The Age-Complicity Hypothesis: A Cognitive Account of Some Historical Linguistic Data

Marcus O'Toole (marcuso@dai.ed.ac.uk)
Division of Informatics, 80 South Bridge
Edinburgh, EH1 1HN Scotland

Jon Oberlander (jon@cogsci.ed.ac.uk)
Division of Informatics, 2 Buccleuch Place
Edinburgh, EH8 9LW Scotland

Richard Shillcock (rcs@cogsci.ed.ac.uk)
Division of Informatics & Department of Psychology, 2 Buccleuch Place
Edinburgh, EH8 9LW Scotland

Abstract

Shillcock, Kirby, McDonald and Brew demonstrate that there is a significant global relationship between word form and meaning across a substantial part of the lexicon of English. Here, 1705 words were studied to establish how their history in the language related to their participation in the correlation between meaning and form. It was found that the meaning-form correlation was significantly stronger for words with earlier dates of entry into the lexicon, implying that an individual word's meaning-form correlation may develop over time. Changes to individual words may be contingent on the word meanings and word forms in the rest of the lexicon.

Introduction

What is the relationship between a word's form and its meaning? And does age matter to the closeness of a relationship?

This paper addresses the second of these questions by building on previous work which addresses the first. The rest of this section introduces general background to the first question; the next section introduces the specific work upon which we build, and our current hypothesis; subsequent sections outline the methods and results of the current study, and draw a general conclusion.

Kelly (1995) suggested that "the hypothesis that phonological cues are unavailable or that people are not sensitive to them have no 'sound' basis in fact". On the one hand, the interaction between phonological and semantic representations has been widely discussed. On the other, the seemingly intuitive idea that there is a structure-preserving relation between these two aspects of words' representations has been largely ignored.

For instance, Dorffner and Harris (1997) report a model predicting that "although the mapping between word form and meaning is arbitrary… novel pseudo-words will prime concepts corresponding to words that are orthographically similar". They go on to discuss findings that showed that when English speakers are presented with pseudo-words, they tend to have associations with English words similar in terms of form. However, Dorffner and Harris dismiss possible relations between orthographic and semantic representations, and this implies that they see no connection between phonological form and the meaning of words. Yet no strong evidence is put forward to support this claim. Whereas priming effects could certainly exist in the absence of meaning-form relations, there is no reason to suggest that useful phonological cues to meaning cannot be utilised.

Indeed, Kelly (1992) had previously investigated how phonological cues—in terms of number of syllables, word duration, and pronunciation of certain syllables—were involved in category assignment. The study indicated that phonetic cues could be used to infer gender in a number of different languages, including French, Hebrew and Russian.

Continuing from this work, Cassidy, Kelly and Sharoni (*in press*) studied how phonological cues can be used to interpret gender, and how this information might be used by English speakers. A connectionist model was trained to classify novel names as male or female, solely on the basis of phonological cues, and succeeded in classifying 80% of names correctly. Experiments were undertaken which showed that four-year-old children had the ability to infer gender from pseudo-names, and that names that are phonologically typical of either gender are classified significantly more quickly than less typical examples.

In addition, Kelly (1998) studied "blend structure", which concerns the manner in which aspects of two words can be combined to produce a fresh word, in terms of cognitive and linguistic principles. Clearly, such cases help enhance the relationship between meaning and form for the words involved. Blended words such as *brunch* may become embedded in the lexicon due to their phonological evocativeness.

The Relationship Between Meaning and Form

Shillcock, Kirby, McDonald and Brew (2001) report a study in which they generated a semantic hyperspace from a large corpus of English, effectively defining the meaning of each word in terms of its contexts of occurrence. The semantic distance between any two words could be quantified using this hyperspace. In addition, they defined the phonological distance between any two words in terms of an edit distance (the number of features that it would be necessary to change to turn one word in the other).

For a set of 1733 monosyllabic, monomorphemic words, they obtained the meaning distance and the form distance between each word and every other word. They demonstrated that there was, overall, a significant relationship between these two distances: words that are phonologically more similar tend to be semantically more similar.

Further, for each word they calculated the correlation between the two pairs of distances between that word and each of the remaining 1732 words. This gave a value of r_{mf} (the correlation between meaning and form) for each word. When these individual values were ranked, important psychological differences between words emerged between different parts of the ranking. A high value for r_{mf} can be seen as the rest of the lexicon conspiring to support the relation between meaning and form for that particular word. Shillcock et al. claim that the communicatively important words predominantly occurred at the top of the list. In contrast, words with a small or negative r_{mf} value are often more specific and "propositional", for example, *priest* or *plight*. Shillcock et al. postulated that this relation is an example of a tendency towards structure-preserving mappings by the brain.

This ranked list of the relation between words' meaning and form provides the basis for the current study.

With reference to studies on the role of phonology and similarities between lexical neighbours, Shillcock et al. suggested that even very different words can be related in "a model that assumes the whole lexicon may influence the processing of any one word". They go on to demonstrate how the variability of monomorphemic, monosyllabic words in terms of length and phonological similarity can relate to semantic meanings, due to a tendency whose results resemble the compositionality normally present only at the higher levels of linguistic structure.

The Age-Complicity Hypothesis

Suppose the meaning-form correlation is, as hypothesised, a quantifiable aspect of a representational strategy employed by the brain. Then we can suggest that the correlation for each word is open to change over time, as groups of phonetically similar and semantically similar words become established. Further to this effect, it is also likely that words with strong individual meaning-form correlations are more likely to *remain* established, whereas words with weak individual meaning-form relations could be subject to semantic drift, and would be more likely to take on new meanings, losing their original ones. Therefore, the meaning-form correlation would tend to be strongest for words that have been longest in the lexicon. We call this prediction the Age-Complicity Hypothesis.

Method

The current study was based on the list of 1733 words ranked according to their individual relationship between meaning and form (r_{mf}), produced by Shillcock et al. Using information obtained from the Oxford English Dictionary on the first non-obsolete date of entry, a total of 1705 of these words were analysed, once words with no dictionary entry had been discarded. This produced a database of 1705 words with information on first non-obsolete date of entry and r_{mf} scores.

Correlation between r_{mf} rank and date of first entry was measured using Kendall's Tau B. The data were further analysed by comparing how the r_{mf} correlation differs within sections of the ranked list partitioned in terms of r_{mf} rank and age.

Results and Discussion

The Age-Complicity hypothesis predicted that words with long established meanings would have high r_{mf} values. This prediction was convincingly supported by these results. Figure 1 gives an overall view of the data, but although certain interesting features are visible, it is not possible to discern the trends in the data. Figure 2 charts averaged data, and trends become visible.

As Table 1 shows, there was a highly significant correlation ($\tau = 0.08$, $p < .001$) between the meaning-form relation and the first date of entry over all the data.

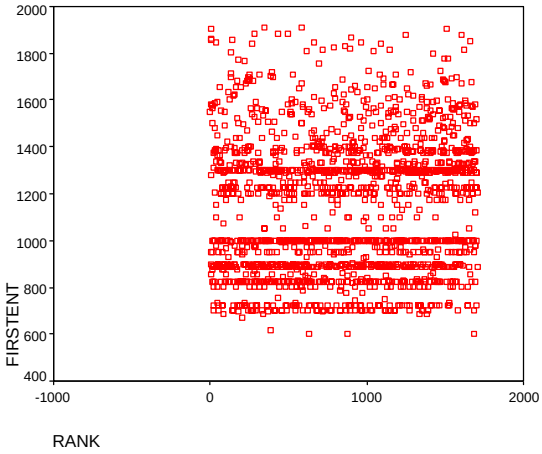

Figure 1: Graph of r_{mf} rank versus date of entry

On the basis of r_{mf} rank, the data were divided into three sections for further analysis: the top ranked 500 words, the middle ranked 500 words and the lowest ranked 500 words. The correlations between first date of entry and r_{mf} are given in Table 1.

There was a negative correlation for the top 500 words in the ranking: the words that were higher in the ranking by r_{mf} had the more recent dates of entry for this subset, contrary to the overall correlation. This feature of the results is probably a reporting phenomenon. The very top of the ranking by r_{mf} contains a number of items such as speech editing terms (*um*, *er*), swear words, and shortened proper names (*Mick*). These kinds of items may have been relatively unlikely to be written down—and hence given a date of first entry—in earlier times.

The middle section showed a non-significant correlation between rank order of r_{mf} value and date of first entry, although this time with a positive correlation. Finally, the lowest entries showed a correlation between rank and first date of entry which mirrored that in the total survey. In summary, despite the anomaly at the top of the ranked list, the results displayed an overall pattern of words' individual meaning-form relationships correlating with age, with words with a high value of r_{mf} having an earlier date of entry.

Table 1: Correlation (τ) between r_{mf} values and date of entry, sorted by r_{mf} rank

Entries 1-1705	Entries 1-500	Entries 600-1099	Entries 1206-1705
0.080 ($p < 0.000$)	-0.056 ($p < 0.033$)	0.032 ($p < 0.148$)	0.080 ($p < 0.004$)

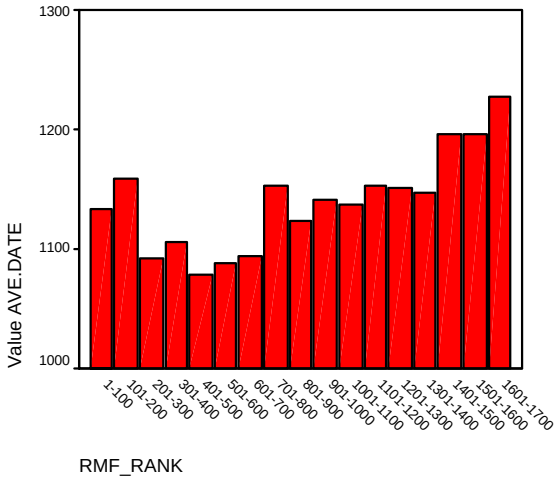

Figure 2: Graph of average first date of entry versus ranked r_{mf} entries

What happens when we think primarily in terms of dates? To provide another view onto the hypothesis, the data were again split into three sections, this time according to dates of entry. The three sections correspond to dates of entry between the years 601 and 897, 972 and 1297, 1303 and 1911, respectively.

Table 2 shows a non-significant (but negative) correlation between r_{mf} rank and date of entry for the section of words with the earliest dates of entry. The positive overall correlation is best reflected in the "middle-aged" words.

We have just seen, in Table 1, the reversal of the correlation for the more recent words, and the same explanation applies to Table 2. A good example is the new entry *yeah* (a colloquial form first recorded in 1905) with a very high r_{mf} value (ranked 8th). If this is due to a reporting phenomenon, then *yeah* could have been important in speech for a long time, and simply not captured in text, to be reported by the OED. On the other hand, it remains possible that *yeah* is just a recent, successful innovation. Then, the entry into the language of such a modified word could be attributed to their meaning-form correlation being stronger than their competitor, (*yes* is ranked 89th in terms of r_{mf}). Figure 3 charts the averaged data.

Table 2: Correlation (τ) between r_{mf} values and date of entry, sorted by date

All Dates 1-1705	Dates ranked 1-476	Dates ranked 599-1100	Dates ranked 1238-1705
0.080 ($p < 0.000$)	-0.043 ($p < 0.091$)	0.064 ($p < 0.023$)	-0.053 ($p < 0.045$)

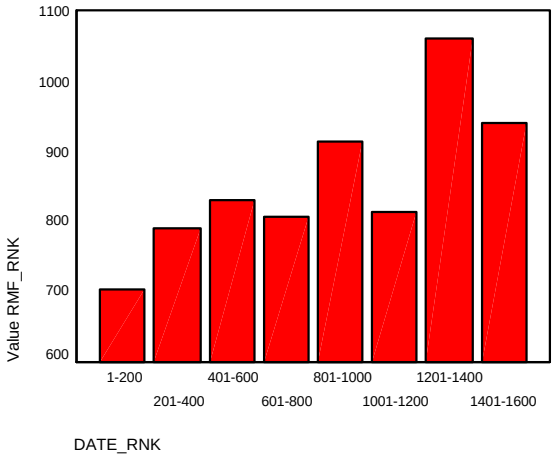

Figure 3: Graph of average r_{mf} rank versus date of entry.

Conclusion

From the results, it seems that there are anomalies amongst the (apparently) youngest, and oldest recorded words in our study. Nonetheless, a moderate interpretation of the Age-Complicity hypothesis has real support: the closeness of the relationship between form and meaning is related to its age.

More specifically, this study has produced evidence of a relationship between the history of an individual word in the language and that word's participation in the overall relationship between meaning and form in the lexicon. If a word has a high value of $r_{mf,}$ then it may be that it resists any change in its own meaning-form relationship; the rest of the lexicon is in effect supporting that relationship. Such a meaning-form relationship is adaptive; it means that the form and the meaning of the word can be partly inferred, one from the other—a clear advantage in language acquisition and processing. At the same time, that individual word may be helping to change the form and/or the meaning of words with weaker values of r_{mf}.

These findings have implications for studies of a number of aspects of human language. Principally, they offer data to substantiate an explanation of why some words become established in the lexicon while others do not. In other words, one of the contributing factors which can help a word become established is "sounding right", i.e. that its form resembles that of words with similar meanings.

We suggest that computational modelling might be used to simulate the data we have presented. It is not possible to obtain sufficient historical data to resolve all the possible reporting biases that may be present in data of the kind we have considered. Computational modelling may help to resolve some of these issues.

Finally, we have shown that it is possible to construe some of the data about language change from historical linguistics in cognitive terms, and specifically in terms of an adaptive relationship between meaning and form in the mental lexicon.

References

Cassidy, K.W., Kelly, M.H., & Sharoni, L. (in press). Inferring gender from name phonology. *Journal of Experimental Psychology: General*

Dorffner, G., & Harris, C. L., (1997) When pseudoword become words – effects of learning on orthographic similarity priming. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 185-189). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*, 349-364.

Kelly, M.H. (1995). The role of phonology in grammatical category assignments. In J.L. Morgan and K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 249-262). Mahwah, NJ: Lawrence Erlbaum Associates.

Kelly, M.H. (1998). To "brunch" or to "brench": Some aspects of blend structure. *Linguistics, 36*, 579–590.

Shillcock, R., Kirby, S., Mcdonald, S. & Brew, C. (2001). The relationship between form and meaning in the mental lexicon. Unpublished manuscript.