Louwerse, M.M., McCarthy, P.M., McNamara, D.S. & Graesser, A.C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

Variation in Language and Cohesion across Written and Spoken Registers

Max M. Louwerse (mlouwers@memphis.edu)
Department of Psychology / Institute for Intelligent Systems, 202 Psychology Building
Memphis. TN 38152

Philip M. McCarthy (pmmccrth@memphis.edu)
Department of English, Patterson 467
Memphis. TN 38152

Danielle S. McNamara (dsmcnamr@memphis.edu)
Department of Psychology, 202 Psychology Building
Memphis. TN 38152

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building
Memphis. TN 38152

Abstract

This paper investigates the variation in cohesion across written and spoken registers. The same method and corpora were used as in Biber's (1988) study on linguistic variation across speech and writing; however instead of focusing on 67 linguistic features that primarily operate at the word level, we compared 236 language and cohesion features at the text-level. Variations in frequencies across these features provided evidence for six dimensions: (1) speech versus writing, (2) informational versus declarative, (3) factual versus situational, (4) topic consistency versus topic variation, (5) elaborative versus constrained, (6) narrative versus non-narrative. Our cohesion and linguistic analysis showed most variation in speech and writing, whereas the linguistic feature analysis operating at the word level did not yield any difference.

Introduction

One way to investigate similarities and differences between speech and writing is by using corpus linguistic methods. The most common and largest investigation of this kind is Biber (1988). Biber used 23 spoken and written registers. These registers are language varieties mediated by social situations and are similar to genres. Biber took these registers from the Lancaster-Oslo-Bergen (LOB) corpus and the London-Lund corpus, and computed the frequency of 67 linguistic features in these registers (see Table 1 for an overview of registers).

The linguistic features used for Biber's analysis primarily operate at the word level (e.g., parts-of-speech) and can be categorized as (1) tense and aspect markers, (2) place and time adverbials, (3) pronouns and pro-verbs, (4) questions, (5) nominal forms, (6) passives, (7) stative forms, (8) subordination features, (9) prepositional phrases, adjectives and adverbs, (10) lexical specificity, (11) lexical classes, (12) modals, (13) specialized verb classes, (14) reduced forms and dispreferred structures, and (15) coordinations and negations.

Table 1. The 23 registers used in Biber (1988)

Corpus	Register
Lancaster-Oslo-Bergen corpus	Press reportage, editorials, press reviews, religion, skills and hobbies, popular lore, biographies, official documents, academic prose, general fiction, mystery fiction, science fiction, adventure fiction, romantic fiction, humor
London-Lund corpus	Face-to-face conversation, telephone conversation, public conversations, debates, and interviews, broadcast, spontaneous speeches, planned speeches
(Additional)	Personal letters, professional letters

In Biber's study the normalized frequencies of these features in each of the registers were then entered in a factor analysis, from which six factors emerged. These factors can be seen as dimensions on which registers can be placed. Biber's analysis showed that no single dimension comprised a difference between speech and writing; As such, Biber defined the sets of relations among texts as follows:

1. Involved versus informational production
2. Narrative versus non-narrative concerns
3. Explicit versus situation dependent reference
4. Overt expression of persuasion
5. Abstract versus non-abstract information
6. On-line informational elaboration.

For example, registers such as romantic fiction, mystery fiction and science fiction were positioned high on the second dimension (narrative); whereas registers such as academic prose, official documents, hobbies, and broadcasts scored low (non-narrative).

Biber's (1988) study and the multi-feature, multidimensional approach have become a standard in corpus linguistics (McEnery, 2003), leading to various extensions (Biber, Conrad & Reppen, 1998; Conrad &

Biber, 2001), as well as to assessments of the validity, stability, and meaningfulness of the approach and its findings (Lee, 2004).

Measuring cohesion

Texts obviously consist of a large variety of linguistic features, many of which can be identified at a word level (e.g. morpho-semantics, syntactic category, frequency). Biber's study has shown that these linguistic features are powerful determiners of similarities and differences between registers. But despite these impressive results, the theoretical question that remains lurking is to what extent these linguistic features fully capture the nature of a text and thereby the nature of a register.

Although linguistic features operating at the word level may identify several register characteristics, we also know that one of the key features of a text is that it is not just a concatenation of words and sentences. Instead, there is a structure in the text that glues the various text components together. In comprehending the text, the reader or listener constructs a coherent, mental representation of the situations which have been cohesively described by the text. We have used the term "coherence" for the representational relationships and "cohesion" for the textual indications through which coherent representations should be built (Louwerse & Graesser, 2004). Cohesion, it should be noted, cannot be captured only by linguistic features at the word level. Instead, cohesion stretches to the inter-clause, inter-sentence and inter-paragraph level.

But if a key component in the nature of text consists of cohesion, a practical issue related to the theoretical question needs to be addressed. Linguistic features that operate at a word level can currently be reliably identified by regular expressions, part-of-speech taggers, and syntactic parsers. However, there is the practical question of whether automated techniques can also capture the cohesion of text. Recent landmark progress in computational linguistics has indeed allowed us to go far beyond surface level components into automating deeper and global levels of text and language analysis (Jurafsky & Martin, 2001). This progress has resulted in the cohesion and coherence measurement tool Coh-Metrix.

Coh-Metrix

Coh-Metrix was initially developed in order to replace readability formulas that exclusively focus on simple and shallow metrics. Instead, Coh-Metrix is sensitive to a broader profile of language and cohesion characteristics. It analyzes texts on 236 types of cohesion relations and measures of language, text, and readability (McNaramara, Louwerse, & Graesser, 2002; Graesser, McNamara, & Louwerse, in press). For this paper, we will only focus on the textual features (cohesion) of the tool.

The modules of Coh-Metrix use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics. For example, the MRC database (Coltheart, 1981) is used for psycholinguistic information; WordNet (Miller, Beckwith, Fellbaum, Gross & Miller) for underlying lexical concepts; Latent Semantic Analysis (Landauer & Dumais) for the semantic similarities between words, sentences and paragraphs; the ApplePie parser (Sekine & Grishman, 1995) and the Brill (1995) part-of-speech tagger for a variety of syntactic categories.

Spatial restrictions do not make it possible to discuss all of the measures Coh-Metrix makes available. As such, a brief summary of key measures will have to suffice, whereas Graesser et al. (in press) has a more complete overview.

1. *Word information* includes word familiarity, word concreteness, word imageability, meaningfulness, and age of acquisition.
2. *Word frequency* includes four corpora-based standards: CELEX from the Dutch Centre for Lexical Information (Baayen, Piepenbrock & Van Rijn, 1993); the Kucera-Francis norms (Francis & Kucera, 1982); Thorndike-Lorge norms (Thorndike & Lorge, 1942) and the Brown norms (Brown, 1984).
3. *Part of speech* categories are adopted from the Penn Treebank (Marcus et al., 1993) and the Brill (1995) POS tagger.
4. *Pronoun density* is computed by taking the ratio of pronouns and nouns.
5. *Logical operators* are the incidence score of logical operators *or*, *and*, *not*, and *if-then* phrases
6. *Interclausal relationships* are the additive, temporal and causal cohesion based on connectives between clauses. These can be positive (extending) and negative (adversative), as outlined in Louwerse (2001).
7. *Type-token ratio* refers to the number of unique words divided by the number of tokens of the words.
8. *Polysemy* and *hypernym*: Polysemy is measured as the number of senses of a word in WordNet; whereas the hypernym count is defined as the number of levels in a conceptual taxonomic hierarchy that is superordinate to a word.
9. *Concept clarity* is a composite of multiple factors that measure ambiguity, vagueness, and abstractness.
10. *Syntactic complexity* refers to the noun-phrase density, the mean number of modifiers per noun phrase, the number of high-level constituents per word and the incidence of word classes that signal logical or analytical difficulty.
11. *Readability* scores are computed according to the Flesch Reading Ease formula and the Flesch-Kincaid Grade Level formula, the standard readability formulas.
12. *Coreference*: Three forms of coreference between sentences are computed, namely noun overlap between sentences, stem overlap, and stem-noun overlap.
13. *Causal cohesion* is interpreted as the ratio of causal particles to causal verbs.
14. *Latent semantic analysis*: LSA is a statistical, corpus based, technique used to represent world knowledge that computes similarity comparisons for terms and documents by taking advantage of word co-occurrences. LSA scores

can be computed for sentence to paragraph, sentence to text, paragraph to paragraph and paragraph to text. These measures can be used for measuring the local and global cohesion of the text (see Kintsch, 2002; Landauer & Dumais, 1997).

The advantage of the use of this wide range of computational linguistic tools is that Coh-Metrix is sensitive to variations in language, discourse, and cohesion. Such an analysis may not only help us to determine text difficulty, but may also help us with determining variations across registers.

Multi-dimensional study on cohesion

In our multi-feature, multi-dimensional approach we carefully followed Biber's study to allow comparison of his findings. We used the same fifteen written registers from the LOB corpus and the same six spoken registers from the London-Lund corpus. Two further non-published registers (professional and personal letters), which Biber had generated himself, were substituted with the *Compilation of Messages and Letters of the Presidents* Richardson (2003/1801) and *The Upton Letters* from Christopher Benson (1905), both downloaded from the Gutenberg text archives.

All textual coding other than alphanumeric characters and punctuation was removed. The 23 spoken and written registers were then processed through Coh-Metrix and the normalized frequencies for each of the 236 cohesion, language, and discourse features were saved. We followed Biber's approach in standardizing all frequencies to a mean of 0.0 and a SD of 1.0 and entered them in a factor analysis using the Promax rotation. Whereas Biber used a principal factor analysis to account for the shared variance, we opted for a principal component analysis to account for all the variances. Loadings with an absolute value of less than .35 were excluded from the analysis (Biber, 1988; Comrey & Lee, 1992). The scree plot of eigenvalues, illustrating the amount of variance accounted for by each factor, showed a clear break after six factors, explaining 88.3% of the total variance.

In order to translate the factor scores to the registers, we followed Biber by adding together the standardized scores of all linguistic features responsible for a factor in a particular text. This provides a measure of register's salience on a particular dimension given the presence of the linguistic features in that register. We deviated from Biber's approach in one important way: Whereas Biber removed the linguistic features from subsequent factors once it was used by a previous factor, we preferred to include these linguistic features in additional factors to account for the interactions between language, discourse, and cohesion features.

Dimensions

Space limitations dictate us to summarize the findings by presenting tables in which we have translated the ratio scale to an ordinal scale, thereby not serving full justice to the actual differences between the 23 registers.

Dimension 1: Speech versus writing. This dimension significantly accounts for 53.5% of the variance, (F (1, 22) = 35.61, p < .001, MSE = .721). When looking at the grouping of the registers, it immediately becomes apparent that spoken registers are distinct from written registers. In addition, the registers clearly show *the degree* to which the registers are speech-dependent. For example, fiction includes, or more closely reflects, spoken discourse, whereas this is far less likely to be the case with press reviews or professional letters.

The linguistic features with positive loadings are presented in the first data row of the table, signifying the higher presence in the register. They consist of concreteness, imageability, meaningfulness, polysemy, and frequency in the spoken discourse. Negative loadings relate to ambiguous quantification, pronoun density, argument overlap, and semantic similarity between sentences and paragraphs. Registers with a higher score on this dimension (like public conversations and face-to-face conversations) are characterized by frequent occurrences of concrete, imaginable, and meaningful language, together with higher pronoun density and ambiguous quantification. At the same time, occurrences of argument overlap and semantic similarities between text units are less prevalent. Registers with negative scores, presented in the second data row, have the opposite characteristics.

In Table 2 results are given for the Dimension 1. The first column presents the registers ranked by total scores and the second column presents the linguistic features ranked by factor loadings. Row separators mark the difference between positive and negative factor loadings. The same format is used for the remaining five tables representing the remaining five dimensions.

Table 2: Distribution registers and summary Coh-Metrix Dimension 1 (speech versus writing)

public conversations, face to face conversation, spontaneous speeches, telephone conversations, planned speeches, broadcast, mystery fiction	frequency, concreteness, imageability, meaningfulness, polysemy, Flesch Reading Ease, ambiguous quantification, pronoun density, higher level constituents per word, abstract nouns, hypernym, polysemy
Personal letters, general fiction, romantic fiction, religion, adventure fiction, skills and hobbies, official documents, humor, academic prose, editorials, popular lore, biographies, science fiction, press reportage, press reviews, professional letters	LSA sentence to sentence, ratio of causal particles to causal verbs, LSA paragraph to paragraph, paragraph to text, vague adverbs, type-token ratio for nouns, concreteness, argument overlap, average paragraph length, age of acquisition, average syllables per word, mean number of modifiers per noun-phrase, stem overlap, Flesch Kincaid Grade Level

Dimension 2: Informational versus declarative. The second dimension accounts for 16.3 % of the variance, but without significant differences between the registers (F (1, 22) = .93, p = .56, MSE = .968). This dimension shows many similarities with Biber's Dimension 6, with the majority of the registers positioned similarly along the axis in both studies. Biber tentatively labeled this "on-line informational elaboration marking stance" with registers such as planned speeches and public conversations being informational in focus and conveying the speaker's attitudes and beliefs. We come to a similar conclusion, interpreting the difference as informational and subjective versus declarative and objective. Informational registers are characterized by a higher occurrence of temporal cohesion, imageability, and concreteness, but a low occurrence of causality, whereas the opposite characterizes declarative registers.

Table 3: Distribution registers and summary Coh-Metrix Dimension 2 (informational versus declarative)

mystery fiction, religion, skills and hobbies, romantic fiction, spontaneous speeches, official documents, general fiction, popular lore, telephone conversations, adventure fiction, biographies, face to face conversation, broadcast, humor	Positive temporal connectives, polysemy (adjectives), meaningfulness, LSA paragraph to paragraph, familiarity, LSA sentence to sentence, negative temporal connectives, paragraph length, argument overlap, LSA sentence to paragraph, LSA paragraph to text, ratio of causal particles to causal verbs, LSA paragraph to paragraph, type-token ratio for nouns, LSA paragraph to text, imageability, concreteness, LSA sentence to sentence, LSA sentence to sentence, concreteness
planned speeches, public conversations, academic prose, personal letters, editorials, science fiction, professional letters, press reportage, press reviews	Negative causal connectives, frequency, (verbs), causal particles, average syllables per word, positive causal connectives, age of acquisition

Dimension 3: Factual versus situational. This dimension, explaining 7.7 % of the variance, shows similarities with Biber's Dimension 3: "explicit versus situation-dependent reference." Biber argues that the situation-dependent site of the dimension refers to places and times outside of the text (imaginary and real world), whereas the opposite side of the dimension has registers with elaborated explicit reference. Although we do not find evidence for the time or place reference, we do find a higher frequency of imageability and a lower frequency of clarification and causal connectives, with the opposite trend evident for the registers on the factual side of the dimension (F(1, 22) = 5.88, p < .001, MSE = .871). The labels "factual" and "situational" refer to the presentation, rather than the content. For instance, religion is located high on the factual dimension because this register is generally presented as factual. On the other hand, press reviews, reportages and fiction are presented in a less transparent way, often requiring the reader to imagine a situation.

Table 4: Distribution registers and summary Coh-Metrix Dimension 3 (factual versus situational)

academic prose, official documents, religion, skills and hobbies, popular lore, biographies, spontaneous speeches, personal letters, face to face conversation	Clarification connectives, causal particles, negative causal connectives, noun overlap, ratio of causal particles to causal verbs, vague adjectives, negative additive connectives, positive causal connectives, ambiguous quantification, argument overlap, vague verbs, vague nouns,
telephone conversations, humor, editorials, public conversations, press reviews, press reportage, professional letters, planned speeches, general fiction, broadcast, mystery fiction, romantic fiction, adventure fiction, science fiction	polysemy, imageability, causal verbs, mean hypernym of verbs

Dimension 4: Topic consistency versus topic variation. This dimension explains 4.6% of the variance with significant differences between the registers (F(1, 22) = 3.76, p < .001, MSE = .870). It marks the consistency of topics across and within instances of a particular register. For instance, personal and professional letters often have a similar set of topics that are used, as do biographies and spontaneous speeches. Face-to-face conversations, interviews, public debates, press reportages and editorials on the other hand, have more topics and are less predictable, often switching between different instances. In the registers located high in the topic consistency (e.g. personal letters and professional letters), semantic similarities marking global cohesion and local cohesion are higher, but noun density and type-token ratio are lower than the topic variation registers (e.g., reportages and editorials).

Table 5: Distribution registers and summary Coh-Metrix Dimension 4 (topic consistency versus topic variation)

personal letters, spontaneous speeches, professional letters, biographies, broadcast, academic prose, religion, official documents, skills and hobbies, romantic fiction, mystery fiction	frequency conditionals, frequency negations, causal verbs, positive additive connectives, polysemy, LSA paragraph to paragraph, positive causal connectives, LSA sentence to text, LSA paragraph to paragraph, LSA paragraph to text
telephone conversations, general fiction, press reviews, popular lore, planned speeches, humor, adventure fiction, science fiction, face to face conversation, public conversations, press reportage, editorials	type-token ratio, noun density

Dimension 5: Elaborative versus constrained. This dimension is harder to interpret and explains only 3.7 % of the variance. Differences between registers ($F(1, 22) = 3.55$, $p < .001$, $MSE = .866$) suggest that personal letters and press reviews for instance are more opinion-based and have a closer distance between writer and reader, whereas professional letters and press reportages, are more fact and evidence driven. It is almost as if there is more space in personal letters and press reviews to compare ideas. This conclusion is supported by the factor loadings of the linguistic features, which show a prominent role for additive cohesion, vague adjectives and adverbs, along with a high type-token ratio and an accompanying low semantic similarity in the case of the personal letters and the press reviews. It is as if many ideas are juxtaposed within these registers.

Table 6: Distribution registers and summary Coh-Metrix Dimension 5 (elaborative versus constrained)

personal letters, press reviews, biographies, skills and hobbies, religion, humor, popular lore, academic prose, official documents, editorials, general fiction	type-token ratio, negative additive connectives, vague adjectives, vague verbs, positive additive connectives
mystery fiction, science fiction, romantic fiction, telephone conversations, broadcast, adventure fiction, face to face conversation, press reportage, planned speeches, public conversations, spontaneous speeches, professional letters	LSA paragraph to text, LSA paragraph to paragraph, LSA sentence to text

Dimension 6: Narrative versus non-narrative Although significant differences were found between registers (F (1, 22) = 1.64, p = .037, MSE = .991) only 2.5 % of the variance was accounted for by this dimension. Dimension 6 is virtually identical to Biber's Dimension 2. In registers such as fiction and biographies, a narration of events is prominent, whereas narration is less obvious in press reviews and professional letters. Linguistic features like temporal connectives are primarily responsible for this dimension. Despite the similarities with Biber's dimension, there are also some important differences. For instance, in our findings, science fiction scores low on narrative but face-to-face conversations score high, whereas in Biber's analysis the opposite is the case. The clear similarities between the two studies (e.g., the clustering of the fiction texts) support this interpretation of the dimension.

Table 7: Distribution registers and summary Coh-Metrix Dimension 6 (narrative versus non-narrative)

Romantic fiction, mystery fiction, face to face conversation, general fiction, adventure fiction, biographies, religion, public conversations, telephone conversations, official documents	ambiguous temporal relation, vague nouns, positive connectives, temporal connectives
Editorials, academic prose, press reportage, skills and hobbies, humor, spontaneous speeches, popular lore, personal letters, broadcast, planned speeches, science fiction, professional letters, press reviews	LSA sentence to text, LSA paragraph to text, LSA sentence to sentence

Discussion and conclusion

The present study has investigated the multi-feature, multi-dimensional corpus linguistic approach initially outlined by Biber (1988). We have used the same corpora and the same methods as Biber, but instead of including linguistic features that primarily operate at the word level, we have included a large variety of language, discourse and cohesion features. These features ranged from the word level, to sentence, paragraph and discourse level. Six dimensions emerged from a factor analysis: (1) speech versus writing, (2) informational versus declarative, (3) factual versus situational, (4) topic consistency versus topic variation, (5) elaborative versus constrained, (6) narrative versus non-narrative. Three of these dimensions (Dimension 2, 3 and 6) show strong similarities with the distributions of registers as well as the interpretations of dimensions in Biber's study.

Results showed one crucial difference with Biber's finding. Whereas Biber was not able to find one single dimension that determined the difference between speech and writing, we found a very prominent difference in linguistic features between spoken and written discourse (Dimension 1). The most plausible explanation for this

result is the contrast between Biber's focus on the linguistic features operating at the word level and our study which included a much wider range of language and discourse characteristics that we have called cohesion.

Acknowledgments

The research was supported by the Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Baayen, R. H., R. Piepenbrock, and H. van Rijn (Eds.) (1993). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.

Benson, A.C. (1905). *The Upton letters*. Retrieved January 2004 from the Project Gutenberg Text Archives.

Biber, D. (1988). Linguistic features: algorithms and functions in Variation across speech and writing. Cambridge: Cambridge University Press.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543-566.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioral Research Methods Instrumentation and Computers*, 16, 502-532.

Comrey, A. L. & Lee, H. B. (1992). A first course in factor analysis. Hillsdale, NJ: Lawrence Erlbaum.

Conrad, S. & Biber, D. (2001). *Variation in English: Multi-Dimensional Studies*. Harlow: Longman

Francis, W.N., & Kucera, N. (1982). *Frequency analysis of English usage*. Houghton-Mifflin.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. (in press). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*.

International Computer Archive of Modern and Medieval English (2000). *Lancaster/Oslo/Bergen Corpus of British English* (CD-ROM).

International Computer Archive of Modern and Medieval English (2000). *The London-Lund Corpus of Spoken English* (CD-ROM).

Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice-Hall.

Kintsch, W. (2002) On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.) *Thematics: Interdisciplinary Studies*. Amsterdam: Benjamins.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Lee, D. Y. W. (2004). *Modeling variation in spoken and written English*. London/New York: Routledge.

Louwerse, M.M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 291–315.

Louwerse, M.M. & Graesser, A.C. (2004). Coherence in discourse. In Strazny, P. (ed.), *Encyclopedia of linguistics*. Chicago: Fitzroy Dearborn.

McEnery, T. (2003). Corpus linguistics. In: R. Mitkov (Ed.), *The Oxford encyclopedia of computational linguistics*. Oxford: Oxford University Press.

McNamara, D.S., Louwerse, M.M. & Graesser, A.C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Miller, G. A., Beckwith, R., Fellbaum, C. , Gross, D. & Miller, K. (1990). *Five Papers on WordNet*. *Special Issue of the International Journal of Lexicography, 3*.

Richardson, J.D. (2003/1801). *Compilation of the Messages and Papers of the Presidents* (Vol. 1, John Adams). Retrieved January 2004 from the Project Gutenberg Text Archives.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two nonterminals. *Fourth International Workshop on Parsing Technology*.

Thorndike, E.L. and Lorge, I. (1944). The *teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.