

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

Simon Zuberek
December 13th, 2022

Heritage Language Learners

Students who have been raised with a strong cultural connection to a particular language through family interaction¹.

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

Basic Skills Difficulty

Non-Heritage v. Heritage Learners

Non-Heritage

1. Speaking (Productive)
2. Writing (Productive)
3. Listening (Receptive)
4. Reading (Receptive)

Heritage

1. Writing (Productive)
2. Reading (Receptive)
3. Speaking (Productive)
4. Listening (Receptive)

Where 1 is most difficult and 4 is least difficult

Why is writing hard?

- Polish nominal morphology (i.e. cases)
- Nominal mistakes made by heritage learners of Polish²:
 - Overgeneralize the use of the LOC case after certain bivalent prepositions.
 - Express DOs in the ACC case following the verbs requiring GEN DOs.
 - Express DOs in the ACC case following negated verbs (normally requiring GEN)

2. 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

Research Questions

1. Are these error types enough to differentiate between non-heritage and heritage written output?
(Relevant for software design)
2. Are some of these error types more consequential than others in making that distinction?
(Relevant for instructional design)

Approach

Develop a supervised ML classifier that distinguishes between heritage and non-heritage written output based on the errors committed.

Features:

- Counts of LOC pos-prepositionally
- Counts of GEN objects following verbs that take the GEN case
- Counts of GEN objects following negated verbs
- Per-character entropy

Data

Training

- The corpus of [Heritage Language Variation and Change](#) (HLVC)
- Interviews with Polish heritage speakers
- [PoLKo, the Polish Learner Corpus](#)

Testing

- Essays written by non-heritage learners of Polish as a foreign language (UIC)
- Essays written by heritage learners of Polish as a foreign language (UIC)

Data (continued)

Training

- 41 Heritage Docs:
 - HLVC: 35 interviews
 - In-person: 6 interviews
- 36 Non-Heritage Docs:
 - PoLKo: 36 essays

Testing

- 9 heritage essays
- 9 non-heritage essays

Baseline: 0.5

Results: Test Data

Test Data Accuracy:

Model	LOC after Prepositions	Genitive after Verbs	Genitive of Negation	Per-Character Entropy	All Features Combined
MultinomialNB(alpha=1)	0.5	0.5	0.5	0.5	0.5
ComplementNB()	0.5	0.5	0.5	0.5	0.5
DecisionTreeClassifier()	0.389	0.556	0.5	0.5	0.5
RandomForestClassifier()	0.389	0.556	0.5	0.5	0.5
SVC(kernel='linear')	0.5	0.5	0.5	0.5	0.5

Results: Cross-Validated Train Data

Cross-Validated Training Data Accuracy:

Model	LOC after Prepositions	Genitive after Verbs	Genitive of Negation	Per-Character Entropy	All Features Combined
MultinomialNB()	0.489	0.489	0.489	0.489	0.668
ComplementNB()	0.5	0.5	0.5	0.5	0.668
DecisionTreeClassifier()	0.853	0.794	0.839	1	1
RandomForestClassifier()	0.901	0.744	0.826	1	1
SVC()	0.864	0.622	0.596	1	0.864

Discussion

Challenges

- Data scarcity
- Difference in text genres (training vs. testing)

Test Data

- Same as the baseline
- Tree-based algorithms
- Feature Ranking:
 1. GEN post Vs
 2. Negation
 3. LOC post PPs

Cross-Validated

- Better than the baseline
- Tree-based algorithms
- Feature Ranking:
 1. LOC post PPs
 2. Negation
 3. GEN post Vs

Research Questions Revisited

1. Are these error types enough to differentiate between non-heritage and heritage learner output?

A cautious “yes”, given more data representing the same genre.

2. Are some of these error types more consequential than others in making that distinction?
(Relevant for instructional design)

A definitive “yes”, with concrete pedagogical ramifications.

Thank you.

