A Comparative Study on English and Chinese Word Uses with LIWC

Haiying Li, Zhiqiang Cai, Arthur C. Graesser, Ying Duan

University of Memphis

Department of Psychology

hli5@memphis.edu

Abstract

This paper compared the linguistic and psychological word uses in English and Chinese languages with LIWC (Linguistic Inquiry and Word Count) programs. A Principal Component Analysis uncovered six linguistic and psychological components, among which five components were significantly correlated. The correlated components were ranked as Negative Valence (r=.92), Embodiment (r=.88), Narrative (r=.68), Achievement (r=.65), and Social Relation (r=.64). However, the results showed the order of the representative features differs in two languages and certain word categories co-occurred with different components in English and Chinese. The differences were interpreted from the perspective of distinctive eastern and western cultures.

Introduction

The Language of Thought Hypothesis (LOTH, Aydede 1999) assumed that thought was tokenized by the syntactic and semantic representations of language. Furthermore, features of the language highly represent the individual's or the group's characteristics and styles from the perspectives of phonology (Mulac, Hanley, and Prigge 1974), lexicon (Bradac, Konsky, and Davies 1976), and words (Wood, Yamauchi, and Bradac 1971). The word as an essential component of language plays a crucial role in the communication processes (Vick and Wood 1969). Bradac, Bowers, and Courtright (1982) specified lexical variations such as intensity, immediacy and diversity had some causal relationships with cognitive and emotional states.

The theory of linguistic relativity, also called Sapir-Whorf hypothesis, or Whorfianism (Sapir 1921, Whorf, 1956) claimed that the linguistic features determined or influenced speakers' conceptualization or cognitive categories. Consequently, the research discovered the speakers' or writers' usages of words were related to their psychological processes. Research on different categories of words such as content and functional words (Pennebaker and Chung 2009, 2011), and social and psychological meaning of words (Tausczik and Pennebaker 2010) demonstrates that the choices of the specific category of words reflects speakers' or writers' psychological processes, emotions, and social relationships.

In the past decades, the tasks for manually analyzing texts were time-consuming, expensively, and less efficiently with the lower inner-rater reliability (Tausczik and Pennebaker 2010). Newly emerging interdisciplinary subjects such as computational linguistics, corpora linguistics, and discourse processes promoted the development and innovation of advanced technologies in the domain of language and discourse processes (Graesser and McNamara 2010). With over two decades' experiments, Pennebaker, Booth and Francis (2007) issued a computer-based text analysis program called a Linguistic Inquiry and Word Count (LIWC) program that automatically detects the links between the words and the psychology-relevant categories.

The LIWC tool provides a text processing module and a dictionary, and counts the percentage of words mapping a specific dimension of a language to study speakers' or writers' psychological states efficiently and economically (Pennebaker and Graybeal 2001). The LIWC program generated 80 word categories representing the various types of words such as personal pronouns, verb, tenses etc. in linguistic categories, and cognitive, affective, biological and social, spatial and temproral words in psychological categories (Pennebaker, Booth, and Francis 2007).

Subsequently, the Chinese LIWC dictionary was developed by National Taiwan University of Science and Technology based on the dictionary in the 2007 English LIWC program (Huang et al in press). The Chinese LIWC dictionary adopted the traditional Chinese characters prevalent outside mainland China, such as Taiwan, Hongkong, Macau, and overseas Chinese communities (Keller 1997). Our Memphis group converted the traditional Chinese characters to the simplified Chinese characters. The Chinese LIWC dictionary embedded in LIWC program generated 72 categories, representing the linguistic and psychological words used in Chinese.

As the categories of words in Chinese LIWC dictionary were constructed based on those in English LIWC dictionary; therefore, LIWC program is used to compare the word use in the Chinese and English languages represented by the culture and ideology.

This paper compares the categories of words with English and Chinese LIWC programs with large corpora using

the method of the Principal Component Analysis. The goal of this paper is to answer two questions: 1) whether the representative linguistic and psychological features are the same in Chinese and English, and 2) to what extent the English and Chinese languages share the common linguistic and psychological categories of word use and to what extent they differ.

Method

Corpora

The TASA (Touchstone Applied Science Associates, Inc.) corpus is adopted as the English corpus. It encompasses academic textbooks for students from kindergarten to the first year of college in the United States (Graesser and McNamara 2010; Zeno et al 1995). Documents in TASA are pre-processed in the form of the text segments or excerpts at varied levels of difficulties and with different genres. The Chinese corpus was collected according to the same genres in TASA (See Table 1) and its readability ranges from elementary to college levels. However, the documents in the Chinese corpus were not segmented into excerpts as TASA, but keep its original length.

		N(%)	MW	Std.W	SL	Std.SL
Econo-	E	1504(4)	283	21	18	4
my	C	513(11)	4795	6429	28	7
Lan-	E	16044(43)	286	25	17	10
guage	C	2050(44)	5222	4060	31	21
Arts						
Social	E	10501(28)	293	25	20	9
Studies	C	1878(40)	5892	5464	26	8
Science	E	6715(18)	282	22	16	6
	C	238(5)	6791	7361	29	8
Other	E	2887(8)	289	23	20	7
	C	X				
Total	E	37651(100)	288	25	18	9
	C	4679(100)	5383	5295	27	17

Table 1: Descriptive statistics of words and sentence length in TASA and Chinese corpora.

Notes: N=numbers of documents; %=percentage of documents in the English and Chinese corpora separately; MW=means of words in each document; Std.W= standard deviation of means of words in each document; SL=sentence length (words per sentence); Std.SL=standard deviation of sentence length.

Both corpora cover the same genres, except unknown genre labeled as "Other" in TASA corpus. The majority of the genre is languages arts in both languages up to 43-44%, whereas the Chinese includes more documents in social studies (40% vs. 28%) and economy (11% vs. 4%) than those in English, but fewer scientific documents than English (5% vs. 18%). The unknown or miscellaneous documents are not included in the Chinese corpus, but they are in TASA (0 vs. 8%). The total amount of documents is 4,679 in Chinese, but 37,651 in English. This unequal amount of documents, however, is compensated by the number of words in two corpora.

The Chinese corpus has 25,184,754 words in 4,679 documents and the mean of amount of words in each document is 5,383 words with 5,295 standard deviations averagely. The English corpus has 10,829,757 words in 37,651 documents, and the mean of words in each document is 288 with 25 standard deviations averagely. Thus, the Chinese corpus has eight times fewer documents, but more than twice amount of words than those in English corpus. This is because the words in each document in the Chinese corpus are averagely over 18 times more than those in TASA.

Another reason is that Chinese texts are all entire and complete chapters from books, whereas TASA documents are short random excerpts from the complete texts (Zeno et al 1995). Even if the document lengths are apparently distinct, the sentence lengths show more satisfactory expectation. In TASA, the number of words per sentence is 18 with 8.9 standard deviations, and 27 words per sentence with 16.8 standard deviations in the Chinese corpus. Meanwhile, the LIWC program computes the frequency of word use in each category with percentage; thus, the unequal number of words will not much influence the results.

As both corpora are large and have diverse genres, they enable to represent the typical word uses in both languages. The English and Chinese corpora, therefore, are comparable.

Procedure

The 2007 English LIWC program including a text processing module and an internal English dictionary, counts the percentage of 80 word categories for given texts efficiently in several minutes (Pennebaker, Booth, and Francis 2007), which encompass linguistic word categories such as diverse parts of speech and tenses, and psychological categories such as social processes, affective processes, cognitive processes, perceptual processes, biological processes, relativity, and new current concerns (Pennebaker, Booth, and Francis 2007).

The Chinese LIWC program applied the same text processing model as the 2007 English LIWC and the simplified Chinese character dictionary, which generates 72 categories of the Chinese words (Huang et al in press) including the linguistic and psychological word categories.

The English 2007 LIWC program was performed on the English corpus, and the Chinese LIWC Chinese corpus. 80 categories of words in English and 72 in Chinese were obtained from two corpora as the English and Chinese output respectively. One category "word per sentence (WPS)" was in the English LIWC output, but not in Chinese. Thus, WPS in Chinese computed by the Chinese LSA tool (Graesser et al 2007) was supplemented in Chinese dataset.

Therefore, 61 identical linguistic and psychological word categories in both English and Chinese were adopted as the independent variables (See Table 2). 19 word categories in English and 12 categories in Chinese were removed from the data set due to various reasons.

One of the reasons is that the categories of words represent the unique and exclusive linguistic features merely in

English or Chinese. For example, word categories "article" (*a*, *an*, or *the*) and "six-string words" (The number of letters in a word is no less than six) are unique in English, but not in Chinese, because Chinese is a language of strokes symbolized by ideograms and the pictograms, not the alphabetic language.

Another type of removed English variables included 12 punctuation categories and one category of the dictionary word, because these 13 categories were generated in English LIWC, but not in Chinese.

Matched (61)	Removed	
	English(19)	Chinese(12)
funct, pronoun, ppron, i, we, you, shehe, they, ipron, verb, auxverb, adverb, preps, conj, negate, quant, number, WPS, swear, social, family, friend, humans, affect, posemo, negemo, anx, anger, sad, cogmech, insight, cause, discrep, tentat, certain, inhib, incl, excl, percept, see, hear, feel, bio, body, health, sexual, ingest, relativ, motion, space, time, work, achieve, leisure, home, money, relig, death, assent, nonfl, filler	article, Sixltr, Period, Comma, Colon, Semic, Qmark, Exclam, Dash, Quote, Apostro, Parenth, OtherP, AllPct, Dic, WC, past, present, future	TenseM, ProgM, PastM, PresesntM, FutureM, Youpl, WC, PrepEnd, SpecArt, QuanUnit, Interjuction, MultiFun

Table 2: Matched and removed categories in LIWC output

The reason for removing the categories of tenses in both languages is that the representations of syntax and tenses are different in the two languages. Therefore, categories relevant to tenses (present, past and future tenses in English, and tense marker, continuation marker, present marker, past marker and future marker in Chinese) were removed from both the English and Chinese datasets, because of the grammatical and syntactic discrimination in English and Chinese tense communication. In English, the tenses are morphologically indicated by the verb inflections (Bache 2008). In Chinese, the tenses are expressed through the aspect- or tense-particles serving as the time adverbials, so people do not need to depend on covert semantic features aided by a tense node to interpret time (Lin 2005).

Another removed category in Chinese is "you", of which the plural form (nǐmen/nínmen) differs from "you" single form (nǐ/nín). Nǐmen functions similarly as English plural "you"; nínmen is used to show deference or respect to the audience (Chao 1956). However, in English, the word "you" represents both the single and the plural forms.

Moreover, categories of words only generated by the Chinese LIWC but not in English were also removed from the dataset, which includes categories of preposition phrase end, specifying article, quantity unit, interjections, and multiple functions.

Only one removed category occurring in both languages was "word count", because its loading was very low in each language in the initial exploratory Principal Component Analysis.

Thus, the entire 61 identical categories in both English and Chinese respectively were utilized as the independent variables and run through the Principal Component Analysis.

Results and Discussions

A Principal component analysis (PCA) was performed with the Promax rotation with an absolute value of loading more than .10. Based on the screeplots in English and Chinese dataset, six components that apparently represented the characteristics of word uses in both languages were extracted from 61 categories. These six components instead of using the myriad 61 categories uncover fewer unobservable and comprehensive features from the dataset in order to find out more meaningful and interpretable phenomena representing the characteristics of two languages.

The PCA results showed that six components explained 39.3% of the total variance in English and 56.3% in Chinese. The six components were ranked and labeled as Narrative (13.3%), Achievement (8.8%), Social Relation (5.5%), Negative Valence (4.7%), Embodiment (3.6%) and a hybrid component (3.4%) in English; and Narrative (25.2%), Social Relation (12.1%), Space and Time (5.7%), Embodiment (5.1%), Negative Valence (4.6%) and Achievement (3.6%) in Chinese according to the representative characteristics of linguistic and psychological word categories in each component.

Correlation

Pearson correlations based on the loadings in PCA showed five components have the significant correlation between Chinese and English (α>.01), except component three in Chinese and six in English (See Table 3). 1-6 Roman numerals refer to the first component to the sixth in the extracted sequence. E and C respectively represent English and Chinese. For example: E1 means the first component in English, and C1 means the first component in Chinese.

	E1	E2	E3	E4	E5	E6
C1	.68**	0.23	-0.23	-0.23	-0.16	.33**
C2	0.15	-.35**	.64**	-0.14	-0.04	-0.24
C3	-.43**	-.33**	-0.14	-0.18	-0.09	0.02
C4	-0.15	-0.08	-.26*	0.04	.88**	-0.14
C5	-0.17	0.05	-0.06	.92**	0.03	-0.11
C6	0.04	.65**	0.06	-0.10	-0.13	-0.14

*Table 3. The correlation of components in English and Chinese. Notes: **. Correlations are significant at the .01 level (2-tailed). *. Correlation is significant at the .05 level (2-tailed).*

Component one (Narrative) in Chinese was substantially correlated with component one in English (.68**). The correlation of component two (Social Relation) in Chinese was .64** with the third component in English. And component four (Embodiment) in Chinese was correlated much higher to .88** with component five in English. Component five (Negative Valence) in Chinese had the correlation of .92** with component four in English. The sixth component (Achievement) in Chinese was correlated with component two in English by .65**.

These substantially correlated components provide the evidence of automatically processing the discourses in large corpora (Graesser, Gernsbacher, and Goldman 2003) for the comparative study on English and Chinese. In addition, the results indicate two languages share such common linguistic and psychological features as negative emotions, biological processes, narrative, achievements and social relation.

Categories in Each Component

Negative Valence. The component of Negative Valence in Chinese had the highest correlation with that in Chinese by .92**. In this component, all the loaded categories were identical in both languages (See Table 4) such as negative emotion, anger, sad, anxiety, death, affective processes and swear. In both languages word category "swear" delivering curses like "hell, suck, shit" etc. was loaded in this component besides the common negative emotion words. In addition, the general emotional category "affect" (affective processes) seems to more likely occur with negative emotions, but not positive emotions in both languages.

	Chinese	English
Identical	negemo; anger; affect; death; sad; swear; anx	negemo; anger; affect; sad; anx; death; swear
Non-Identical	X	X

Table 4: Categories in Negative Valence
Note: negemo, negative emotion; anx, anxiety

Embodiment. The second highly correlated component (.88**) was Embodiment, which includes the identical categories such as "biological processes, body, health, and ingestion" relevant to the embodiment, and one perceptual category "feel" in both language (See Table 5). Thus, the perceptual category "feel" representing tactile sensory more likely occur with biological and physiological category.

	Chinese	English
Identical	bio; body; health; ingest; feel	bio; body; ingest; health; feel
Non-Identical	sexual	X

Table 5: Categories in Embodiment
Note: bio, biological processes

However, one biological category "sexual" was loaded in Chinese, but not in English. "Sexual" category more likely co-occurred with the categories of social processes, positive emotion, leisure, assent, and third person pronouns in English. This distinction perhaps results from the restriction of the private and personal topic "sex" to physiological topic in Chinese culture, but more open in English.

Narrative. The component of Narrative in Chinese was correlated with that in English by .68**. Nine identical categories were loaded such as linguistic categories "impersonal pronouns, adverbs, auxiliary verbs and the total functional words", and cognitive categories "cognitive processes, exclusive, tentative, discrepancy, and certainty". Therefore, Narrative component consists of the majority of cognitive words with the co-occurring functional words, adverbs and auxiliary verbs.

Nevertheless, non-identical categories demonstrated the distinction in the word use in narration. In Chinese, the more likely categories in narrative occur, such as "conjunctions, preposition, quantifiers, we, and they", the less likely "numbers and sentence length" occur. Conversely, the more likely occurring linguistic categories in English were "verbs, negations and you". Therefore, the first personal pronoun "we" occur in Chinese, while the second personal pronoun "you" occur in English in Narrative.

In addition, the cognitive categories like "causation and inclusive" moderately occur in Chinese, whereas "insight" in English. Moreover, category of positive emotion more likely occurs in Narrative in Chinese, but in Achievement in English. It is possible that the Chinese culture emphasizes the implicitness and modesty, which probably causes that people choose positive words in narration (Chao 1956).

	Chinese	English
Identical	ipron; funct; cogmech; excl; tentat; discrep; certain; adverb; auxverb	auxverb; cogmech; funct; excl; tentat; ipron; discrep; adverb; certain
Non-Identical	conj; quant; preps; cause; incl; we; filler; posemo; they; number (-); WPS (-)	verb; negate; insight; you

Table 6: Categories in Narrative
Note: ipron, impersonal pronoun; funct, Total functional words; cogmech, cognitive processes; excl, exclusive; tentat, tentative; discrep, discrepancy; certain, certainty; auxverb, auxiliary verbs; conj, conjunctions; quant, quantifiers; preps, prepositions; cause, causation; incl, inclusive; posemo, positive emotion; WPS, words per sentence; negate, negations; (-), the negative loading (the same below)

Achievement. The component of Achievement in Chinese had a correlation of .65** with that in English. The identical categories in both languages were primarily composed of current concerned topics such as work, money and achievement, which more likely co-occur with the cognitive category of "inhibition", but less likely co-occur with the perceptual category of "see" (See Table 7). This perhaps could be interpreted as the achievements are built on the destruction or reconstruction of the old and conquering obstacles (words in inhibition) in both cultures.

	Chinese	English
Identical	work; money; achieve; inhib see (-)	work; achieve; see (-); money; inhib
Non-Identical	relig (-)	relativ; space (-); cause; percept (-); motion (-); time (-); quant; i (-); filler (-)

Table 7: Categories in Achievement
Note: inhib, inhibition; relig, religion; relative, relativity; quant, quantifiers

The non-identical category in Chinese also contains currently-concerned topic "religion", but it is less likely to occur with the achievements. However, in English, non-identical categories vary from categories of relativity such as space, time and motion, perceptual category, cognitive category of causation in psychological processes, to linguistic category of quantifiers and spoken category of filler. Therefore, achievements perhaps are more complexly expressed in English. However, the more achievements are involved, the more likely causation, quantifiers and relativity occur, but the less likely space, motion, time, filler and the first personal pronoun singular "I" would occur in English. This is possibly predicted the English do not talk more about their past achievements.

Social Relation. The component of Social Relation had a correlation of .64** between Chinese and English (See Table 8). The identical categories in both languages consisted of all the social related categories such as social processes, family, friends and humans as well as the relevant concern "home". In addition, it also included linguistic categories such as personal pronouns, pronouns and she/he, perceptual category "hear", and spoken categories "assent" and "non-fluencies". Thus, the categories of words related to social relation co-occurred with pronouns and some oral words in both Chinese and English.

	Chinese	English
Identical	ppron; hear; social; pronoun; shehe; assent; family; humans; friend; home; nonfl	social; ppron; family; shehe; humans; hear; pronoun; home; friend; assent; nonfl
Non-Identical	i; you; percept; insight; verb	posemo; leisure; preps (-); number (-) they; sexual; relig

Table 8: Categories in Social Relation
Note: ppron, personal pronouns; i, I; nonfl, nonfluencies; posemo, positive emotion; preps, prepositions; relig, religion

However, in Chinese, social words co-occurred with the first singular personal pronoun "I" and the second singular personal pronoun "you" as well as "she/he", but in English just the third personal pronoun "they" and "she/he". This is perhaps that the Chinese is the collectivist culture focusing on the speakers and listeners during the conversation, whereas the English concentrates on other individuals. This argues that the use of pronouns is highly related to individual or/and group identity (Gorodnichenko 2011).

Besides, in Chinese, another two perceptual categories co-occurred, "perceptual processes" and "insight". This might be interpreted that the Chinese tend to perceive the utterances with the assist of verb category. Nevertheless, the English prefer to use more prepositions but less numbers, and to express more positive emotions when involved in leisure, sex and religion during socialization. However, these two topics are considered extreme personal and private so that it is not appropriate to talk about them openly in the Chinese culture.

Thus, the style of language implies the information about social relations. The relevant social information will vary profoundly with the change of language and cultures (Maass, Karasawa, Politi, and Suga 2006). Moreover, some of the most striking cultural differences in language like social closeness are inherent in function words rather than content words (Boroditsky, Schmidt, and Phillips 2003) in English.

Non-Correlated Component. The component three in Chinese was not significantly correlated with the component six in English. In Chinese, the categories included relativity, space, time and motion, which are related to the space and time (See Table 9). Meanwhile, this component also encompassed leisure, which could be interpreted when the Chinese talk about the past events or the future, they tend to concern about the leisure, but not the present. This is also proved by the less co-occurrence of the use of negation. However, in English, the space and time categories co-occurred less likely with achievements, not leisure.

Chinese	English
relativ; space; negate(-); time; motion; leisure	incl; conj; WPS; we

Table 9: Categories for Non-correlated component
Note: relative, relativity; negate, negation; incl, inclusive; conj, conjunctions, WPS, words per sentence

In English, this component is of a hybrid, mixed with a cognitive category "inclusive", and three linguistic categories "conjunctions", "words per sentence", the first personal pronoun plural "we". Thus, this component does not show much significant interpretation.

In summary, five components showed that English and Chinese languages do share some common psychological processes relevant to the linguistic usage of words. The most representative shared features are Negative Valence, Embodiment, Narrative, Achievement and Social Relation.

However, Narrative is the first component representing the same feature in both languages. It is perhaps that over 40% documents are of language arts in both languages. However, the second component in Chinese was Social Relation, possibly because the Chinese collectivist culture emphasizes the harmonious relationship (Triandis 2001), and also because dominant documents are of social studies in Chinese, over 10% more those in English. Meanwhile, "Achievement" was ranked in the second in English but last in Chinese. It is perhaps interpreted by the individualist western culture which highly values the personal achievements (Gorodnichenko 2011) than the eastern culture. Another salient component "Space and Time" independently existed in Chinese, but not in English. Perhaps in social studies, more documents are involved in historical documents in Chinese. Therefore, the further study needs to compare the linguistic and psychological word use in the same genre with the similar proportion to find out representative features.

Conclusion and Future Work

This study identified the common psychological and linguistic features with LIWC word categories in English and Chinese. The five components in two languages were cor-

related significantly at the level of .01 from .640** to .918**. Therefore, it is likely to argue that English and Chinese do share some common psychological features represented by words from the perspectives of negative valence, embodiment, narration, achievement and social relation. However, such categories as sexual, social relation, achievements are influenced by and constricted to the eastern and western cultural constraints and ideology (Fodor 1987), whose differences are mirrored in language use. This conclusion possibly to some extent refutes the doctrine of linguistic relativity known as the Sapir-Whorf hypothesis prevalent in the 19th century (Sapir 1921; Whorf 1956). Although the components ranked differently in both languages such as Social Relation in Chinese and Achievement in English as the second component, this is possibly partially due to cultural discrimination and partially due to the corpora selection.

For future work, we aim to enlarge the corpus to balance the documents in both languages to make them more comparable. Ultimately, results from this research will be explored further to evaluate the quality and credibility of the Chinese and English LIWC programs, and also to examine the similarities and differences of the psychological processes through types of word use.

Acknowledgements

This work was funded by the National Science Foundation (BCS 0904909) for the Minerva project: Languages across Culture. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies, cooperating institutions, or other individuals.

References

Aydede, M. 1999. On the Type/Token Relation of Mental Representations. *Facta Philosophica 2:*23-50.

Bache, C. 2008. *English Tense and Aspect in Halliday's Systemic Functional Grammar: a Critical Appraisal and Alternative.* London: Equinox Publishing.

Boroditsky, L., Schmidt, L., and Phillops, W. 2003. *Sex, Syntax, and Semantics.* In D. Gentner and S. Goldin-Meadow (Eds.), Language in Mind: Advances Study of Language and Cognition. Cambridge, MA: MIT Press.

Bradac, J. J., Bowers, J. W., and Courtright, J. A. 1982. Lexical variations in intensity, immediacy, and diversity: an axiomatic theory and causal model. In R. N. S. Clair, H. Giles (Eds.), *The Social and Psychological Contexts of Language.* NC: Baker and Taylor.

Bradac, J. J., Konsky, C. W., and Davies, R. A. 1976. Two Studies of the effects of linguistic diversity upon judgements of communicator attributes and message effectiveness. *Communication Monographs, 43:*70-79.

Chao, Y. R. 1956. Chinese terms of address. *Language, 32(1):*217-241.

Fodor, J. A. 1987. *Psychosemantics: the Problem of Meaning in the Philosophy of Mind*. Cambridge, MA.: Bradford Books/The MIT Press.

Gorodnichenko, Y. 2011. Individualism, innovation, and long-run growth. *PNAS Proceedings of the National Academy of Sciences of the United States of America* 108(4):21316-21319.

Graesser, A. C., and McNamara, D. S. 2010. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 2:371-398.

Graesser, A. C., Gernsbacher, M. A., and Goldman, S. 2003. *Handbook of Discourse Processes*. Mahwah, NJ: Erlbaum.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., and Hu, X. 2007. Using LSA in AutoTutor: learning through mixed-initiative dialogue in natural language. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (Eds.) *Handbook of Latent Semantic Analysis.* N.J.: Erlbaum, Mahwah, NJ.

Huang, C. L., Chung, C. K., Hui, N. H. H., Lin, Y.-C., Seih, Y., Lam, B. C. P., Chen, W.-C., Bond, M. H., and Pennebaker, J. W. (in press). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology*.

Keller, A. T. 1997. *Vernacular Literacy: A Re-Evaluation*. London: Oxford University Press.

Lin, J. 2005. Time in a language without tense: the case of Chinese. *Journal of Semantics* 23:1-53.

Maass, A., Karasawa, M., Politi, F., and Suga, S. 2006. Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality and Social Psychology* 90:734-750.

Mulac, A., Hanley, T. D., and Prigge, D. Y. 1974. Effects of phonological speech foreignness upon three dimensions of attitude of selected American listeners. *Quarterly Journal of Speech* 60:411-420.

Pennebaker, J. W., and Graybeal, A. 2001. Patterns of Natural Language Use: Discourse, Personality, and Social Integration. *Current Directions in Psychological Science* 3:90-93.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. 2007. *Linguistic Inquiry and Word Count: LIWC [Computer software]*. Austin, TX: LIWC net.

Pennebaker, J.W., and Chung, C.K. 2009. Computerized text analysis of Al-Qaeda transcripts. In K. Krippendorff and M. Bock (Eds.), *A Content Analysis Reader*. Thousand Oaks, CA: Sage.

Pennebaker, J.W., and Chung, C.K. 2011. Expressive writing and its links to mental and physical health. In H. S. Friedman (Ed.), *Oxford Handbook of Health Psychology*. New York, NY: Oxford University Press.

Sapir, E. 1921. *Language*. New York: Harcourt, Brace.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 1:24-54.

Triandis, H. C. 2001. Individualism-collectivism and personality. *Journal of Personality* 69(6):907-924.

Vick, C. F., and Wood, R. V. 1969. Similarity of past experience and communication of meaning. *Speech Monographs* 36:159-162.

Whorf, B. L. 1956. Science and linguistics. In J. B. Carroll (Ed.), *Language, Thought and Reality*. Cambridge, Mass.: M. I. T.

Wood, R. V., Yamauchi, J. S., & Bradac, J. J. 1971. The communication of meaning across generations. *Journal of Communication, 21:* 160-169.

Zeno, S. M., Ivens, S. H., Millard, R. T., and Duvvuri, R. 1995. *The Educator's Word Frequency Guide*. Brewster, NY: Touchstone Applied Science Associates.