REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
 54(107) 971-984

Mapping out the road from corpus linguistics to psycholinguistics

Trazando el camino de la lingüística de corpus a la psicolingüística

Max M. Louwerse
TILBURG UNIVERSITY
THE NETHERLANDS
mlouwerse@uvt.nl

Received: 21-VI-2021 / **Accepted:** 25-X-2021
DOI: 10.4067/S0718-09342021000300971

Abstract

Parodi (2007) made the case that corpus linguistics ought to more consider the second most common language spoken in the world (Spanish), and better disseminate the research findings on the structure of that language in the lingua franca of the academic world (English). Moreover, Parodi argued that corpus linguists should use corpora that are heterogeneous in nature, and that corpus linguistics and discourse psycholinguistics should go hand in hand. In the current paper these claims are taken to heart with an overview of how corpus linguistics and discourse psycholinguistics could be linked, by mapping out their relations with the Symbol Interdependency Hypothesis that predicts that language encodes the perceptual information. Built on previous research that shows that word order reveals semantic information that language users can take advantage of, and by showing that the longitude and latitude of cities can be estimated based on the way the city names share the same linguistic context, this paper shows – using examples from the Spanish language and the South American continent – that language creates meaning.

Key Words: Corpus linguistics, computational linguistics, psycholinguistics, symbol interdependency.

Resumen

Parodi (2007) planteó que la lingüística de corpus debería tener más en cuenta la segunda lengua más hablada en el mundo (el español), y difundir mejor los resultados de la investigación sobre la estructura de esa lengua en la lengua franca del mundo académico (el inglés). Además, Parodi sostenía que los lingüistas de corpus debían utilizar corpus de naturaleza heterogénea, y que la lingüística de corpus y la psicolingüística del discurso deberían ir de la mano. En el presente trabajo se abordan estas afirmaciones con una

panorámica de cómo podrían vincularse la lingüística de corpus y la psicolingüística del discurso, trazando sus relaciones con la Hipótesis de Interdependencia de Símbolos que predice que el lenguaje codifica la información perceptiva. Partiendo de investigaciones anteriores que demuestran que el orden de las palabras revela información semántica que los usuarios de la lengua pueden aprovechar, y mostrando que la longitud y latitud de las ciudades pueden estimarse a partir de la forma en que los nombres de las ciudades comparten el mismo contexto lingüístico, este artículo demuestra -utilizando ejemplos del idioma español y del continente sudamericano- que la lengua crea significado.

Palabras Clave: Lingüística de corpus, lingüística computacional, psicolingüística, interdependencia de símbolos.

INTRODUCTION

Few people are WEIRD. That is the conclusion drawn by Henrich, Heine and Norenzayan (2010) on the subject population of most psychological experiments. Generally, participants in psychology experiments tend to come from Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies. When reviewing the academic literature, Henrich et al. found that only 12 percent of the world's population represent almost 80 percent of study participants in the academic literature. They argue that in order to understand human psychology, researchers ought not to focus their attention on the selective 12 percent, but take the heterogeneity of the world population into account. Arnett (2008) voiced the same concern: Psychological research focuses too narrowly on North-Americans, who consist of less than 5% of the world's population. Generalization from data obtained from such a small part of the population may yield invalid conclusions. In other words, we should not focus on the small selected few but consider the heterogeneity of the many.

What is true for psychological research, is also true for corpus linguistics. Linguistics at large, corpus linguistics being no exception, tends to focus on the structure of one language. Those studies published in the lingua franca of the academic community, English, tend to focus on language phenomena from just that one language, namely English. However, the English language is spoken by only 5% of the world population. And English is not even the most common language spoken. There are 2.4 times more Chinese speakers than English speakers. And there are some 25% more speakers of Spanish in the world than there are speakers of English. Parodi (2007) strongly voiced the argument that corpus linguists should focus on languages other than English. Moreover, he argued that studies that investigate the structure of that other language, for instance Spanish, should be disseminated in the lingua franca of the academic community, English, to avoid cross-linguistic findings are not disseminated. That is, Parodi (2007) argued that generalizations from language data obtained from primarily one language (English) may yield invalid conclusions. The recommendations for the psychology community with regards to the WEIRD population can thus be extended

to the linguistic community when it comes to languages: we should not focus on one selected language but consider the heterogeneity of many languages.

Parodi (2007) also emphasized the importance of heterogeneity of corpora (Parodi, 2007, 2010, 2015). Rather than focusing on a small set of texts from a specific genre or register, Parodi emphasized the importance of large texts on different genres and registers to allow for drawing conclusions on a language. Most corpora however are written texts, for the obvious reason that there are better records of written than spoken discourse. They are also primarily modern, again because of electronic records of newspaper articles, email archives and books.

In conclusion, whereas Henrich et al. (2010) cautioned against a WEIRD effect in psychological studies, Parodi (2007) cautioned against a similar effect in corpus linguistics, one I will dub the WHAM effect: a narrow focus on texts that are Written, Homogeneous, American English and that tend to be Modern. Spoken, heterogeneous, cross linguistic, and older texts tend to be forgotten. WHAM texts are not problematic per se, but with the evidence from corpus linguistics that are considerable differences between registers of discourse (Biber, 1988), generalizing from WHAM texts might lead to invalid conclusions for language at large.

In addition to the recommendation to also consider corpora other than English, Spanish in particular, and to consider the heterogeneity of corpora, including written and spoken registers to avoid what I have dubbed the WHAM effect, Parodi (2007) makes another important point that will serve as the thread of this paper: when it comes to studying language, both corpus linguistic as well as a discourse psycholinguistic perspectives need to be considered. To identify the structure of language, a corpus linguistic approach may be needed, but because language is a product of the human mind, a psycholinguistic approach should be embraced.

In the spirit of Parodi (2007) the current paper has two objectives. First, the paper bridges corpus linguistics to (discourse) psycholinguistics by demonstrating how cognitive perceptual processes might be built on language structure. In earlier work I have dubbed the mapping of perceptual representations onto language statistical structures the Symbol Interdependency Hypothesis (Louwerse, 2007, 2011, 2018, 2021) Second, following Parodi's (2007) recommendation, the paper aims to give examples for Spanish and the South American continent. Examples previously reported for English are illustrated with Spanish examples, and examples previously reported for the United States, China and the Middle East are illustrated with an example from South America. The purpose of the study is not to be exhaustive and complete. Instead, the purpose is to be illustrative how to map out the road from corpus linguistics to psycholinguistics thereby operating in the spirit of what Giovanni Parodi has argued for in his work.

1. Corpus linguistics and psycholinguistics

Research in corpus linguistics has provided important information. For instance, using corpus linguistics differences in genres and registers can be identified (Louwerse, McCarthy, McNamara & Graesser, 2004), the readability of text can be estimated (Graesser, McNamara, Louwerse & Cai, 2004) and literary periods can be estimated (Louwerse, 2004).

Differences in genres and registers were investigated by Biber (1988) who performed an extensive corpus linguistic analysis on a large number of heterogeneous texts, focusing on word-level linguistic analyses, and using factor analysis obtained six dimensions of registers in text, ranging from the involvement of production, the narrative concerns, explicit reference, overt expression of persuasion, abstract information, and on-line informational elaboration. Louwerse et al. (2004) extended these analyses by moving beyond the word level and including semantic relations. By considering not only surface-level structural information but also semantic and coherence information similar dimensions were obtained as those found by Biber (1988). Moreover, an additional dimension was obtained that was most prominent, the difference between speech and writing.

With regards to the readability of text, we developed a corpus linguistic tool that measured linguistic features at the syntactic, semantic, and pragmatic levels called Coh-Metrix (Graesser et al., 2004) was used the assess the cohesion in text. Coh-Metrix was developed as an exploration to determine whether readability measures of text could move beyond the traditional type-token ratio computations. Using hundreds of features Coh-Metrix was able to better adapt texts to the reader.

Finally, Louwerse (2004) aimed to identify the idiolect and sociolect of literary texts using both Boolean and vector space models. Based on theories in literary studies, Louwerse used corpus linguistic means to identify semantic fields in literary text in order to make predictions with regards to th literary period a text belonged to, demonstrating the opportunities for empirical studies of literature.

Findings patterns in corpora are useful to better understand a selection of texts, such as argumentative texts (Bolívar & Parodi, 2014) and helps to shape the field of applied linguistics, offering solutions to language-related real-life problems, from textbooks, to second-language learning (Crossley, Louwerse, McCarthy & McNamara, 2007). For applied linguistics purposes the explanation why certain patterns emerge is considerably less important than to demonstrate that they emerge. When measuring linguistic variables on a large corpus of text to perform dimension reduction techniques in order to obtain the most prominent dimensions, it is more important to know what the registers are, than why specific linguistic features yield these dimensions. Similarly, being able to measure the readability of texts reliably is more important than being able to explain why certain (combinations of) linguistic features distinguish different kinds of

texts. And for identifying the idiolect and sociolect in literary texts a test whether corpus linguistics allows for identifying the idiolect of the author or the sociolect of the authors within a literary period is more important than knowing why certain features identify idiolect and sociolect. This is not to say that corpus linguistics and applied linguistics do not provide explanations. The point is that these explanations are less important for the purposes of the investigation.

However, when corpus linguistics techniques are used to identify the structures in language, explanations more than applications become increasingly important. Parodi (2007: xii) is aware of this, arguing that it is important "to find a way to connect both areas with profound and promising links". Explanations why language is structured in a particular way allows for exciting new avenues in corpus linguistics, computational linguistics and cognitive science (Louwerse, 2021). I will give two examples to illustrate this point, one coming from Zipf (1935), the other from Firth (1957).

2. Least effort

One of the most commonly reported phenomena in philology, quantitative linguistics and corpus linguistics is what has been dubbed Zipf's Law. The linguistic law is named after the philologist who popularized it (Zipf, 1932, 1935), even though the language regularity was reported two decades earlier (Estoup, 1912). Zipf's Law is a regularity in language that concerns the inverse relationship between the rank of a word and its frequency. If one creates a word frequency list of a corpus –any corpus in any language– the most frequent word of that corpus ranked first has approximately double the number of words of the second most frequent word, ranked second. The word ranked third frequent has half the frequency of the second ranked word, and so on. A rather consistent power law function emerges that consist of two variables, a constant and a parameter that identifies the slope of the curve. This power law seems rather ubiquitous across corpora, across languages, and across genres. Zipf's Law can be extended to the Zipf's Law of Abbreviation, the inverse relationship between word length and word frequency. More frequent words tend to be shorter (or shorter words tend to be more frequent). From a corpus linguistics approach, Zipf's Law and the Law of Abbreviation shed light on the structure of language.

So far these laws can be viewed as useful corpus linguistic regularities. Zipf however went beyond identifying the structure of language –let's call it the corpus linguistic approach–providing a cognitive explanation for the observed regularity –let's call it the psycholinguistic approach. Zipf argued for a Principle of Least Effort (Zipf, 1949) stating that speakers optimize their behavior by minimizing their effort. Generally speakers (writers) aim for unification. They would prefer using a small number of words as frequently as possible, rather than using a large vocabulary of different words less frequently. However, for the hearer (or the intended reader) diversification is important.

For them more different words with a lower frequency are more advantageous, for instance because frequent words tend to be ambiguous.

There are two reasons Zipf's Law and the Principle of Least Effort are noteworthy. First of all, Zipf's Law and Zipf's Principle of Least Effort are good examples of linking a finding in corpus linguistics to a psycholinguistic explanation. As a consequence of this link between corpus linguistics and psycholinguistics, new predictions emerge. For instance, if the Principle of Least Effort is an explanation for Zipf's Law, we could make predictions which language genres and registers better fit Zipf's Law. And secondly, Zipf's Principle of Least Effort demonstrates something important about language users. They cut corners. They prefer to use linguistic shortcuts in their communicative efforts. And that is not only relevant for psycholinguistics. It turns out to also be important for corpus linguistics.

3. Company it keeps

A second example of regularity in language comes from measuring the similarity in meaning between words, sentences, paragraphs and texts based on how words operate in language. Computational linguistic techniques such as Global Vectors for word representation (GloVe; Pennington, Socher & Manning, 2014), Continuous Bag of Words (CBOW; Mikolov, Sutskever, Chen, Corrado & Dean, 2013) as part of the Word2Vec architecture, and Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis & Kintsch, 2007) all allow for measuring semantic relations between text units. LSA for instance builds a word x paragraph matrix that consists of the frequencies with which words co-occur at first and higher order levels. Applying matrix algebra, the initial sparse matrix is reduced to a concise matrix of which the distance between the vectors of words represents the semantic similarity between the words. LSA has for instance been used to identify the correct answer of Teaching of English as a Foreign Language tests (Landauer & Dumais, 1997). By measuring each of the four answers on a multiple choice test, LSA was able to get a passing TOEFL grade. Similarly LSA has been able to successfully evaluate the quality of summaries provided by students, thus being an automated essay grader (Foltz, Laham & Landauer, 1999). When students submitted their essays to LSA, LSA was able to measure the quality of the essays such that students could make improvements. LSA is also used by AutoTutor, an intelligent tutoring system that measures the similarity between an incoming student answer with the ultimate good and bad answers to a question in order to adjust its pedagogical moves in a conversational tutoring session with a student (Graesser, Lu, Jackson, Mitchell, Ventura, Olney & Louwerse, 2004). The fact that these corpus linguistic techniques allow for predicting answers on a TOEFL test, automatically score essays, or serve as the backbone for an intelligent tutoring system is remarkable and provides valuable tools for applied linguistics. Yet why this is the case is equally important, but less clear.

Vector space models such as LSA are useful computational linguistic tools that can have a variety of valuable applications. But we can also extend the corpus linguistic approach to a psycholinguistic approach and wonder why vector space models such as LSA perform so well on language. The psycholinguistic answer may be found in an adage in Firth (1957): you shall know the meaning of a word by the company it keeps. It seems so obvious but given that language users have full flexibility to decide when to use specific words where, it is remarkable that it is apparently the case that words similar in meaning occur in similar contexts. And this psycholinguistic explanation for why corpus linguistic techniques like LSA perform the way they do allow us to open up new research avenues.

What both examples –Zipf's Law and LSA– demonstrate here is that identification of the language structure may be useful, but asking why the structure emerges is equally useful, as it gets us to the psycholinguistic approach which in turn might yield questions for the corpus linguistic approach. That is, the (psycholinguistic) explanation can tell us something about language, about the human mind, and how these structures come about.

4. Symbol interdependency hypothesis

Over the last two decades much of the cognitive science literature has argued that linguistic symbols must be grounded to be meaningful (De Vega, Glenberg & Graesser, 2008). For words to be meaningful they must refer to referents in the outside world or at least to perceptual experiences with these referents. Linguistic symbols have been considered abstract, amodal and arbitrary (Glenberg & Kaschak, 2002). Take for instance the word 'chair'. It refers to the abstract concept chair (not the wooden or plastic one, but any chair). The word is visual (the written word) or auditory (the spoken word) and therefore amodal. Moreover, there is no relationship between the form of the word and its meaning (the word 'chair' in English sounds different than the word *silla* in Spanish and yet they refer to the same object). When compared to the linguistic symbol, the the picture of a chair refers to an actual chair, that wooden or plastic one. It is modal as it concerns a modality-specific representation. And in the case of pictures the form-meaning relationship is not arbitrary but fixed: it looks like a chair and is a chair. This view of embodied cognition is of interest for psycholinguists, but does not map out a road to corpus linguistics. It seems.

In a number of studies, Louwerse (2011, 2018) and most recently extensively outlined in a popular science book (Louwerse, 2021), I have argued that the grounding argument is incomplete. Of course, linguistic symbols can often be grounded, but the question is whether they always must be grounded. It seems that some of the cognitive effort to ground linguistic symbols may be offloaded the language system itself. Language has evolved in such a way that it represents the world around it. If language

is structured after its meaning, language users can use the least effort to maximize estimating meaning. All language users need to do is recognize the patterns, so they can estimate the meaning of words using linguistic shortcuts. Language thus creates meaning. Language users can create perceptual experiences of the linguistic symbols they encounter, but they can also rely on the relationships between symbols. In other words, symbols are interdependent on one another and on their referents. Louwerse (2021) gives a range of examples, but this paper limits itself to two cases, word order and word context.

To demonstrate that meaning can be extracted from language itself, let me first focus on word order, specifically binomials. It seems that language users have full flexibility over the order in which they place their words. We could say 'corpus linguistics and psycholinguistics' or 'psycholinguistics and corpus linguistics'. There is no linguistic law enforcement that tells us in what form the order of nouns need to be. And yet the order is not as arbitrary as we may think. Louwerse (2008) showed that if two concepts that have a high and low referent in the outside world, for instance 'sky and ground' the higher concept tends to precede the lower concept in language more frequently than the other way around. That is 'sky and ground' occur significantly more frequently than 'ground and sky'. And this is not limited to the binomial itself. Take any two words which have a vertical spatial relationship and count the number of times they co-occur in an *n*-gram window (where *n* is for instance 5), and it is most likely the higher concept precedes the lower concept, rather than the other way around.

This finding is not limited to vertical spatial relationships. They also apply to gender and authority relationships, with masculine words preceding feminine words more frequently than the other way around (ladies and gentlemen being the obvious conventionalized exception), and those in charge preceding those not in charge more frequently than the other way around. Hutchinson and Louwerse (2013) found that positive valence words more commonly precede negative valence words than the other way around. These patterns are not reserved for English, but seem to apply to other languages as well. Let me demonstrate this with a few Spanish examples. When using binomials in the Google *n*-gram corpus for Spanish corpora in 2019 (see Michel, Shen, Aiden, Veres, Gray, Pickett & Aiden, 2011), we can compare the frequency of one word order combination with another. 'Head and shoulders' is more frequent than 'shoulders and head', not only in English but also in Spanish, with the former being four times more frequent than the latter (*cabeza y hombros* versus *hombros y cabeza*). 'North and South' in Spanish (*norte y sur*) is ten times more frequent than 'South and North' (*sur y norte*). Similarly, the binomials 'positive and negative' (*positivo y negativo*) with positive preceding negative are over 6 times more frequent than the binomial 'negative and positive' (*negativo y positivo*). For the Spanish translation of 'plus and minus', the positive-negative order is almost four times more frequent than the negative-positive order.

Table 1. Examples of word order encoding vertical relations.

Spanish valence	English valence	Number of times high-low is more frequent than low-high
alto y bajo	high and low	5.18
arriba y abajo	up and down	5.18
cabeza y hombros	head and shoulders	4.16
cielo y suelo	sky and ground	1.27
norte y sur	north and south	10.23

Table 2. Examples of word order encoding valence relations.

Spanish valence	English valence	Number of times positive-negative is more frequent than negative-positive valence
positivo y negativo	positive and negative	6.16
más y menos	plus and minus	3.94
bueno y malo	good and bad	9.82
cielo e infierno	heaven and hell	9.18
ganar y perder	win and lose	6.16

These examples of vertical order and the valence order effect are not carefully selected. They have been more extensively been reported in Louwerse (2007) and Hutchinson and Louwerse (2013). These examples demonstrate a few things. First, even though language users have the full flexibility to say 'shoulders' followed by 'head', the more iconic order is more frequent. This suggests that some meaning is encoded in the word order relation. Second, the order does not mean that grounding is not needed. I still need to know that the two words have some vertical or valence relation, and the order could be reversed, but with a higher probability than expected by chance meaning can be bootstrapped from the linguistic word order information alone.

Relations that we can see in the outside world are mapped onto language. When translating the world outside in language, language encodes the order in the perceptual world. These patterns are of course not restricted to binomials and word order. We can take this further to linguistic context in general. Language users have full flexibility in choosing any word order, but apparently stick to the word order that best maps onto the perceptual world around them. In Similarly, language users have the flexibility to use any word in any context, but constrain themselves in such a way that the meaning of a word can be estimated by the company it keeps.

To take an extreme example, it is possible to estimate the longitude and latitude of cities on the basis of the way city names are mentioned in language. Louwerse and Zwaan (2009) tested the hypothesis that cities that are talked about together are located together. That is, it is more likely that the words 'Washington' and 'Boston' are

mentioned in the same sentence, than that the words 'Boston' and 'Los Angeles' are mentioned in the same context. We took the 50 largest cities of the United States and computed the LSA cosine values for the semantic similarities. Applying Multidimensional Scaling we then obtained the loadings of the city names on two dimensions. These dimensions correlated with the actual longitude and latitude of the cities in the United States, not only for one corpus, but for three different corpora on which LSA was trained. A bidimensional regression analysis – bringing two correlations for longitude and latitude back to one – confirmed these findings. In other studies this analysis was repeated for cities in China and the Middle East using Chinese and Arabic texts (Louwerse, Hutchinson & Cai, 2012). In fact, the same analysis even predicted the longitude and latitude of locations in Middle Earth using place names in Tolkien's *Lord of the Rings* (Louwerse & Benesh, 2012).

It is important to stress that these findings cannot be explained by the computational model. Instead, the findings must be explained by the language system itself. When Louwerse and Zwaan (2009) repeated their experiment with first-order co-occurrences, their corpus had to be considerably larger, but the findings were more or less similar as those obtained by the vector space model. The magic was less in the computational algorithm and more in the language structure.

As with the word order example given earlier, it is important to keep in mind that these findings do not exclude the grounding of linguistic information. This can be illustrated with a similar analysis predicting geographical information from language, now using cities in South America. For instance, we can select city names such as *Rio de Janeiro*, *Manaus*, *Santiago*, *Lima* and *Bogota*. Using the commonly used Touchstone Applied Science Associate (TASA) corpus (Touchstone Applied Science Associates, Inc.) that consists of 37,651 documents on a variety of different topics (including Literature, Arts, Science, Economics and Social Studies), a matrix of 5 x 5 cosine values were computed that were submitted to a Multidimensional Scaling (MDS) Analysis yielding a *Stress value* of .167 and an R^2 of .812. The plotting of the five city names on the two dimensions yielded the following Figure 1.

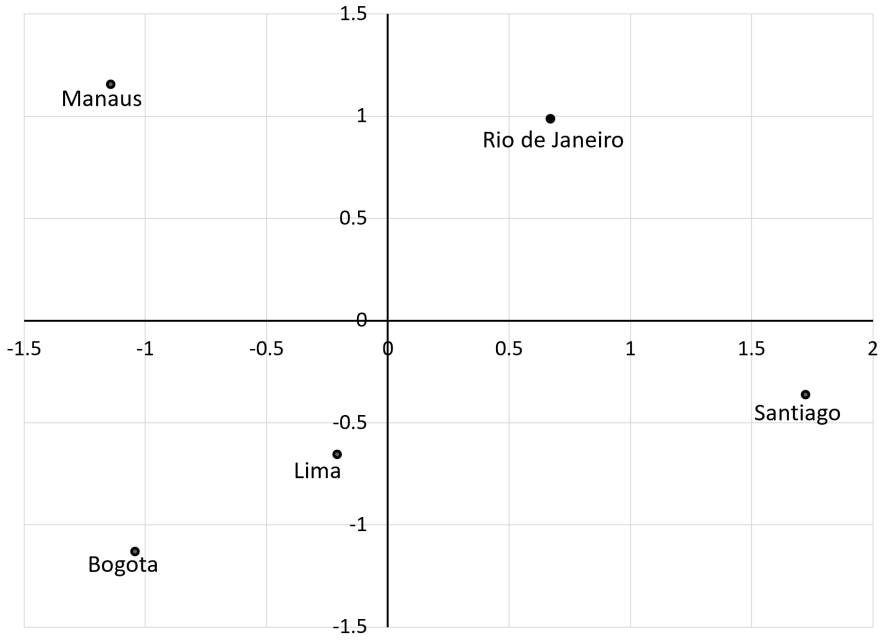

Figure 1. Distributional semantic estimates of the location of cities in South America.

This may not quite be a map of South America. However, with limited grounding, for instance, knowing that Santiago lies in the South of the continent, a map can be bootstrapped that shows interesting similarities with an actual geographical map of South America emerges (i.e., rotate the Figure 1 90 degrees and such an estimated map of South America appears). What these examples show is that language encodes perceptual information that language user can rely on to minimize their cognitive effort in grounding information to the perceptual world.

This is not an exhaustive analysis of city names in South America whose geographical location can be estimated. This is merely an example of linking language structure to psychological processes. Analyses conducted for the United States, China, the Middle East and Middle Earth can also be extended to South America.

CONCLUSION

In this paper, I have provided some initial demonstrations that findings found for the English language can be extended to the Spanish language. Patterns in word order represent perceptual world order and there is currently no evidence that this is a phenomenon restricted to a single language. Secondly, I have provided an initial demonstration that previous findings obtained for a geographical map of the United States can be extended to the South American continent. Patterns in linguistic context map onto the perceptual (geographical) context.

Parodi (2007) argued for a link between corpus linguistics and psycholinguistics and for a focus beyond what is common in corpus linguistics and psycholinguistics, the English language and the English speaking world. In the spirit of Parodi's argument, and in memory of Giovanni Parodi, the current paper has outlined the road from corpus linguistics to psycholinguistics, arguing that language is structured in such a way that the human mind is readily able to pick up on patterns mapped onto the outside world.

REFERENCES

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*(7), 602-614.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bolívar, A. & Parodi, G. (2014). Academic and professional discourse. In M. Lacorte (Ed.), *The Routledge handbook of Hispanic applied linguistics* (pp. 475-492). Routledge.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M. & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30.

De Vega, M., Glenberg, A. & Graesser, A. (2008). *Symbols and embodiment: Debates on meaning and cognition*. Oxford: Oxford University Press.

Estoup, J. B. (1912). *Gammes sténographiques. Recueil de textes choisis pour l'acquisition méthodique de la vitesse*. Paris: Institut Sténographique.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *Edmedia+ innovate learning* (pp. 939-944). Association for the Advancement of Computing in Education (AACE).

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558-565.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers,* 36, 193-202.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A. & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers,* 36, 180-193.

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29-29.

Hutchinson, S. & Louwerse, M. M. (2013). Statistical linguistic context and embodiment predict metaphor processing but participant gender determines how much. *Cognitive Linguistics*, 24, 667-687.

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*(2), 211-240.

Landauer, T. K., McNamara, D. S., Dennis, S. & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities,* 38, 207-221.

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107-120). Mahwah, NJ: Lawrence Erlbaum.

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin* & *Review, 15*(4), 838-844.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science (TopiCS)*, 3, 273-302.

Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science,* 10, 573-589.

Louwerse, M. M. (2021). *Keeping those words in mind: How language creates meaning.* Prometheus Books.

Louwerse, M. M. & Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, 33, 51-73.

Louwerse, M. M. & Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science*, 36, 1556-1569.

Louwerse, M. M., Hutchinson, S. & Cai, Z. (2012). The Chinese route argument: Predicting the longitude and latitude of cities in China and the Middle East using statistical linguistic frequencies. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 695-700). Austin, TX: Cognitive Science Society.

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 26th Annual Conference of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Lawrence Erlbaum.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Parodi, G. (Ed.) (2007). *Working with Spanish corpora*. London: Continuum.

Parodi, G. (Ed.). (2010). *Academic and professional discourse genres in Spanish* (Vol. 40). Amsterdam: John Benjamins.

Parodi, G. (2015). Variation across university genres in seven disciplines: A corpus-based study on academic written Spanish. *International Journal of Corpus Linguistics*, *20*(4), 469-499.

Pennington, J., Socher, R. & Manning, C. (2014). *Glove: Global Vectors for Word Representation*, 1532-1543.

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard: Harvard University Press.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton, Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.