EasyChair Preprint

№ 3912

Using Automatic Measurements of Morphological Features to Distinguish Spoken and Written Discourse.

Rurik Tywoniw and Scott Crossley

July 21, 2020

Using Automatic Measurement of Morphological Features to Distinguish Spoken and Written Discourse.

Rurik Tywoniw[1]

Scott Crossley[1]

[1]Department of Applied Linguistics and English as a Second Language, Georgia State University

Author Note

The authors declare that there no conflicts of interest with respect to this preprint.

Correspondence should be addressed to Rurik Tywoniw. Email: rtywoniw1@gsu.edu

Abstract

Morphological accuracy, complexity, and awareness are often considered important benchmarks in language acquisition and performance. Though morphology is underexplored in natural language processing, automatic measurement of morphological complexity in English can lend insights into various aspects of text and discourse processing. This study introduces a tool to automatically process morphological complexity in texts. Spoken and written English-learner corpora were analyzed using the tool to explore the relationship between morphological complexity and language mode.

Keywords: Morphology, learner corpora, natural language processing

Using Automatic Measurement of Morphological Features to Distinguish Spoken and Written Discourse.

This paper presents a study integrating Natural Language Processing (NLP) and morphology complexity. Morphological features of language have historically been seen in linguistics as a gateway to understanding implicit language knowledge and acquisition processes (DeKeyser, 2000; Nagy, Berninger, Abbot, Vaughn and Vermeulen, 2003). However, morphology is often overlooked as an element of variation in discourse. This may stem from morphology being strongly tied to local syntactic constraints. While this may be the case for English inflectional morphology, the use of complex derivational morphology is constrained by how words and phrases relate to one another across larger units of text.

Recent research in natural language processing has provided linguistic analysis innovations in the form of powerful and efficient automatic text analysis tools (Graesser, McNamara, Louwerse, and Cai, 2004; Kyle & Crossley, 2015). However, few, if any, tools focus on morphological complexity and corpus-based studies of morphology are rare. In one case, Brezina and Pallotti (2019) introduced a Morphological Complexity Index which examines diversity and density of inflections in a text finding strong positive correlations between inflectional complexity and proficiency.

This paper extends a young line of research by using the newly developed Tool for Automatic Measurement of Morphological Information (TAMMI). Corpora of spoken and written English language learner production were analyzed to examine to ask the research question: how do morphological complexity features vary between modes of production in English learner texts?

Method

Data

Data in this study came from two learner corpora: 125 texts from the spoken subcorpus from the National Institute for Information and Communications Technology Japanese Learner English corpus (Izumi, 2004), and 125 texts from the L1 (first-language) Japanese subcorpus from ICNALE (International Corpus Network of Asian Learners of English) written corpus (Ishikawa, 2013).

Linguistic analysis

We used the Tool for Automatic Measurement of Morphological Information (TAMMI) to measure the incidence of inflectional and derivational morphology in texts. Using a reference list of affixes, TAMMI parses each word in a text and provides the word's lemma (word stem without inflectional morphology), and derivational base, the smallest meaningful unit within a word stripped of any inflectional or derivational affixes. It can then calculate per text indices for types and tokens which complex morphological information. TAMMI calculated each text's number of inflected word tokens per word and word types per type and derived word tokens per word and word types per type.

Statistical Analysis

The above indices were analyzed using group-wise comparison (t-tests) between the written and spoken corpora. Indices that demonstrated differences between the two registers were selected for use in a logistic regression to predict mode of production (speaking or writing). The logistic regression was structured using a training set (100 out of 125 texts in each corpus) and a test set (the remaining 25 out of 125 texts in each corpus). The results of the pairwise tests and the accuracy of the logistic regression are presented below.

Results

Table 1 presents mean measurements for each TAMMI index in the JLE speaking corpus and the Japanese ICNALE writing corpus, and results from t-tests comparing the indices between the spoken and written texts. The proportion of derived words and types were greater in written texts. Types of inflected words per word type were more associated with spoken texts, for which there was marginally significant difference.

Table 1

Pairwise comparisons of TAMMI indices between English learner speaking and writing.

Index	Written M (SD)	Spoken M (SD)	*t*	*p*	*d*
Inflected word tokens per word	0.156 (.027)	0.161 (.045)	-0.905	0.366	-0.114
Derived word tokens per word	0.169 (.023)	0.144 (.046)	5.465	0.000	0.691
Inflected word types per type	0.194 (.031)	0.209 (.055)	-2.571	0.011	-0.325
Derived word types per type	0.082 (.022)	0.048 (.029)	10.597	0.000	1.340

The three indices found to significantly differ between written and spoken texts were standardized and used in a logistic regression to predict mode of production (Table 3). Each predictor was found to contribute significantly to the model, though Inflected Types did so only marginally. The column of coefficients (B) shows the direction of prediction that each index contributed to the model, with positive predictors predicting writing and negative predictors predicting speaking. Log odds show how many times more likely a text was to be a written text as opposed to a speaking text given one standard deviation increase.

Table 2.

Logistic regression model predicting text mode using TAMMI indices in training data.

Predictor	B	*Log odds*	SE	χ^2	*p*	VIF*
(Intercept)	-0.042	0.959	0.191	0.047	0.828	
Derived Types per type	1.884	6.581	0.282	44.758	< .001	1.172
Inflected Types per type	-0.497	0.608	0.222	5.002	.025	1.077
Derived tokens per token	0.888	2.430	0.213	17.355	< .001	1.092
Pseudo R^2	.348					

*Variance Inflation Factor: values over $1/(1-R^2) = 1.533$ indicate predictor variables with high multicollinearity to other predictors and inaccurate coefficient calculations.

The accuracy of prediction is presented as a confusion matrix in Table 4 below. The model's overall accuracy was significantly more predictive than chance, $\chi^2 = 13.718$, $p = .0002$, $\kappa = .56$ (a moderate effect size).

Table 3.

Confusion matrix for logistic regression predictions of L2 production mode in the test set using morphological complexity frequency.

	Predicted mode		
Actual mode	Writing	Speaking	
Writing	21	4	84.00%
Speaking	7	18	72.00%
Overall % Correct			78.00%

Note: Overall % correct by chance = 50%

Discussion

The results from the pairwise tests and the logistic regression indicate that morphological features of language production differ significantly between learner speaking and learner writing. The proportion of word types with inflectional morphology was higher in spoken texts than in written texts, even though the overall proportion of word tokens with inflectional morphology was roughly equal between spoken and written texts. Written texts were found to have a higher proportion of word types and tokens with derivational morphology than spoken texts, suggesting that written texts involve greater employment of complex word construction processes. Logistic regression results further indicated that morphological features were predictive of mode of production. Future studies can compare morphological features across more narrowly defined registers. Beyond investigations of discourse between registers, learner corpus investigations of morphology could unearth information about the interaction of proficiency with morphological features.

References

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. Second *Language Research, 35*(1), 99–119. https://doi.org/10.1177/0267658316643125

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499–533.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers, 36*(2), 193-202.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (pp. 91-118). Kobe, Japan: Kobe University.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly, 49*(4), 757-786.

Nagy, W., Berninger, V., Abbott, R., Vaughan, K., & Vermeulen, K. (2003). Relationship of morphology and other language skills to literacy skills in at-risk second-grade readers and at-risk fourth-grade writers. *Journal of educational psychology, 95*(4), 730.