FLOW

Foreign Language Opportunities *in* Writing

Edited by
Jan Majer and Łukasz Salski

Łódź University Press • Łódź 2011

Przemysław Krakowian
University of Łódź

INVESTIGATING THE NATIVE SPEAKER PHENOMENON – A PILOT CORPUS STUDY OF NATIVE AND NON-NATIVE WRITING

Abstract

The aim of this report is to provide a preliminary account of the investigation of two general corpora of written English, which was prompted originally by interest in an analytical tool designed to assess the propositional density in the utterances of learners of English. Since corpora of written language are easier to obtain and to procure in comparison with corpora of spoken language, the procedure was honed and fine tuned on a written corpus with the aim to investigate spoken utterances in an attempt to validate a scoring procedure. Propositional density was envisaged at the onset of the study as an instrumental factor in determining the relative merit of an assortment of samples. A computer program called CPIDR (a Computerized Propositional Idea Density Rater, pronounced "spider") involves a relatively straightforward procedure and produces results which are easy to interpret for most purposes

1. Overview

The aim of this report is to provide a preliminary account of the investigation of two general corpora of written English, which was prompted originally by interest in an analytical tool designed to assess the propositional density in the utterances of learners of English. Propositional density was envisaged at the onset of the study as an instrumental factor in determining the relative merit of an assortment of samples. A computer program called CPIDR[4] (a Computerized Propositional Idea Density Rater, pronounced "spider") involves a relatively straightforward procedure and produces results which are easy to interpret for most purposes.

In an attempt to overcome obstacles in reaching suitable convergence in sample assessment in the WebCEF project[5], where oral proficiency of non-

[4] http://www.cs.uga.edu/~wcb/cpidr/
[5] http://webcef.eu

native samples was assessed collaboratively in an online environment, several promising avenues of research were explored.

In order to test the suitability of the program for analysing spoken language, a purpose for which its designers feature a separate algorithm, existing small corpora of written language were used, in the hope that there will be adequate discriminating differences between native and non-native language.

Interest in CPIDR can be described in relation to two separate phenomena: (i) work conducted in connection with the English Showcase and the discussion on convergence; and (ii) the work carried out in relation to the samples collected for the Polish Showcase. Before the samples for the Polish Showcase were selected, approximately 200 samples of monologues and dialogues involving non-native and native speakers were recorded, reviewed and analysed before consensus was reached. In an attempt to reach convergence the researchers looked at distinguishing features of good, fluent and proficient speech. This necessitated establishing some selection procedure and criteria where one of the premises used in the process was the notion of propositional and idea density, a notion which was postulated to be a distinguishing feature of samples that could be deemed native-like.

The notions of propositional density and idea density are an attractive premise in attempting to clearly define the rating scales for oral performance under various examination schemes. CPIDR (a Computerized Propositional Idea Density Rater, pronounced "spider") is a computer program that allows the researcher to establish the propositional idea density of a transcribed spoken text without human intervention (Brown, Snodgrass, Kemper, Herman and Covington, 2008). The authors of the program claim that it has been validated against human raters and the convergence is sufficiently high to lead to further applications in machine aided assessment (MAA). Recent neurological studies (Caplan, Alpert and Waters, 2008) of how people handle multiple propositions and how the mechanism changes both with age and when handling languages other than native, as well as studies dealing with propositional and idea density in the native speaker speech (Krakowian in print) tentatively suggest that there is perhaps another explanation accounting for the phenomenon of the native speaker and they point in the direction of the role of external context and implicature used by native speakers and more proficient non-native speakers, something which is not explicitly covered by CEF scales.

2. Propositional density, propositional idea density.

Propositional density, also known as proposition density, or P-density, but sometimes referred to as propositional idea density, and understood as in Kintsch (1974) and Turner and Greene (1977), can be determined by the total number of content words such as verbs (but not auxiliaries), adjectives, adverbs, prepositions, and conjunctions against by the total number of words (Snowdon, Kemper, Mortimer, Greiner, Wekstein and Markesbery 1996). In a research study by Brown, Snodgrass, Covington, Herman, and Kemper (2007), a computer algorithm was conceived and perfected allowing the researchers to obtain accurate idea density measures. The implementation of this algorithm, the CPIDR program (or a Computerized Propositional Idea Density Rater, pronounced "spider") was vetted against human raters, and according to its creators, it agrees with them better than they agree with each other, r = 0.97 vs. 0.82 respectively (Brown et al., 2008: 2).

Started by Kintsch and Keenan (1973) and Kintsch (1974), research into propositional density and idea density posits that propositions are the elements of the utterance involved in the process comprehending and recall of texts, both spoken and written. Following the Kintsch's paradigm (Kintsch and Keenan 1973; Kintsch 1974), with subsequent revisions of Turner and Greene (1977), the verb of the main clause alongside the subject, object, indirect object, and any other elements present form a single proposition. Additional descriptive elements such as modifiers in the form adjectives, adverbs which qualify the main verb, and qualifier phrases need to be seen as additional propositions.

The authors of the CPIDR program somewhat depart from Kintsch's ideas, as those differ from propositions in logic or logical semantics. The first, and probably most quantitatively important, point of departure concerns the fact that most of the information about the main verb in the main clause such as verb tense, aspect, and its modality is reduced in Kintsch's model of propositional density (Turner and Greene 1977). The second reason being that common nouns are not propositions in Kintsch's understanding of propositional density. As a result the model produces deflated measures of propositions as can be seen in the examples below. Following Kintsch's model both sentence (1) and sentence (2) contain the same number of propositions, namely three, and the propositional density measured in Kintsch's model, therefore, would respectively be 0.375 and 0.333 as the number of word in the examples is 8 and 9. Paradoxically, the sentence expressing a more complex meaning scores lower on the measure propositional density as it expresses the same number of propositions in more words.

(1) I would like to go to the cinema.
(2) I would like to have gone to the cinema.

The CPIDR algorithm accounts for the propositions differently:

CPIDR 3.2.2695.24633
"I would like to have gone to the cinema."
002 PRP W i
002 MD W P would
200 RB W P like
510 TO W to
402 VB W P have
200 VBN W P gone
200 TO W P to
201 DT W the
002 NN W P cinema
000 . .
6 propositions
9 words
0,666 density

"I would like to go to the cinema."
002 PRP W i
002 MD W P would
200 RB W P like
510 TO W to
200 VB W P go
200 TO W P to
201 DT W the
002 NN W P cinema
000 . .
4 propositions
8 words
0,500 density

The measures in CPIDR are obtained based on the notion of idea density, which the authors of the program tend to use over the term propositional density, which is understood as the number of expressed propositions divided by the number of words. In terms of semantics, idea density constitutes a gauge of the extent to which the speaker is making claims or for that matter making requests

rather than just referring to entities. Propositions here include the notions of verb tense, aspect, and its modality, as well as account for the common nouns.

There exists a body of empirical research in the field of psychology looking at the connection between idea density to readability (Kintsch and Keenan, 1973; Kintsch, 1998), its relation to memory and retention (Thorson and Snyder, 1984), evaluation of the writing by students, both native and non-native (Takao, Prothero, and Kelly, 2002). A number of studies attempted to explain how propositions are handled depending on age (Kemper, Marquis, and Thompson, 2001; Kemper and Sumner, 2001) as well as various neurological conditions connected with old age (Caplan, Alpert and Waters, 2008) and finally the role other languages play in relation to idea density (Altarriba and Heredia, 2008).

3. The phenomenon of the native speaker

Numerous attempts such as Voss (1979), Embretson and Gorin (2001), Takao, Prothero, and Kelly (2002), Milfont (2006), Sayeed (2007), to name but a few, have been made to ascertain the qualities of the native speaker in comparison with the L2 learners, some for evaluation and validation purpose, some to establish a uniform measure of oral proficiency and fluency, and some to dispel popular myths that native speakers speak faster, construct longer and more complex sentences, use longer, rarer and/or more sophisticated vocabulary and manage to involve a greater number of propositions in their speech.

In a recent research project, Krakowian (in print) claims that such empirical measures as those that were obtained in a corpus study of non-native speech, when compared with the performance of native speakers, contrary to popular beliefs, and what is more important to a evaluative study of speech perception by both native and non-native speakers, do not indicate a greater rate of speech in native speakers. A small transcribed corpus of over 50 samples of speech by Polish speakers of English as a Second and Foreign language, amounting to nearly 17,000 words, indicates that the rate of speech of non-native speakers is approximately 20 per cent higher than that of native English speakers. Other indices seem to be pointing in a similar direction. From the point of view of mere word length, vocabulary sophistication understood as the proportion of words belonging to different frequency groups, native speakers are statistically significantly behind non-native users of the language. The same observation applies to sentence length and the number of embeddings.

The only exception lies in propositional and idea density measures, but it seems to be only statistically significant when they are obtained by hand rather than by the CPIDR algorithm. This could be owing to a rather striking

observation made in the process of investigation concerning a commonsensical impression about some of the sentences analysed using the program. Consider the sentences below; sentence (3) is deemed in this mode of analysis to be more idea or propositional dense than sentence (4). The first of the sentences is a genuine sentence from a written task, an information brochure, from a corpus of written student work, the second one is prepared for the purpose of analysis by CPIDR.

> (3) Conveniently located for outdoor sports, surrounded by picturesque green forests and within easy reach of the Lublinek airport, Lodz is definitely a place to go.
>
> (4) Lodz is definitely a place to go as it is conveniently located for outdoor sports, and it is surrounded by picturesque green forests and lies within easy reach of the Lublinek airport.

The CPIDR calculations yield density of 0.520 based on 13 propositions in 25 words for the former and 0.469 based on 15 propositions in 32 words for the latter. The difference is negligible, and from a statistical point of view within measurement error, and based on the calculations the sentences may be considered comparable. Yet even a very superficial look at the sentences by a human rater will have to regard the first as a much more complex, mature and altogether more elegant and native-like. It would seem that the secret of the native speaker may lie somewhere else, or that automatic density measures are not necessarily yet the determining index to rely on.

WebCEF, a Socrates Minerva project involving a creation of a multimedia database of speech samples to be assessed within the Common European Framework of Reference for Languages offers sufficiently diverse material for analysis. Within the project a number of threads of discussion concerned the issue of convergence and adequate measures and criteria to be used in establishing valid instruments in language assessment (Krakowian in print). Principally, the samples were collected to build a showcase of samples to illustrate key concepts in CEF scales. With lack of consensus and convergence in the case of a large proportion of collected samples, several of the project partners investigated the issue of sample comparability with native oral performance on the same tasks.

The findings point to similar conclusions as were noted by Krakowian (in print) on the basis of the above mentioned small transcribed corpus, namely that while the differences in the rate of speech and language sophistication indices do sometimes favour native speakers in some of the selected samples in the

analysis, there is no statistically significant difference between empirical measures of native and non-native oral language production. Native language production, however, is perceived, following the analytical scales of the CEF to be consistently and statistically significantly better than the selected upper-scale non-native performance used in the analysis.

It has been tentatively noticed that the source of such perception may lie in the fact which escapes the corpus methodology and is beyond the grasp of the CPIDR model, but which may be impressionistically acknowledged by the raters, even when they allegedly use the analytical CEF scales. At closer inspection at least some of the samples involving native speakers contain propositions which are not only implied, but implied in a manner that makes the inference by their interlocutors easy.

The samples under investigation consist of several types of interaction including (i) native speakers talking with other natives speakers; (ii) native speakers talking with non-native speakers; (iii) native speakers talking to native speakers; and (iv) native speakers talking to non-native speakers.

The implicature used when addressing other native speakers, either as part of a monologue or interaction, seems to be more frequent, which would reflect how propositions are handled by native speakers, as well as the recognition that a non-native speaker will have greater difficulties handling multiple prepositions and requires language adjustment (Altarriba and Heredia, 2008), which apparently happens not only at the level of syntax, a phenomenon known as foreigner talk, but also at the level of propositional density both explicitly and through implicature.

4. Conclusion

Further research, preferably of quantitative type involving explicit marking/tagging to be included in the corpus of spoken language of comparable samples from native and non-native speakers is required to claim conclusively that implicature in native speech influences the perception of propositional density. Implicature, however, and more precisely the capacity for the quantitative control of implicature seems to be one of the factors distinguishing native and non-native speech.

References

Altarriba, J. and R. R. Heredia. 2008. *An Introduction to Bilingualism: Principles and Processes.* Taylor and Francis LLP

Anderson, R. C. 1982. "How to construct achievement tests to assess comprehension." *Review of Educational Research,*42: 145–170.

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R. and M. A. Covington. 2008. "Automatic Measurement of Propositional Idea Density from Part-of-Speech Tagging." *Behavioral Research Methods.* 40(2): 540–545. L&B

Caplan, D., Alpert, N. and G. Waters. 1988. "Effects of Syntactic Structure and Propositional Number on Patterns of Regional Cerebral Blood Flow." *Journal of cognitive nuroscience* , July 1998, Vol. 10, No. 4: 541-552 1998 Massachusetts Institute of Technology

Embretson, S. E. and J. S. Gorin. 2001. "Improving construct validity with cognitive psychology principles." *Journal of Educational Measurement*, 38: 343–368.

Flesch, R. 1948. "A new readability yardstick." *Journal of Applied Psychology*, 32: 221–233.

Freedle, R. and I. Kostin. 1991. "The prediction of GRE reading comprehension item difficulty for expository prose passages" (ETS Research Report #RR-91-29). *Educational Testing Service*; Princeton, NJ.

Johnson, D. K., Storandt, M. and D. A. Balota. 2003. "Discourse analysis of logical memory recall in normal aging and in dementia of the Alzheimer type." *Neuropsychology,*17: 82–92.

Kemper, S., Marquis, J. And M. Thompson. (2001) "Longitudinal change in language production: Effect of aging and dementia on grammatical complexity and propositional content." *Psychology and Aging,*16: 600–614.

Kemper, S. and A. Sumner. 2001. "The structure of verbal abilities in young and older adults." *Psychology and Aging,* 16: 312–322.

Kintsch, W. and J. Keenan. 1973. "Reading rate and retention as a function of the number of propositions in the base structure of sentences." *Cognitive Psychology,* 5: 257–274.

Kintsch, W. 1998. *Comprehension: A paradigm for cognition*. Cambridge University Press; Cambridge, UK.

--- 1974. *The representation of meaning in memory*. Erlbaum; Hillsdalex, NJ.

Krakowian, P. (in print). "A corpus study of L2 speech – in search of fluency determinants".

Milfont, T. L. 2006. "Native speaker biases." *PsycheD! - The University of Auckland PhD Newsletter of the Psychology Department, 1(8)*: 3-4.

Perfetti, C. A. and M. A. Britt. 1995. “Where do propositions come from?” In: Weaver, C. A., Mannes, S. and Fletcher, C. R. (eds) *Discourse comprehension: essays in honor of Walter Kintsch*. Erlbaum; Hillsdale, NJ: 11–34.

Sayeed, S. A. 2007. “The Notion ‘Native Speaker’: A Philosophical Response.” *Annual Review of South Asian Languages and Linguistics*.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R. and W. R. Markesbery. 1996. “Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study.” *JAMA*, 275: 528–532.

Takao, A. Y., Prothero, W. A. and G. J. Kelly. 2002. “Applying argumentation analysis to assess the quality of university oceanography students' scientific writing.” *Journal of Geoscience Education,* 50: 40–48.

Thorson, E. and R, Snyder. 1984. “Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts.” *Journal of Marketing Research,* 21: 127–136.

Turner, A. and E. Greene. 1977. *The construction and use of a propositional text base (Technical Report 63).* University of Colorado, Institute for the Study of Intellectual Behavior; Boulder, CO.

Voss, B. 1979. “Hesitation Phenomena as Sources of Perceptual Errors for Non-Native Speakers.” *Language and Speech*, 22: 129

Please visit our website at
www.wydawnictwo.uni.lodz.pl

ISBN 978-83-7525-564-5