EXPLORING INTRALEXICAL MEANING: SEMANTIC NEIGHBOURHOOD AND TRANSPARENCY EFFECTS ON THE READING OF COMPOUND WORDS

Katherine R. Matchett[1] and Lori Buchanan[1]
1 – University of Windsor, Department of Psychology
matche11@uwindsor.ca, buchanan@uwindsor.ca

ABSTRACT

This exploratory archival analysis investigates the relationships probabilistic co-occurrence measures of semantic association and semantic richness have with subjective measures of semantic transparency in English compound words. We also examine their correlations with behavioural measures. Specifically, the present study establishes links between nominal transparency classification [11]; participant ratings of whole word transparency and lexeme meaning dominance (LMD) [9]; vector-based measures of semantic distance, semantic neighbourhood density, and semantic neighbourhood dispersion [5, 6]; and lexical decision data from the English Lexicon Project [1]. We show that semantic distances between the whole compound and its constituent lexemes may capture associative information involved in transparency and LMD ratings. Further, evidence is presented for the semantic neighbourhoods of constituent morphemes playing a role in compound recognition.

Keywords: transparency, semantic neighbourhood density, compound words

1. INTRODUCTION

Morphologically complex words contain semantic relationships at multiple levels. With respect to concatenated English compounds (e.g., *strawberry*), the two constituent morphemes may relate to each other, as well as to the meaning of the whole compound. There are multiple ways to conceptualize and quantify these meaning-based relationships, but perhaps the most studied of these has been semantic transparency [11]. Transparency in this context refers to how predictable the meaning of the whole word is from the meanings of its constituent morphemes. The methods used to determine semantic transparency are typically subjective, prone to bias, and may not always correlate with other operationalizations of semantic relatedness in a manner that one would expect if they are indeed representative of the same underlying construct [7]. Here we examine various metrics assumed to describe the intralexical semantic relationships of English compounds. Specifically, we determine whether values obtained from a probabilistic global co-occurrence model of semantics [6] predict transparency ratings. We also assess the predictive ability of all examined constructs with respect to archival lexical decision RT data.

1.1. Semantic transparency, headedness, and LMD

Semantic transparency has been most commonly operationalized through expert classification into discrete categories [11, 12] and participant ratings of transparency [9]. Transparency of morphemes has been shown to facilitate the processing of compound words across a variety of task paradigms, including both primed and unprimed lexical decision tasks and typing tasks [15].

Other subjectively derived methods of understanding the meaning-based relationships within compound words exist. Most notable among these constructs is that of headedness, wherein the head of a compound is the lexeme that determines the semantic category to which the word belongs [8]. In 2008, Juhasz et al. assessed the dominance in meaning between constituents using participant ratings and showed that this method of classifying words into "headed" (first constituent dominant) and "tailed" (second constituent dominant) had robust and novel effects on behavioural measures [10].

1.2 Modelling semantic space

Significant variation exists in the conceptualization and design of models aiming to objectively characterize the semantic organization of the mental lexicon. (see Buchanan et al., [2] for a review). The present study, however, is concerned with language-based models of semantics that class words and concepts together based on various statistical co-occurrence properties of a corpus. It has been shown that well-designed models of this type can mimic object-based models (e.g., feature based models) in many contexts and are much easier to implement in an objective fashion [13, 2, 5]. Models of this type generate a high-dimensional semantic matrix, in which each word

has a defined location and its similarity to another word is measured as a function of the distance between them.

1.2.1. Semantic richness and the distributional characteristics of the semantic neighbourhood

The collection of words with which a target shares association can be referred to as its "semantic neighbourhood", with the "neighbours" being the associated words therein [2]. These neighbours vary in semantic distance from a given word; close neighbours are those with a stronger level of association to the target, whereas distant neighbours may share little association with the target. As such, target words may differ in the distributional characteristics of their neighbourhoods, with some neighbourhoods being more "dense" than others. That density of the neighbourhood is used as a measure of semantic richness. Language-based co-occurrence models have been used to generate a measure of semantic neighbourhood density (SND). Originally, SND was operationalized as the mean semantic association between a target word and its *n*th closest neighbours [2]. More recently, with the evidence that the relative positioning of neighbours may play a larger, or at least different, role in lexical processing than simple mean distance, [3, 14], measures that capture the relative distributions of neighbours across semantic neighbourhoods have come in favour. We investigate the effects of both in this study.

1.3 The present study

In accordance with questions raised by Wang, Hsu, Tien, & Pomplun [18] the present study employs a global co-occurrence model to predict transparency values. Semantic distances derived from the WINDSORS model [5] are used to this end. This model has been shown to effectively capture associative information, as well as feature and category information [5]. SND values derived from this model have demonstrated relationships with behavioural variables across task types [3]. In this study we examine two methods of deriving semantic richness, the second being a novel measure of semantic neighbourhood dispersion, to assess if the semantic neighbourhoods of constituent lexemes play a role in transparency and LMD ratings and/or lexical decision performance for their respective compounds.

2. METHODS

In this study, we examine semantic distances between the first constituent (C1), second constituent (C2), and whole word (WW) lexemes of English compound words and assess their ability to predict subjective measures of semantic transparency and LMD. Finally, we assess the ability of all semantic variables to predict archival lexical decision data from the English Lexicon Project (ELP) [1]; this dataset is used to avoid inter-stimulus compound priming effects [12].

2.1. Data sources and measures

This study makes use of four publicly available databases that can be found by accessing the relevant reference material. Words with extreme frequency values or that were not common between necessary datasets were omitted from analysis. Transparency classifications of 124 compounds are taken from Stathis [16]. Participant-rated transparency and lexeme meaning dominance (LMD) of 445 compounds are taken from Juhasz et al. [9]. For each of these two lists of compounds, semantic distances between each whole word and its constituents are taken from the WINDSORS database [5]. Measures of semantic richness for each constituent lexeme of each word are also derived from this database. For analyses not examining transparency, the full 445 compound list from Juhasz et al. are used. Finally, the lexical decision times for each whole compound is taken from the ELP [1].

2.1.1. Subjective semantic variables

Classification of transparency comes from a stimulus set developed by Stathis [16] in which words were classified into the four categories of transparent-transparent (TT), transparent-opaque (TO), opaque-transparent (OT), and opaque-opaque (OO) [11]. 124 noun-noun compounds from this dataset, along with their constituent lexemes, were used in all analyses involving nominal transparency classification.

Participant-rated semantic transparency and LMD are operationalized using ratings gathered by Juhasz et al. [9]. These ratings are based on the mean in-lab rating of compound words as part of a larger norming study. Whole word transparency was rated by participants on a 1-7 scale, where higher ratings indicate that a word's lexemes represent a word transparently. LMD was rated on a 0-10 scale, where 0 represents first constituent dominance and 10 represents second constituent dominance. Each word was rated by an average of 14 participants. A total of 445

compounds along with their constituent lexemes were included in all analyses involving these variables.

2.1.2. Semantic distance and semantic neighbourhood variables

Semantic distances between the first constituent (C1), the second constituent (C2), and the whole word (WW) are operationalized as the similarity cosine of the two lexemes as included in the WINDSORS database [6]. These values range from 0-1, with higher values representing higher association or closer semantic distance. The SND of C1, C2, and WW is operationalized as the mean distance to a target constituent and its 200 closest neighbours in the WINDSORS model, with values closer to 1 indicating a denser neighbourhood. A novel measure intended to better capture the distribution of semantic distances is also included. We call this measure dispersion of semantic association (DSA) and is operationalized as the standard deviation of the semantic distances to the closest 200 neighbours of a target word. Both SND and DSA were converted to z-scores before analysis.

2.1.3. Behavioural data

Mean lexical decision response time (RT) data for each whole compound noted in section 2.1.1 was gathered from the ELP [1], with 124 (for analyses involving nominal transparency) and 445 (for all other analyses) response times (RTs) being used in this study. These RTs were normed on 816 undergraduate participants, with each individual word in the dataset receiving 34 responses. Whole-word RTs are assessed with log orthographic frequency [4] as a covariate.

3. RESULTS

3.1 Predicting transparency and LMD

An analysis of variance comparing the semantic distance between WW and C2 across transparency classes revealed a large effect [$F(3,120) = 5.255$, $p = .002$, $\omega^2 = .087$]. Semantic distances between WW-C1 did not differ across transparency classification, see Figure 1. Multiple regression analyses found WW-C1 and WW-C2 semantic distances to be significant predictors of both participant-rated whole word transparency and LMD with large effect sizes (See Table 1).

Figure 1: Mean semantic distance between lexeme pairs compared across transparency classifications.

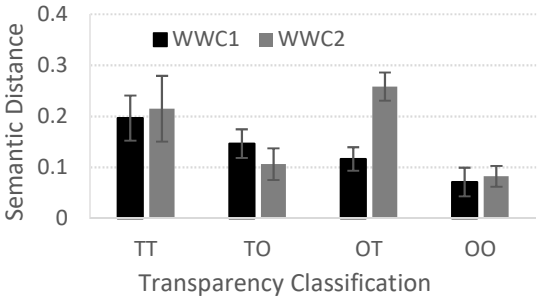

Table 1: Multiple regression analyses predicting whole word transparency ratings and LMD.

Variable	R^2	*B(SE)*	B	*F/t*	sr^2
Trans.	.16			43.82*	
WW-C1		1.87(.28)	.29	6.68*	.08
WW-C2		1.63(.29)	.24	5.54*	.05
LMD	.25			72.88*	
WW-C1		-2.05(.30)	-.28	-6.69*	.09
WW-C2		3.44(.32)	.45	10.85*	.20

***** *p<.001*

3.2. Predicting lexical decision performance

Analyses of variance revealed no differences in lexical decision times between nominal transparency classifications. Hierarchical regression assessed whether SND and DSA of all lexemes predicted response latency after frequency was entered. Frequency was entered into the first step of the equation, followed by SND of each lexeme for the second step, and finally DSA was added into the third step (see Table 2).

Table 2: Hierarchical multiple regression analysis predicting lexical decision response times.

Step	R^2	*B(SE)*	β	*F/t*	sr^2
First	.03			13.26*	
Freq.		-1.34(.37)	-.17*	-3.64	.03
Second	.05			5.78*	
Freq.		-1.27(.373)	-.17*	-3.42	.03
SND					
C1		12.03(5.42)	.11*	2.33	.01
C2		8.61(5.39)	.08	1.59	.00
WW		-7.15(4.55)	-.08	-1.59	.00

(continued)

Table 2. (continued)

Step	R^2	*B(SE)*	β	*F/t*	sr^2
Third	.15			7.85*	
Freq.		-.96(.37)	-.13*	-2.62	.01
SND					
C1		14.46(5.61)	.14*	2.69	.02
C2		14.26(5.46)	.13*	2.62	.01
WW		12.95(6.42)	.14*	2.02	.01
DSA					
C1		-8.21(4.94)	-.08	-1.68	.00
C2		-14.03(4.74)	-.14*	-2.97	.02
WW		-27.06(6.22)	-.30*	-4.36	.04

* $p<.05$

4. DISCUSSION

Our findings suggest that the WINDSORS-derived semantic distances may capture the underlying associative information that informs transparency and headedness judgements. Further, SND and DSA appear to contribute uniquely to the variance in lexical decision performance. Having closer-on-average neighbours appears to be facilitative to word recognition, whereas greater neighbourhood dispersion confers an inhibitory effect. Further, these effects arise not simply at the level of the whole word; SND of the constituents are at least as, and possibly more, facilitative than whole word SND. Notably, DSA of the whole word appears to have a far stronger effect then any SND variable, providing support for models of semantic processing that account for the variability of a neighbourhood [3, 14]. These results have implications for theories of morphological decomposition [17] and for attractor dynamics models of semantic processing [14] and as such, further research using these variables will likely prove worthwhile.

REFERENCES

1. Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459. Available: http://elexicon.wustl.edu/userguide.pdf
2. Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*(3), 521-544.
3. Danguecan, A. (2015). Towards a new model of semantic processing: *Task-specific effects of concreteness and semantic neighbourhood density in visual word recognition* (Doctoral dissertation).
4. Durda, K., Buchanan, L. (2006). *WordMine2* [Online] Available: http://web2.uwindsor.ca/wordmine
5. Durda, K. (2013). *Embodied properties of semantic knowledge acquired from natural language* (Doctoral dissertation).
6. Durda, K., Buchanan, L. (2008). WINDSORS: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods, 40*(3), 705-712.
7. Gagne, C. L., Spalding, T. L., & Nisbet, K. A. (2016). Processing English compounds: Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics, 12*(2), 2-21.
8. Hoeksema, J. (1992). 'The head parameter in morphology and syntax'. In D. G. Gilbers & S. Looyenga (eds.) *Language and cognition 2.* Groningen, Netherlands.
9. Juhasz, B. J., Lai, Y., & Woodcock, M. L. (2015). A database of 629 English compound words: Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research, 47*, 1004-1019.
10. Juhasz, B. J., Starr, M. S., Inhoff, A. W., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology, 94*(2), 223-244.
11. Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language, 61*(1), 30-44.
12. Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language, 84*(1), 50-64.
13. Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28(*2), 203-208.
14. Mirman, D., Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 65-79.
15. Smolka, E. & Libben, G. (2017). 'Can you wash off the hogwash?' – semantic transparency of first and second constituents in the processing of German compounds. *Language, Cognition, and Neuroscience, 32*(4), 514-531.
16. Stathis, A. (2014). How partial transparency influences the processing of compound words (Masters thesis).
17. Taft, M. & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal Verbal Learning and Verbal Behavior, 14*(6), 638-647.
18. Wang, H.-C., Hsu, L.-C., Tien, Y.-M., Pomplun, M. (2013). Predicting raters' transparency judgments of English and Chinese morphological constituents using latent semantic analysis. *Behavioral Research, 46*(1), 284-308.