use them in the current version of the project.

My thanks and gratitude go to Dr. Adrian Jan Zasina of Charles University in Prague for granting me access to the PoLKo corpus, as well as Dr. Naomi Nagy of the University of Toronto for allowing me to access the HLVC corpus.

The *HLVC Corpus* is funded by research grants from the *Social Sciences and Humanities Research Council of Canada* 410- 2009-2330 (2009-2012) and 435-2016-1430 (Naomi Nagy, PI).

References

Hoyos, J. (2021). *PLPrepare: A Grammar Checker for Challenging Cases* (MA Thesis, East Tennessee State University).

Koppel, M., Schler, J., & Zigdon, K. (2005, August). Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 624-628).

Laskowski, R. (2014). *Language Maintenance – Language Attrition: The Case of Polish Children in Sweden (Warschauer Studien zur Germanistik und zur Angewandten Linguistik)* (New). Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Mędak, S. (2011). *Praktyczny słownik łączliwości składniowej czasowników polskich*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych UNIVERSITAS.

Możdzierz, T. (2020). Długość przeciętnego polskiego wyrazu w tekstach pisanych w świetle analizy korpusowej. *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców*, *27*, 177-192.

Nagy, N. (2011). A multilingual corpus to explore geographic variation. *Rassegna Italiana di linguistica applicata*, *43*(1-2), 65-84.

Sadowska, I. (2012). *Polish: A comprehensive grammar*. Routledge.

Van Deusen-Scholl, N. (2003). Toward a definition of heritage language: Sociopolitical and pedagogical considerations. Journal of language, identity, and education, 211-230.

Wolski-Moskoff, I. (2019). *Case in Heritage Polish. A Cross-Generational Approach* (Doctoral dissertation, Ohio State University). OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=osu1573395670224938

Zasina, A. J., & Kaczmarska, E. (2020). *Infrastructure of the Polish Learner Corpus PoLKo*. Retrieved from https://www.researchgate.net/publication/342888260_Infrastructure_of_the_Polish_Learner_Corpus_PoLKo. https://doi.org/10.13140/RG.2.2.23874.40648