34 Managing Second Language Acquisition Data with Natural Language Processing Tools

Scott A. Crossley and Kristopher Kyle

1 Introduction

Second language (L2) data in the form of learner corpora are linguistically untidy when compared to the majority of language data corpora because L2 corpora are rife with grammatical errors, neologisms and borrowed terms, unique phrasal items, misspellings, punctuation problems, and ill-formed syntactic constructions. Nonetheless, much of second language acquisition (SLA) research investigates patterns of learning within large learner corpora that may be composed of either naturalistic or elicited language production. Working with L2 learner corpora introduces unique sets of qualifiers and limitations that may not exist in more standardized data sets (such as first-language production samples). Thus, the management of L2 data requires specific insight to avoid potential pitfalls that may make data analyses less robust, leading to erroneous conclusions. This becomes even more important when researchers rely on natural language processing (NLP) tools to automatically assess the content of L2 learner corpora because these NLP tools process language literally without knowledge of text content, learner background, or the data collection context.

The purpose of this chapter is to introduce proper data management techniques specific to L2 data analyses that rely on NLP tools and learner corpora. Specifically, we introduce and discuss the limitations of both NLP tools and learner corpora. We then provide details about proper data management workflow unique to L2 corpora and NLP tools (section 2). We next present a hands-on case study to guide the reader through the process of an NLP analysis of a learner corpus (section 3). We follow this with a discussion of the analysis and a guide to interpreting NLP results (section 4). We conclude with a general statement about the strengths and limitations of both NLP analyses and learner corpora (section 5).

1.1 Natural language processing

NLP encompasses all computerized approaches to analyzing language in order to measure linguistic features to better understand various features of language use (e.g., developmental trajectories, register variation). All NLP tools are computer programs that rely on a sequence of instructions that tell the program how to complete a task. NLP also requires, at some level, knowledge of language that can either be derived explicitly or implicitly. Explicit language knowledge can be provided to computer programs as databases of lexical items or rules for part-of-speech tagging. Implicit language knowledge can be derived by through neural network models that then inform NLP approaches. The strength of NLP programs is their efficiency in analyzing massive amounts of data by repeating analyses objectively and literally, something that is time consuming and difficult for humans to accomplish.

The two main purposes of NLP analyses are to better understand language and cognition (i.e., gather information on how we understand language and use language; a cognitive science approach) or to respond appropriately to humans using natural language (i.e., an artificial intelligence approach). For example, researchers may want to study the quality and the content of L2 writers' essays longitudinally to better understand L2 writing development. Alternatively, researchers may want to develop NLP algorithms to provide feedback to (L2) writers about the quality of their texts to help guide the writers through the revision process. NLP tools are generally not used in isolation and are instead mixed with statistical methods that are based on inferential techniques or machine learning algorithms based on

probabilistic models to increase the reliability and validity of their output. In addition, NLP tools can be used to complement qualitative analyses by providing objective support for more nuanced examinations of language.

NLP tools can also be used to assess a number of domains within language studies, including language acquisition, reading ability, speaking proficiency, and writing development (Crossley et al. 2011; Kyle, Crossley, & Berger 2018; Crossley & McNamara 2013). NLP tools can also be used to assess the mental states of students in terms of engagement, boredom, confidence, and openness as well as individual differences related to prior knowledge, reading skills, persistence, language ability, native language used, and working memory (Allen & McNamara 2015; Allen, McNamara, & McCrudden 2015; Allen, Mills, et al. 2016; Allen, Perret, & McNamara 2016; Crossley & McNamara 2012; Crossley, Salsbury, & McNamara 2012). Beyond assessing student and learning development and characteristics, NLP has many other educational applications, including but not limited to, assessing text readability, predicting math proficiency, categorizing text disciplines and registers, and predicting task types (Biber 1988; Biber & Conrad 2009; Crossley et al. 2018; Crossley, Skalicky, et al. 2017; Crossley, Russell, et al. 2017; Guo, Crossley, & McNamara 2013; Kyle & Crossley 2016).

There are a number of freely available NLP tools that have been developed recently to help non-specialists conduct NLP analyses (Crossley, Kyle, & Dascalu 2018; Kyle & Crossley 2017; Kyle, Crossley, & Berger 2018). Most of these are in English, but some tools are multilingual (MacWhinney 2014; Dascalu et al. 2015). Linguistically, the tools can provide information about text cohesion, lexical attributes of a text, syntactic complexity metrics, and emotion and affective features. In turn, these features can be used to better understand L2 production in learner corpora.

1.2 Learner corpora

Learner corpora are electronic collections of spoken, signed, and/or written language produced by individuals who are learning a particular second (third or fourth, and so on) language. Learner corpora vary widely in their characteristics and in the metadata available. Learner corpora are often characterized by the nature of the data collection (e.g., natural conversations, semistructured interviews, timed production tasks), the particular L2(s) that are represented, the first-language (L1) group(s) that are included, and the types of metadata included (production quality scores, standardized proficiency scores, age, gender, educational settings, time spent learning the L2). Furthermore, learner corpora are distinguished with regard to whether the data are cross-sectional or longitudinal in nature. Given the difficulty in collecting longitudinal data (e.g., attrition), most learner corpora are cross-sectional in nature, though a number of longitudinal corpora have been recently made available (e.g., LANGSNAP [Languages and Social Networks Abroad Project]; Tracy-Ventura et al. 2016). Cross-sectional and quasi-longitudinal corpora allow researchers to examine trends across (usually a large number of) learners of different proficiency levels. These corpora help researchers create and refine assessment criteria, set learning benchmarks, and discover potential developmental trends. Longitudinal corpora, which are usually much smaller due to participant recruitment and attrition issues, allow researchers to examine the actual developmental trajectories of language learners. These trajectories can be non-linear (Verspoor et al. 2017), which cannot be captured in cross-sectional and quasi-longitudinal corpora, demonstrating their limits.

2 SLA data management workflow

2.1 Corpus collection

Most SLA NLP analyses depend on the use of learner corpora. These corpora form the foundation for NLP analyses and need to be developed and/or selected carefully and conscientiously. Mistakes in collecting wide enough corpus parameters, specifically metadata, can prove extremely problematic to interpretations of results. Some elements that need to be controlled for when collecting and/or analyzing learner corpora are L1 influence, age-related variables, gender, proficiency level, task, prompts and topics, and parallelisms between corpora that are combined in analyses.

Some analyses will require research to control for L1 influence. This will be the case when a corpus includes data from speakers of multiple L1s. Researchers can use a corpus that is composed of L2 users from the same L1. However, such an approach will not allow for findings to be generalized beyond the L1 represented in the corpus, which will limit any interpretations of the data. If multiple L1s are represented in a corpus, researchers have a

few options. First, a corpus or a subcorpus of a larger corpus can be analyzed wherein the participants speak similar languages within a larger language family. This would allow for some generalizability of the findings beyond a single language group. However, the best approach is to use a corpus that affords representation from a variety of L1s and not a subsection of that corpus or corpus that represents only a single L1 background. Using a corpus that includes multiple L1s allows for greater generalizability of findings and can be robust if principled statistical analyses are used to factor in crosslinguistic influence of the L1s. For instance, the L1 of a participant can be included as a categorical fixed effect in mixed-model analyses, or a continuous variable such as linguistic distance can be used (Chiswick & Miller 2005).

Age is another important factor to consider when using NLP tools on learner corpora. Much research has demonstrated differences in L2 performance based on the age at which individuals began learning a particular L2 (referred to as age of onset; Abrahamsson & Hyltenstam 2009; Hyltenstam & Abrahamsson 2003). While it is feasible to analyze a corpus that contains participants of the same age of onset and even potentially the same learning experience background, it is unlikely. A robust learner corpus will contain this metadata for all participants such that researchers can control for age-related effects in mixed-effects models.

The most important variable to control for in a corpus (and one that likely supersedes data on L1 or age-related variables) is L2 language proficiency. A large body of research has indicated that the language produced by L2 users varies greatly by proficiency (Kyle & Crossley 2015, 2017; Lu 2011; Laufer & Nation 1995). In the absence of proficiency level information, it is difficult to interpret findings from NLP tools.

Beyond controlling for individual differences such as L1, age-related variables, and proficiency, NLP analyses need to carefully control for context. The two most obvious contextual factors are the task and the topic, both of which can exert influence on language production (Biber & Gray 2013; Kyle & Crossley 2016; Kyle, Crossley, & McNamara 2016). In terms of task, a good example is the difference in linguistic output that occurs between independent tasks that require participants to rely on existing knowledge or integrated tasks that require participants to incorporate information from outside sources. Multiple studies have shown that these two tasks produce different linguistic output from participants (Guo, Crossley, & McNamara 2013; Kyle & Crossley 2016). If task is not controlled for, interpretation of results may be problematic. Similarly, topic can exert a strong influence on linguistic production with many studies indicating that the formality of the topic or the difficulty of the topic matter can lead to different types of linguistic production (Biber & Gray 2013; Hinkel 2009). In some cases, researchers will focus on a single task with a single topic, but this limits the generalizability of findings. It is best to control for task and topic in statistical analyses.

Another key consideration is parallelism in corpora. Often, researchers will combine corpora to develop as large a sample as possible to better reflect L2 learning. While this is not problematic on its own, and, in fact, can be good practice, researchers need to ensure that combined corpora are comparable (Sinclair 2005). In addition, often researchers will have to select which reference corpus in a specific NLP tool they want to use to best match the corpora at hand. That is to say, there are many instantiations for NLP indices based on different corpora and researchers need to select the appropriate corpus for their analysis. For instance, if a researcher is interested in examining lexical growth in classroom writing in terms of word frequency, the reference corpus used to develop the frequency measure is an important consideration. If the class is an English for Academic Purposes class, it would be best to use frequency counts from an academic corpus. However, an academic corpus may not be appropriate for a class that focuses on journal writing or narrative writing.

2.2 Corpus preprocessing

Once a corpus has been rigorously developed and/or selected, a number of preprocessing decisions need to be made. Chief among these are decisions about text formatting, text cleaning, and spelling correction. In terms of text formatting, most NLP analyses need to be conducted on texts that are in plain text formats (e.g., .txt). Plain text formats contain no special formatting (e.g., images, objects, and other elements such as headings, footnotes, or special font characteristics) and, as such, preserve the basic linguistic information in the text in a format that is machine-readable.

Researchers also need to make decisions about cleaning texts of unneeded and unwanted information. For instance, many times during transcription, text

conversion, or when text files are switched between operating systems, a number of opportunities for non-Unicode characters to permeate texts occur. These characters will cause difficulties for many NLP tools because they will not recognize the characters, possibly causing the tools to shut down. Researchers need to carefully examine a percentage of texts to make sure they only include Unicode characters. In addition to non-Unicode characters, researchers need to make principled decisions about the inclusion or exclusion of disfluencies (i.e., false starts, attention signals, word repetition, and filler words). Such disfluencies are common in speech and NLP tools will process them regardless of structural concerns. However, the inclusion of disfluencies could strongly influence outcomes, especially lexical outcomes. If L2 learners use "yea" or "okay" to signal that they are paying attention, an NLP tool will assume the learner is using very frequent words more often, even though these words may not represent actual production. The same may be true if false starts are included or filler terms such as "you know," "I mean," or "like" are included.

Lastly, researchers need to make a principled decision about spell-correcting L2 data (specifically written data). L2 written data is rife with spelling errors and these errors may not represent wider language knowledge, but only knowledge of spelling conventions. However, they do represent actual production and researchers may want to maintain the independence of the data. Problems also arise with spelling correction at the conceptual and practical levels. Conceptually, it may be difficult to accurately fix all misspellings because of context. Additionally, many L2 learners may use words from their L1 (either cognates or non-cognates) and it will not be clear to researchers which words represent actual production and which ones are guesses, estimates, or non-words. Practically, when dealing with large data sets, it is difficult to spell-correct texts without the aid of computer code. While spelling correction engines exist or can be programmed for specific purposes, they will likely be less accurate than human judgments.

2.3 Text processing

Once the corpora have been selected or developed and preprocessed, the easy part of using NLP tools comes into play: the actual text processing. Many NLP tools are extremely user-friendly, while some require a certain amount of familiarity with computer science or fluency in a computer science language. For the purposes of this chapter, we introduce the readers to the Suite of Automatic Linguistic Analysis Tools (SALAT) developed by the authors of this chapter.

SALAT contains freely available desktop tools that work on Windows, Mac OSX, and (in many cases) Linux operating systems. They allow users to process large numbers of texts in a (relatively) short period of time. All texts must be formatted in ASCII or eight-bit Unicode transformation format and saved as plain text (.txt) files to preserve Unicode formatting. The tools provide different processing options, but generally allow users to select specific indices (e.g., content word frequency) or index categories (e.g., clausal complexity). After index selection has been completed, the user indicates where the desired files to process are located, selects a name and location for the output file(s) (i.e., where the results of the analyses will be written), and presses the Process Texts button. The internal processes of each tool differ somewhat, but generally speaking each text is lemmatized, tagged for part of speech, and possibly parsed using Stanford CoreNLP (Manning et al. 2014). The processed texts are then analyzed using Python scripts that either look up each item of interest (e.g., word, *n*-gram, or syntactic structure) in a database (e.g., of word naming response times) or conduct a formal analysis (e.g., number of words per clause) and provide average scores for each text. After the program has processed each text, values are written to a machine (and human) readable output file in .csv format. Users can then analyze their data in other programs such as R.

2.4 Text length

Once the tools are processed through the NLP tools, it is likely that the length of each text will become available. The length of a text needed for robust NLP analyses is difficult to pin down because the number of words needed to provide an accurate profile of a learner's language skills is an open question. A common rule of thumb is that texts should be long enough to provide enough linguistic coverage to accurately portray the linguistic knowledge of the person that produced that text. However, the exact number of words needed is unknown and may fluctuate among learners and contexts. Little work in this vein has been conducted in L2 learner research. Work in educational sciences and other fields indicates that language analyses benefit from longer text samples

(Allen, Snow, & McNamara 2015; Crossley 2018; Varner et al. 2013; Vyas & Uma 2018; Young et al. 2018; Zhang, Huang, & Zhao 2018) because longer texts provide stronger and more reliable NLP results that better represent language profiles. Of the studies conducted using NLP tools with L2 learner data, an agreed upon threshold seems to be around one hundred words in order to support reliable results (Crossley et al. 2011; Crossley & McNamara 2013).

2.5 Data organization

One problem with many NLP tools is that their output can be overwhelming. For instance, Tool for the Automatic Analysis of Lexical Sophistication (TAALES) will produce a spreadsheet that may have over five hundred columns for every text for which it calculates features. Many of the features calculated may not be of interest to the user and many may be redundant. Thus, it is generally a good idea to preselect indices of interest that will help address the research questions in a study. Preselecting indices will help control the number of features in a spreadsheet and will also help control for type 1 errors in one's statistical analysis (Larson-Hall 2015).

It is generally recommended to put data in the format that best answers a particular research question. For instance, a long (or narrow) format comprises a column that contains the context of the values found in another column containing the values of interest. Long formats may contain multiple variable types in a single column. Most L2 NLP data are best fit into wide formats wherein each column represents an NLP feature of interest and each row represents a text from which those data were collected.

Data can also be represented in various file types, the most common being comma separated values (.csv) and Excel files. In practice, it is easier to use .csv files to store data because they are more easily accessible to statistical and machine learning programs. For instance, R (R Core Team 2014) will read .csv files quite easily while Excel files need more manipulation. As well, machine learning packages such as WEKA (Frank, Hall, & Witten 2016) will also read .csv files (but not Excel files).

2.6 Statistical analysis

Once the data from the NLP tools is suitably organized, analyses of the data can begin. Because researchers will likely be looking at thousands of data points and each point is represented by an incidence or ratio score, statistical or machine learning approaches are required. In addition, with such large samples, there are a number of statistical concerns that must be addressed including normality of data, multicollinearity of the data (i.e., when various features measure the same construct), overfitting, and suppression effects (for more information, see Field 2013).

While normality of the data is not an assumption in all analyses, it is important to examine the incidences of zero features found in a text. While not common, it is possible that some features measured by an NLP tool will not be found in texts and these texts will report a zero for that feature. As an example, consider the argument overlap features reported in Tool for the Automatic Analysis of Cohesion. A researcher interested in overlap between paragraphs (i.e., global cohesion) in L2 writing may find that many students don't produce writing samples that contain more than one paragraph. In these cases, the tool will report zeroes for paragraph argument overlap. While this may not affect statistical analyses, it will make generalizations to new data or new tasks problematic. A similar concern can be found with the academic word sublists found in TAALES. In some instances, the words in these lists may not occur in L2 data either because of the rarity of the words or the shortness of the L2 texts. If data are non-normally distributed, certain statistical and machine learning approaches cannot be used and alternatives will have to be used. Researchers may also decide to remove these features because they do not accurately represent learners' language profiles.

With hundreds of variables reported by NLP tools and many variables being conceptually redundant (e.g., multiple frequency features), researchers have to be careful not to include similar variables. The easiest solution to this problem is to control for multicollinearity among variables, which can be done with simple correlational analyses. Generally, multicollinearity is set at $r \geq .700$ or researchers use variance inflation values below 5 (Field 2013). These thresholds ensure that variables are not strongly correlated, but allows similar variables to be included (i.e., lexical variables such as concreteness and frequency, which measure different lexical constructs, will be included).

Overfitting may also be a problem in certain contexts. Briefly, overfitting occurs when a statistical model (e.g., a linear regression) attempts to use all variables to predict

an outcome regardless of whether the derived model predicts future observations. Simply put, the regression model will force unwarranted variables to fit the data and explain a greater amount of variation regardless of whether the variation explained represents the model or represents statistical noise. Because NLP tools calculate hundreds of variables, researchers may try to incorporate all of them into a model for a relatively small corpus, producing a model that contains more variables/parameters than are justified. A general rule of thumb is that researchers should include one variable for every fifteen items (see, e.g., Tabachnick & Fidell 2013). Thus, if your learner corpus contains samples from three hundred learners and the outcome variable is learner level, the researcher should limit the number of NLP features to twenty (i.e., one variable for every fifteen items in the analysis).

Researchers need to also be careful with including variables that show flipped signs (i.e., suppression effects) in statistical models. A statistical model may include a variable that raises the amount of variance explained because it accounts for residual variance left over after other variables have been added. It is possible that the residual variance explained by this variable is not due to its association with the dependent variable (in this case writing quality score), which can often be seen because of flipped signs between the correlation of the variable on its own and within a model. For example, when examining links between L2 writing quality and lexical variables, it is likely that frequency will negatively correlate with quality such that more frequent words are found in lower quality writing samples. However, if frequency is added into the model late, perhaps after other lexical variables have been added such as concreteness, the reported co-efficient may be positive instead of the expected negative seen in the initial correlation. This is a strong indicator that there is a suppression effect in which case the frequency variable should be removed. To control for suppression effects, researchers should check initial statistical patterns (e.g., correlations, t-tests, or ANOVAs) between the independent variables and the dependent variables in isolation to the patterns reported in statistical models.

3 Case study

3.1 Data

As a case study, we present an examination of L2 lexical development over the course of a semester of study. The data in this case study has been previously used in two other studies examining lexical development in terms of word concreteness, familiarity, meaningfulness, age of acquisition (Crossley & Skalicky 2019) and word frequency (Crossley et al. 2019). In this case study, we examine lexical development in terms of word-naming times for words produced by L2 learners during naturalistic oral output with both English L1 and L2 users. The users included 50 non-matriculated L2 English learners enrolled in an intensive English program. These 50 intensive English program students conversed with 50 matriculated L1 and L2 speaking students (25 L1 and 25 L2) enrolled in undergraduate- and graduate-level TESOL courses over the course of a semester. The L2 students were recorded at minimum twice during the semester and at maximum four times. More information about these data can be found in Crossley and Skalicky (2019) and Crossley et al. (2019).

We specifically use these data because they afford control for a number of demographic, task, and individual differences in the learners as we have discussed. These include distance from the users' L1s to English using a language distance measure (Chiswick & Miller 2005) to control for potential crosslinguistic influence as well as age, gender, L2 proficiency level (as measured by standardized language assessment tests), order of data collection session, and whether the interlocutor was an L1 or L2 user.

3.2 Word-naming index

The word-naming norms used in this study were derived from a standard psychological task for measuring word recognition. In word-naming tasks, participants are presented with an orthographic word that they then must name aloud. The time it takes for the participant to begin pronouncing the word is measured as the response time. The word-naming response times used in this case study were derived from the English Lexicon Project, a large publicly available behavioral and descriptive data set (Balota et al. 2007). This data set includes word-naming response norms from 816 participants, with 2,500 per participant for the word-naming tasks. In total, response times were calculated for 40,481 words. All participants were native English users. Previous research has demonstrated that L2 learners produce words that are named more quickly over time and that learners rated as more lexically proficient by human raters produced words that are named more slowly (Berger, Crossley, & Kyle 2019).

However, this previous research did not control for demographics, individual differences, or task features.

3.3 Statistical analysis

We used linear mixed-effects models in R (R Core Team 2014) using the lme4 package (Bates et al. 2015) to examine any potential for changes in word-naming response times over time. The model tested whether any independent variables significantly predicted changes in word-naming response times (i.e., the dependent variable). In the model, we entered age, gender (female or male), language distance, English proficiency (advanced, high intermediate, low intermediate, or high beginning), session order (1–4), and pair language match (i.e., L1/L2 or L2/L2) as predictor variables (i.e., fixed effects).

The baselines comparison for our proficiency variable was advanced. Baseline for gender was female and baseline for pair language match was L2/L2. We also included subjects as random effects as well as a random slope of session order fit to participants. Lastly, within the model, we fitted an interaction between session order and English proficiency to measure whether English proficiency influenced word frequency production over time. We used lmerTest (Kuznetsova, Brockhoff, & Christensen 2017) to derive p values from the models, multcomp (Hothorn, Bretz, & Westfall 2008) to obtain full pairwise comparisons between the four proficiency levels (advanced, high intermediate, low intermediate, and high beginner), the effects package to retrieve model information for our figures, and the MuMIn package (Nakagawa & Schielzeth 2013) to obtain two measures of variance explained: marginal R^2, which measures the variance explained by the fixed effects only, and conditional R^2, which measures the variance explained by both the fixed and random effects.

4 Results

The linear mixed-effects model reported a significant main effect for age, and a significant interaction between English proficiency (high intermediate learners compared to advanced learners) and session order. In the case of the interaction, the high intermediate participants' word-naming response times for the words they produced was significantly lower during earlier sessions when compared to advanced participants, but then increased significantly as session order increased. This interaction is visualized in figure 34.1. This model reported a marginal R^2 of .281 and a conditional R^2 of .393. Table 34.1 displays the estimates, standard errors, t values, and p values for the fixed effects entered into this model.

5 Discussion and conclusion

We present an overview of how data derived from NLP tools are best managed and provide a sample analysis to highlight the overview. The sample analysis follows the suggestions in the overview by examining a longitudinal learner corpus in terms of lexical development while controlling for a number of variables known to interact with linguistic features including age, gender, language proficiency, and L1 considerations. The findings from

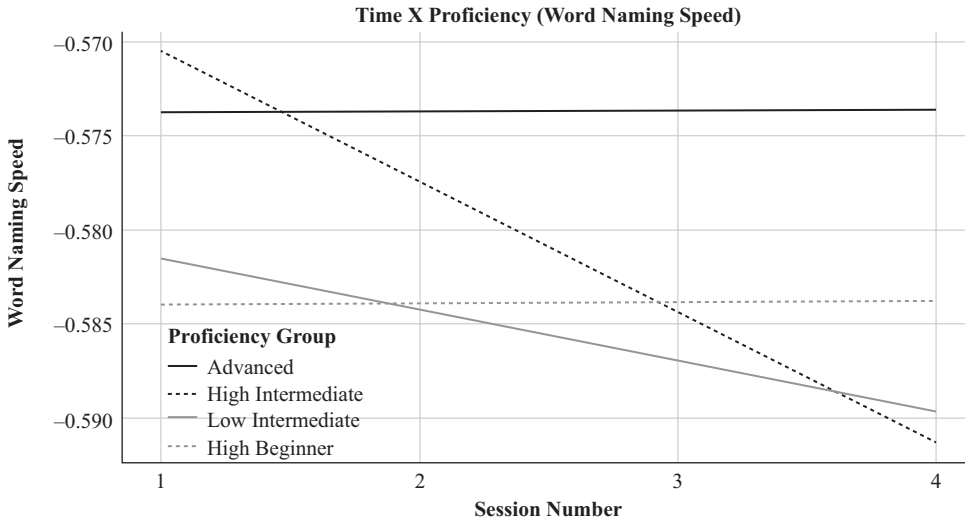


Figure 34.1
Interaction between proficiency and session order.

Table 34.1
Linear mixed-effects model predicting word-naming response times

Fixed effects	Estimate	Standard error	*t*	*p*
(Intercept)	–0.602	0.009	–65.181	<.001
Sex: male	0.002	0.004	0.599	.552
Match: L1-L2	–0.005	0.004	–1.124	.267
Age	0.001	0.000	5.649	<.001
Proficiency				
High intermediate	0.010	0.010	1.057	.294
Low intermediate	–0.005	0.009	–0.582	.563
High beginning	–0.010	0.011	–0.932	.354
Session order	0.000	0.002	0.015	.988
Language distance	–0.003	0.004	–0.914	.365
Proficiency × session order interaction				
High intermediate	–0.007	0.003	–2.143	<.050
Low intermediate	–0.003	0.003	–0.949	.345
High beginning	0.000	0.004	0.006	.995

Baselines for categorical variables are as follows: sex = female; match = L2-L2; proficiency = advanced.

the sample analysis indicate that L2 learners show a significant interaction with time and proficiency such that high intermediate learners show significant decreases in word-naming speed across time while learners from other proficiency levels do not. Age was also a significant predictor, indicating that older participants produced words with higher word-naming times. In total, the analysis indicates that older participants produce more sophisticated words and that high intermediate learners show a trend toward producing less sophisticated words over time when compared to other proficiency level learners.

This simple case study demonstrates some of the strengths of using NLP tools to examine SLA data. First and foremost, the measures reported by the NLP tool are objective and are calculated similarly across learners. This provides credibility to the study in terms of reproducibility (i.e., the ability to reproduce this analysis) and replication (i.e., the ability to reproduce this study using new data following similar research methods). Concerns about subjective rating scales, differences in material designs, and human coding are mitigated by NLP tools' intrinsic designs. In addition, the measures reported by NLP tools are generally transparent. The word-naming scores used in this study can be traced back to the original data set released by Balota et al. (2007), which provides extensive information about the collection techniques (although this is not always the case). In addition, a tool such as TAALES provides additional information to ensure greater transparency including output that shows word-naming scores for each individual word and percentage of words in each text that have word-naming scores. Lastly, the website that supports the tool (linguisticanalysistools.org) provides supplementary information related to how the indices are calculated.

There are limitations to using NLP tools for SLA analyses. Among these are depth of features reported by NLP tools; the production of words, phrases, or structures not recognized by an NLP tool; and the breadth of indices available. From a depth perspective, it should be noted that NLP tools often only measure surface level features of text (though there are many exceptions) and the tools do not have the capability to measure the context in which words are used and whether the language structures produced are appropriate for the discourse at hand. Thus, while NLP tools can measure many linguistic features, they cannot assess features related to style, strength of argumentation, effect on interlocutor, or other contextual variables. NLP tools often rely on databases that provide the linguistic knowledge needed to analyze a text. If the database does not contain the words or structures produced by the participant, it may not be possible to present a fully accurate linguistic profile. In the case of L2 learners, this is especially true because L2 student data may contain neologisms, unique phrasal combinations, or uncommon syntactic and grammatical constructions.

These data may not be captured by NLP tools. Another problem with NLP tools is the sheer breadth of features they produce, which may overwhelm researchers. For instance, TAALES reports over 150 frequency features each with their own specific calculations, domains, and dictionaries. Without experience, researchers may select incorrect frequency features that do not match their research questions. In many cases, the NLP tools in SALAT contain component scores developed using statistical analysis that convert related indices into linearly uncorrelated, aggregated variables. For instance, Kim, Crossley, and Kyle (2018) combined hundreds of lexical indices into twelve core lexical components that are available in the newest version of TAALES. These and similar component scores can help researchers narrow down the number of features they analyze.

References

Abrahamsson, N., and K. Hyltenstam. 2009. Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59:249–306.

Allen, L. K., and D. S. McNamara. 2015. Promoting self-regulated learning in an intelligent tutoring system for writing. In *AIED 2015: Artificial Intelligence in Education*, ed. C. Conati, N. Heffernan, A. Mitrovic, and M. Verdejo. Lecture Notes in Computer Science, vol. 9112. Cham: Springer. https://doi.org/10.1007/978-3-319-19773-9_125.

Allen, L. K., D. S. McNamara, and M. T. McCrudden. 2015. Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, ed. D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, and P. Maglio. Austin, TX: Cognitive Science Society.

Allen, L. K., C. Mills, M. E. Jacovina, S. A. Crossley, S. K. D'Mello, and D. S. McNamara. 2016. Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In *Proceedings of the 6th International Learning Analytics and Knowledge (LAK) Conference*, 114–123. https://doi.org/10.1145/2883851.2883939.

Allen, L. K., C. Perret, and D. S. McNamara. 2016. Linguistic signatures of cognitive processes during writing. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, ed. J. Trueswell, A. Papafragou, D. Grodner, and D. Mirman, 2483–2488. Philadelphia: Cognitive Science Society.

Allen, L. K., E. L. Snow, and D. S. McNamara. 2015. Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, ed. P. Blikstein, A. Merceron, and G. Siemens, 246–254. New York: Association for Computing Machinery.

Balota, D. A., M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, et al. 2007. The English Lexicon Project. *Behavior Research Methods* 39 (3): 445–459.

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67 (1): 1–48.

Berger, C. M., S. Crossley, and K. Kyle. 2019. Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics* 40 (1): 22–42.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., and S. Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., and B. Gray. 2013. *Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-grammatical Analysis*. TOEFL iBT Research Report 19. Princeton, NJ: Educational Testing Service.

Chiswick, B. R., and P. W. Miller. 2005. Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development* 26 (1): 1–11.

Crossley, S. A. 2018. How many words needed? Using natural language processing tools in educational data mining. In *Proceedings of the 10th International Conference on Educational Data Mining*, 630–633.

Crossley, S. A., K. Kyle, and M. Dascalu. 2018. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods* 51:1–14.

Crossley, S. A., and D. S. McNamara. 2012. Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge. In *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*, ed. S. Jarvis and S. A. Crossley, 106–126. Bristol, UK: Multilingual Matters.

Crossley, S. A., and D. S. McNamara. 2013. Applications of text analysis tools for spoken response grading. *Language Learning and Technology* 17 (2): 171–192.

Crossley, S. A., J. Ocumpaugh, M. Labrum, F. Bradfield, M. Dascalu, and R. Baker. 2018. Modeling math identity and math success through sentiment analysis and linguistic features. In *Proceedings of the 11th International Conference on Educational Data Mining*, ed. K. E. Boyer and M. Yudelson, 11–20. International Educational Data Mining Society.

Crossley, S. A., D. Russell, K. Kyle, and U. Römer. 2017. Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics* 1:48–81.

Crossley, S. A., T. Salsbury, and D. S. McNamara. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing* 29 (2): 243–263.

Crossley, S. A., T. Salsbury, D. S. McNamara, and S. Jarvis. 2011. What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly* 45 (1): 182–193.

Crossley, S. A., and S. Skalicky. 2019. Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley, and McNamara (2011). *Language Teaching* 52 (3): 385–405.

Crossley, S. A., S. Skalicky, M. Dascalu, D. McNamara, and K. Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes* 54 (5–6): 340–359.

Crossley, S. A., S. Skalicky, K. Kyle, and K. Monteiro. 2019. Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition* 41 (4): 721–744.

Dascalu, M., S. Trausan-Matu, D. S. McNamara, and P. Dessus. 2015. ReaderBench—Automated evaluation of collaboration based on cohesion and dialogism. *International Journal of Computer-Supported Collaborative Learning* 10 (4): 395–423.

Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistics*. Thousand Oaks, CA: Sage.

Frank, E., M. A. Hall, and I. H. Witten. 2016. The WEKA workbench. Online appendix for *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. Burlington, MA: Morgan Kaufmann.

Guo, L., S. A. Crossley, and D. S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18 (3): 218–238.

Hinkel, E. 2009. The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics* 41 (4): 667–683.

Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50 (3): 346–363.

Hyltenstam, K., and N. Abrahamsson. 2003. Maturational constraints in SLA. In *The Handbook of Second Language Acquisition*, ed. C. Doughty and M. Long, 539–588. Malden, MA: Blackwell.

Kim, M., S. A. Crossley, and K. Kyle. 2018. Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal* 102 (1): 120–141.

Kuznetsova, A., P. B. Brockhoff, and R. H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82 (13). doi:10.18637/jss.v082.i13.

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

Kyle, K., and S. A. Crossley. 2016. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing* 34:12–24.

Kyle, K., and S. A. Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing* 34 (4): 513–535.

Kyle, K., S. Crossley, and C. Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods* 50 (3): 1030–1046.

Kyle, K., S. A. Crossley, and D. S. McNamara. 2016. Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing* 33 (3): 319–340.

Larson-Hall, J. 2015. *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. London: Routledge.

Laufer, B., and P. Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16:307–322.

Lu, X. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45:36–62.

MacWhinney, B. 2014. *The CHILDES Project: Tools for Analyzing Talk*. Volume 2: *The Database*. New York: Psychology Press.

Manning, C. D., M. Surdeanu, J. Bauer, F. Finkel, S. J. Bethard, and D. McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Nakagawa, S., and H. Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4:133–142.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Salsbury, T., S. A. Crossley, and D. S. McNamara. 2011. Psycholinguistic word information in second language oral discourse. *Second Language Research* 27 (3): 343–360. doi: 10.1177/0267658310395851.

Sinclair, John. 2005. Corpus and text: Basic principles. In *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne, 1–16. Oxford: Oxbow Books.

Tabachnick, B. G., and L. S. Fidell. 2013. *Using Multivariate Statistics*. Boston: Pearson.

Tracy-Ventura, N., R. Mitchell, and K. McManus. 2016. The LANGSNAP longitudinal learner corpus: Design and use. In *Spanish Learner Corpus Research: State of the Art*, ed. M. Alonso Ramos, 117–142. Philadelphia: John Benjamins.

Varner, L. K., G. T. Jackson, E. L. Snow, and D. S. McNamara. 2013. Does size matter? Investigating user input at a larger bandwidth. In *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, ed. C. Boonthum Denecke and G. M. Youngblood, 546–549. Menlo Park, CA: AAAI Press.

Verspoor, M., W. Lowie, H. P. Chan, and L. Vahtrick. 2017. Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle* 14 (1): 1–27.

Vyas, V., and V. Uma. 2018. An extensive study of sentiment analysis tools and binary classification of tweets using rapid miner. *Procedia Computer Science* 125:329–335.

Young, T., D. Hazarika, S. Poria, and E. Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13 (3): 55–75.

Zhang, T., M. Huang, and L. Zhao. 2018. Learning structured representation for text classification via reinforcement learning. In *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence Conference*, 6053–6060.

This is a section of doi:10.7551/mitpress/12200.001.0001

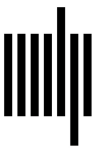
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363