

Authorial Voice Analyzer 1.0 for Mac (released on October 2, 2016)

Hyung-Jo Yoon

Michigan State University

yoonyu9@msu.edu

When using AVA, please cite:

Yoon, H. (in press). Textual voice elements and voice strength in EFL argumentative writing. *Assessing Writing*.

Instructions:

1. Start the program.
2. Locate the input folder that contains the text files (.txt). The folder name should not include special characters (e.g., “, ’, |, \, /, <, >, ^) or spaces.
3. Provide a name of the output file. The output file will be automatically saved in the comma separated value format (.csv), which can be easily imported to Excel or other statistical programs.
4. Start analysis!

Authorial Voice Analyzer (AVA) calculates the normalized frequency values of Hyland’s (2005) metadiscourse categories (e.g., *hedges*, *boosters*, *attitude markers*, *self-mention*, *reader pronouns*, and *directives*) and then calculates composite scores (e.g., stance and engagement) (see the table at the end of this document). For the identification of textual voice elements, AVA relies on the lists of voice elements built through *regular expression* patterns that enable us to

address a range of language variations in the text. For example, one expression included in the hedge list is “in (my|our)\s\w*\s?opinion” in which “|” means *or* and “\s\w*\s?” indicates the possible insertion of an optional word. Accordingly, this one expression counts a wide range of relevant hedging phrases such as *in my opinion, in our opinion, in my personal opinion, in my own opinion, in our common opinion*. Another expression “(?<!in|un)(arguable|arguably)” includes “?<!” that indicates *not following*. This expression thus counts the occurrences of *arguable* and *arguably* as hedging but not *inarguable, unarguable, or unarguably* because a lexical item, *arguable*, does not function as a hedge anymore when following a prefix *un* or *in*.

Hedges and boosters. The hedge list adopted in AVA includes a total of 164 hedging expressions, and the booster list contains a total of 174 expressions. The hedge and booster lists were compiled drawing on the lists and explanations provided in the literature. The items collected were modified in the format of regular expressions.

Attitude markers. The list of attitude markers was based on the existing lists of emotion and attitude words (Hu & Liu, 2004; Mohammad & Turney, 2013), as well as Hyland’s (2005) metadiscourse item list. Specifically, the datasets examined for inclusion in AVA were the lists of positive and negative opinion words (Hu & Liu, 2004) and the National Research Council Canada emotion lexicon containing a list of words related to various emotions and sentiments (Mohammad & Turney, 2013). The final list of attitude markers contains 640 expressions.

Self-mentions and reader pronouns. The categories of self-mentions and reader pronouns include personal pronouns: first person singular pronouns (*I, my, me, and mine*) for self-mentions, and first person plural (*we, our, us, and ours*) and second person pronouns (*you, your, and yours*) for reader pronouns. Given the extensive use of inclusive *we* (and *our, us, and ours*) in

academic writing (particularly in single-authored essays), first person plural pronouns were included in the reader pronoun list.

Directives. For directives, AVA counts necessity modals (e.g., *should, ought, have to*) and predicative necessity-related adjectives controlling a to-clause complement (e.g., *it is critical to, it is essential to*) reported in Hyland (2002). These predicative adjectives of necessity are all constructed with relevant regular expressions. AVA also counts the number of imperative constructions based on sentence parsing by the Stanford parser (Klein & Manning, 2003).

For the lists of hedges, boosters, and attitude markers (i.e., categories with many different items), AVA produces both token (total number of uses in a text) and type (repeated use of each item counted once) values in order to identify potentially different roles of the extensive and varied use of voice elements. For each of the measures, AVA generates two different normalized frequencies that have been extensively used in previous research: (1) frequency per 1,000 words and (2) frequency per T-unit. For T-unit parsing, I used the Tregex pattern reported in Lu (2010).

Table

AVA measures

Category		
Hedges	Token	1) Frequency per 1,000 words 2) Frequency per T-unit
	Type	
Boosters	Token	
	Type	
Attitude markers	Token	
	Type	
Self-mentions	Token	
Reader pronouns	Token	
Directives	Token	
Stance (sum of hedges, boosters, attitude markers, and self-mentions)	Token	
Engagement (sum of reader pronouns and directives)	Token	

References

- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Seattle, Washington.
- Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics*, 23, 215–239.
- Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7, 173–192.
- Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.
- Lu, X. (2010). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14, 3–28.
- Mohammad, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29, 436–465.