Phonaesthemic and Etymological effects on the Distribution of Senses in Statistical Models of Semantics

Armelle Boussidan (armelle.boussidan@isc.cnrs.fr)
L2C2, Institut des Sciences Cognitives-CNRS,
Université Lyon II, Bron, France

Eyal Sagi (ermon@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Sabine Ploux (sploux@isc.cnrs.fr)
L2C2, Institut des Sciences Cognitives-CNRS,
Université Lyon I, Bron, France

Abstract

This paper uses methods based on corpus statistics and synonymy to explore the role language history and sound/form relationships play in conceptual organization through a case study relating the phonaestheme *gl-* to its prevalent Proto-Indo European root, **ghel*. The results of both methods point to a strong link between the phonaestheme and the historical root, suggesting that the lineage of a language plays an important role in the distribution of linguistic meaning. The implications of these findings are discussed.

Keywords: Corpus statistics, Synonymy, Historical Linguistics, Sound/form relationships.

Introduction

Recent years have seen a surge in the use of statistical models to describe the distribution and inter-relation of concepts at the cognitive level and meanings at the linguistic level.[1] These models have been applied to a wide range of tasks, from word-sense disambiguation (Levin et al., 2006) to the summarization of texts (Marcu, 2003) and the tracing of semantic change (Sagi, Kaufmann, & Clark, 2009). They have also been used to model a variety of cognitive phenomena, such as semantic priming (Burgess, Livesay, & Lund, 1998) and categorization (Louwerse, et al., 2005).

In this paper we will explore the role that language history and sound/form relationships might play in conceptual organization using two methods – one based on corpus statistics (Infomap, Schütze, 1996) and the other based on synonymy (Semantic Atlases, Ploux & Victorri, 1998). Importantly, the use of both corpus-based and lexicon-based statistics allows us to examine these phenomena at two different levels – lexical meaning and language in use. This examination will highlight that even though a language can undergo drastic changes over time, some aspects of the underlying cognitive organization remain stable.

Many models based on corpus statistics (e.g., LSA, Landauer & Dumais, 1997; Infomap, Schütze, 1996; Takayama, et al. 1999; HAL, Lund & Burgess, 1996) are built around the assumption that related words will tend to co-occur within a single context with higher frequency than unrelated words. As a result, this pattern of word co-occurrence can be considered an approximation of the underlying organization of concepts.

The relationship between words and concepts can also be described in terms of closest semantic equivalents, synonyms. (Wordnet, Fellbaum, 1998; Semantic Atlases, Ploux, 1997; Ploux & Victorri, 1998). The Semantic Atlas (SA) is a geometrical model of meaning based on fine grained units of meaning called ‘cliques’. Each clique contains a series of terms all synonymous with each other.

While models that rely on measuring word co-occurrence might seem to be very different from those that are based on identifying clusters of synonyms in dictionaries, both approaches are distributional in nature and rely on very similar methods of investigation. Nevertheless, these approaches take somewhat different perspectives and examine different aspects of word distribution. Therefore, they may complete each other so as to reach a more complex and complete picture of how word meanings are anchored in language on the one hand, and how they relate to concepts on the other. Both synonymy and context participate in the architecture of meaning and in relating lexical items to a conceptual network.

We can use different types of data to enhance our understanding of language. For instance, following work by Firth (1930), Otis and Sagi (2008) demonstrate that the distribution of terms in a corpus is also related to the phonetic features of words known as *phonaesthemes*, sub-morphemic units that have a predictable effect on the meaning of a word as a whole. For instance, non-obsolete English words that begin with *gl-* are, more often than not, related to the visual modality (e.g., *gleam*, *glitter*, *glance*) whereas words that begin with *sn-* are usually related to the nose (e.g., *snore*, *sniff*, *snout*). More generally, it appears that some phonetic aspects of word form might be related to meaning and indicative of its conceptual underpinnings.

However, to properly utilize this new information it is important to understand how it relates to conceptual organization. For instance, phonetic similarity may be used as a cue for conceptual similarity. This suggests that phonaesthemes may be a specific case of a more general principle and that in contrast with the Saussurian tradition,

[1] As Jackendoff (1983: 95) notes, it is possible that “semantic structure is conceptual structure”. However, for the purpose of this paper we will assume that these two levels of representation are distinct.

language might incorporate an abundance of non-trivial relations between word form or sound and word meaning.

Another factor that governs these similarities is the history of the language – For instance, reconstructions of Proto-Indo European, the ancestor of many of the languages spoken in Europe and western Asia, suggest that it was a root-based language and as such incorporated many meaningful morpho-phonological clusters. Some of these may have survived through the generations and formed the basis for phonaesthemes. In this case, the survival of these specific clusters might indicate that they are linked with important aspects of cognitive organization. As a result, identifying and cataloging these phonaesthemes might provide interesting insights into some of the basic dimensions underlying the organization of concepts. In this paper we examine this question by contrasting the influence of phonetic similarity and the historical roots of words in the case of the *gl-* phonaestheme and its prevalent Proto-Indo European root, **ghel*.

**ghel/gl-*: A case study

Indo European (IE) or Proto-Indo European (PIE) is a reconstructed common original language covering almost all languages spoken from Europe to India and dated around the fifth millennium BC. It gives birth to ten families of languages including the Germanic branch, of which English is a descendant. 19th century comparative linguists carried out PIE's reconstruction by observing similarities across languages and with the help of mutation rules. They determined a semantic common denominator for each root. As a consequence, root definitions are often vague, imprecise and all-encompassing. This calls for caution on the semantic plane: while the senses of PIE roots might seem more vague than those used in modern day English word definitions, this could be an effect of the reconstruction process rather than a real semantic difference.

In English, the vocabulary inherited from PIE appears to form the genuine core of the language even though it represents a small proportion of it compared to loan words. For example, Watkins (2000) reports that the 100 most frequent words in the Brown corpus are PIE based. PIE was an inflected language following the structure Root + Suffix + Ending. Some derivations were made on the basis of inflected words. The root is thus the most stable unit although roots can undergo extension and words can derive directly from these extensions. In PIE consonant alternation conveys semantic content whereas vowel change is apophonic, that is, it expresses morphological functions (Philps, 2008a). Although sound patterns and orthographic patterns follow laws of change which are quite regular, the semantic content attached to them often survives these changes and re-establishes a connection with the new sound forms and orthographic forms. This pattern seems to be central in language change processes.

Watkins (2000) identified **ghel* as a PIE root meaning "to shine" with derivatives referring to colors, bright materials, gold (probably yellow metal) and bile or gall[2]. It produces a series of words denoting colors (e.g., *yellow* from the extended root **-ghel-wo-*), words denoting gold (e.g., *gold* from the zero grade[3] form **ghl-to-*), words denoting bile and gall (*gall* from the o-grade form **ghol-no-*) and most interestingly a bag of Germanic words related to light and vision starting with *gl-* (e.g., *gleam, glass*).

Researchers identified the phonaestheme *gl-* as relating to the "phenomena of light", to "visual phenomena" (Bolinger, 1950, pp. 119 & 131) and to the concepts "light" and "shine" (Marchand, 1960, p. 327). However, while many English words that feature this phonaestheme seem to have a meaning that is obviously related to the visual modality (e.g., *glow, glare, glisten*), some other words (e.g., *glue, glucose*) appear to be unrelated. Therefore, it seems that phonaesthemes are not absolute – not all words that feature them fit the conceptual pattern of the phonaestheme. A phonaestheme is therefore more likely to be a statistical cue to some general conceptual features of meaning.

However some apparently unrelated items may be associated to the central meaning of the *gl-* phonaestheme via the process of antonymy ("fire, to be warm", balanced by "cold" in *glace*, and "light" balanced by "dark" in *gloom*) or other similar processes. Concepts related to the tongue and swallowing appear in words such as *glottis*, or *glutton* which might be explained by a conceptual mapping from mouth to eye in terms of their open-close characteristics as described in Philps (2008b). Similarly there are *gl-* words that do not have a meaning related to light (e.g., "to cut" from the **kel-* root, " sweetness" from **dlk-u-*, "clay" from **glei-*, and "cold" from **gel-*).

Otis and Sagi (2008) demonstrated that it is possible to statistically validate the internal consistency of meaning that is at the core of phonaesthemes – i.e., that the group of words which feature a specific phonaestheme are also closer in meaning than a similarly-sized group of words that do not share a phonaestheme. Furthermore, priming experiments conducted by Bergen (2004) suggest that cognitive processing of linguistic stimuli is affected by phonaesthemes and that these effects cannot be fully explained as the result of either semantic or phonetic similarity.

As a result, it appears that there are two possible factors that might explain the relationship between phonaesthemes and word meaning – the historical root of the words, and cognitive processes that relate phonetic and semantic similarity. Importantly, these hypotheses are not mutually exclusive. One way to compare them is to examine how much of the relatedness between sound and meaning that

[2] **ghel-*, to call, shout and **ghel-*, to cut, are homonymic roots which do not appear in the '*gl-*' set of words and therefore will not be investigated in this paper.

[3] There are three grades in Indo-European grammar: the full grade in -e-, the o-grade, and the zero-grade (without vowel). Here the zero grade form of **ghel-* (full grade) is **ghl-*, and its o-grade is **ghol-*.

identifies a phonaestheme is attributable to the historical root and how much is attributable to phonetic similarity.

In other words, if the observed effect is due to the historical root **ghel* then it should extend equally to all words that resulted from that root, but not to words that resulted from other roots. Similarly, if the effect of phonaesthemes is primarily due to their phonetic similarity then the effect exhibited by the phonaestheme *gl-* should be restricted to words that begin with *gl-*, regardless of their PIE root, but should not extend to other words that originated from the **ghel* root. We will test this hypothesis using two different approaches. Firstly, we will employ the method developed by Otis and Sagi (2008). Because the cohesiveness of a word cluster is a measure of its inter-relatedness, we can use this measure to examine the relative role of the PIE root **ghel* and the phonaestheme *gl-* by comparing their relative cohesiveness. Specifically, we hypothesize that if the historical root **ghel* is the source of the phonaestheme *gl-* then the cluster of words belonging to the root should be more cohesive than the cluster of words that begin with *gl-*, and vice versa.

Secondly, we will examine clusters generated from the Semantic Atlases synonym database (Ploux & Victorri, 1998) and investigate whether *gl-* and non *gl-* sets have independent semantic status and sound/form within the **ghel* space and conversely for the **ghel* set within the *gl-*space.

Following our hypothesis, if the phonaestheme *gl-* has its roots in the PIE root **ghel*, then we would expect the average distance between words that come PIE root **ghel* and begin with *gl-* to be small compared to the average distance between words in other sets. In addition, we predict that the *gl-* set will be more cohesive within the **ghel* space than the whole, due to its phonetic unity, and that the **ghel* set will be more cohesive within the *gl-* space than the whole due to its historic unity.

Method

Materials

We identified PIE roots based on the work done by Watkins (2000). The lists of words starting with *gl-* were generated on the basis of the dictionary database for the SA and on the basis of the corpus for Infomap. A sample of words used in this study as well as their PIE roots (if known) can be found in Appendix A.

Using Infomap to measure cluster cohesiveness

The corpus

We used a corpus based on Project Gutenberg (http://www.gutenberg.org/). Specifically, we used the bulk of the English language literary works available through the project's website. This resulted in a corpus of 4034 separate documents consisting of over 290 million words. Infomap analyzed this corpus using default settings (a co-occurrence window of 15 words and using the 20,000 most frequent content words for the analysis) and its default stop list.

Computing Word Vectors

For our computational model we used Infomap (http://infomap-nlp.sourceforge.net/; Schütze, 1996), which represents words as vectors in a multi-dimensional space based on the frequency of word co-occurrence. In this space, vectors for words that frequently co-occur are grouped closer together than words that rarely co-occur. As a result, words which relate to the same topic, and can be assumed to have a strong semantic relation, tend to be grouped together. This relationship can then be measured by correlating the vectors representing those two words within the semantic space.[4] Importantly, as mentioned in Buckley, et al. (1996), the first factor identified by Infomap is somewhat problematic as it is monotonically related to the frequency of the term. Because of this we elected to omit it when computing word vector correlations.

For each occurrence of a target word type under investigation, we calculated a context vector by summing the vectors for the content words within the 15 words preceding and the 15 words following that occurrence. The vector for a word is then simply the normalized sum of the vectors representing the contexts in which the word occurs.

Measuring the cohesiveness of a word cluster

We measured the cohesiveness of a word cluster in a similar manner to that used by Otis and Sagi (2008). The cohesiveness of a cluster was defined as the average correlation of the vector pairs comprising the cluster – a higher correlation value represents a more cohesive cluster (*r* below). It is also possible to directly test whether the cohesiveness of a cluster is greater than that of another. For this purpose we used Monte-Carlo sampling to repeatedly choose 50 pairs of words from the hypothesized cluster and 50 pairs of words from a similarly size cluster chosen from the corpus as a whole. We used an independent sample t-test to test the hypothesis that the one of the clusters was more cohesive (had a higher average cosine) than the other. This procedure was repeated 100 times and we compared the overall frequency of statistically significant t-tests with the binomial distribution for $\alpha=.05$. After applying a Bonferroni correction for performing 50 comparisons, the threshold for statistical significance of the binomial test was for 14 t-tests out of 100 to turn out as significant, with a frequency of 13 being marginally significant. Therefore, if the significance frequency (*#Sig* below) of a candidate cluster was 15 or higher, then one of the clusters was judged as being more cohesive than the other.

Synonym clustering

Clustering was conducted using the Semantic Atlas synonym database, which is composed of several dictionaries and thesauri enhanced with a process of symmetricality (available at http://dico.isc.cnrs.fr/). For each list of words, one comprised of all words that start with *gl-*, and one comprised of all words derived from the PIE **ghel*, a semantic space is built on the basis of all synonyms and near-synonyms of the words. For *gl-* this resulted in a

[4] This correlation is equivalent to calculating the cosine of the angle formed by the two vectors.

list of 2198 words, and for words derived from PIE this resulted in a list of 1130 words.

The set of cliques containing all these synonyms is calculated. Correspondence factor analysis is applied to the matrix composed of words in the columns and cliques in the lines to obtain the coordinates for each clique (Ploux & Ji 2003). To split the space into clusters, a hierarchical classification is obtained via the calculation of the Ward's distance of cliques' coordinates. A word belongs to a cluster if all the cliques that contain it belong to this cluster.

Results

Word Cluster Cohesiveness with Infomap

We first computed the cohesiveness of the cluster of all words that have been identified as descendents of **ghel* and that of all words that feature the *gl-* phonaestheme. We also computed the cohesiveness of the cluster formed by their intersection, that is, the cluster of words that start with *gl-* and are descended from the **ghel* root. The results of these computations, as well as the cohesiveness of related clusters are given in table 1. Interestingly, all of these clusters show a higher cohesiveness than would be expected by chance alone, as is evident by the fact that all of the *#Sig* measures are above the chance threshold of 15.

Table 1 - The cohesiveness of the **ghel* PIE root and the *gl-* phonaestheme clusters.

N – cluster size; *r* – cohesiveness;
#Sig – number of significant t-tests compared to baseline

Cluster	*N*	*r*	*#Sig*
****ghel* words**	38	.15	100
***gl-* phonaestheme**	88	.097	75
****ghel* words starting with *gl-***	25	.25	100
****ghel* words not starting with *gl-***	13	.046	22
Non-ghel* words starting with *gl-***	17	.15	95

In order to answer our research question, we also compared the clusters to one another. Overall, the results follow the pattern indicated by the relative cohesiveness of the clusters as seen in table 1. The *gl-* phonaestheme as a whole forms a less cohesive cluster than either part of it that is descended from words with a **ghel* PIE root (*#Sig*=28, *p*<.0001) or the part of it that is descended from words with PIE roots other than **ghel* (*#Sig*=28, *p* < .0001). However, that same cluster is more cohesive than the cluster comprised of words with a **ghel* PIE root that do not begin with *gl-* (*#Sig*=30, *p*<.0001). Finally, the cluster formed by words that begin with *gl-* and whose PIE root is **ghel* is stronger than any of the other clusters. More specifically, it is stronger than both the cluster formed by words with a **ghel* PIE root (*#Sig*=55, *p*<.0001) and that formed by words with a PIE root other than **ghel* (*#Sig*=45, *p*<.0001).

The most cohesive part of the *gl-* phonaestheme therefore seems to be formed by words with a **ghel* PIE root. Nevertheless, it appears that the set of words starting with *gl-* with other PIE roots also form a cohesive cluster of meaning, even if it is somewhat weaker. This suggests there is more to the phonaestheme than merely a historical root.

Interestingly, the weakest cluster identified in this analysis was formed by words with a PIE root of **ghel* that do not begin with *gl-*. One possible interpretation is that those words having gone through a variety of languages (eg., Greek, Sanskrit) have been subjected to many semantic and morpho-phonological changes creating a disparity in the set. However *gl*-words that relate to light and vision have mostly gone through Germanic, which may explain their high semantic and morpho-phonological cohesiveness.

Word Cluster Cohesiveness and Prototypicality with the SA

***ghel clustering**

Our analysis of the **ghel* data resulted in three main clusters (and a plethora of weak ones). For **ghel*'s main cluster we obtained 649 synonyms of which 609 were relevant[5]. This main cluster is further divided into three sub-clusters and included the central senses of **ghel*: The first sub-cluster (362 terms) relates to the visual modality and to shining. It also contains most *gl-* items (with the exception of terms related to *glide* in cluster 3 as well as *gladden* and *gloaming* in separate clusters). The second sub-cluster (149 terms) relates to melancholy and colors. The third sub-cluster (98 terms) relates to bile, gall and emotional states mapped onto them metaphorically. The last two sub-clusters are significantly separated from the first one.

gl- clustering

From the unstemmed total of 230 *gl-* words, 74 come from PIE **ghel* (32,17%) while in the stemmed list of 106 items 23 do (21,69%). The higher percentage of *gl-* words coming from the root **ghel* in the unstemmed list shows that these items are highly productive in terms of derivation and composition.

The strongest cluster of *gl-* was comprised of 1048 synonyms and was divided into three sub-clusters that form a total of 883 relevant synonyms. The strongest sub-cluster (678 terms) relates to the visual modality. The second sub-cluster (124 terms) relates to gloom and melancholy, and the third (81 terms) relates to the globular shape. Other significant clusters relate to the meanings “glide”, “glue” and “glove”. All other clusters are small and specialized.

Prototypicality

In the **ghel* space, one sub-cluster gathered most of *gl-*based words (38 out of 43) and the other two gather most of non-*gl*-based words. The meanings of light and vision are clearly correlated with the *gl-* phonaestheme, while non-*gl-*item clusters inherit the bulk of other semantic contents associated with **ghel*. The historic root clearly evolved into a *gl*-based conceptual network related to light and vision,

[5] 'Relevant' synonyms are in the cliques that only belong to one given cluster. Conversely some highly polysemous cliques belong to several clusters.

Table 2 - Prototypicality in the strongest clusters of the **ghel* space and the *gl-* space

***ghel**	*Sub-Cluster*	*In # of cliques*	*%*	***gl-***	*Sub-Cluster*	*In # of cliques*	*%*
glow	1	55	19%	**gleam**	1	59	10%
glitter	1	48	16%	**glow**	1	55	10%
glowing	1	48	16%	**shine**	1	51	9%
melancholy	2	52	53%	**gloomy**	2	85	64%
sad	2	25	25%	**dismal**	2	39	29%
yellow	2	17	17%	**dark**	2	37	28%
gall	3	58	81%	**globe**	3	20	50%
virulence	3	19	26%	**ball**	3	13	33%
bitterness	3	18	25%	**orb**	3	11	28%

while secondary meanings were distributed across non-*gl-* items.

Clusters classify words in decreasing order of importance: the ones that belong to a high number of cliques are considered to be more prototypical. Table 2 shows the 3 most prototypical items of **ghel* and *gl-*'s main clusters. The percentage denotes the number of cliques the item belongs to on the total of cliques composing the cluster.

In the *gl*-space, one sub-cluster gathers most **ghel*-based words (65 out of 82), while the two others gather a smaller number of then (8 in sub-cluster 2, and 9 in sub-cluster 3). Again the first sub-cluster is the largest and corresponds to the central meaning of the *gl-* phonaestheme, while the two others relate to antinomic and secondary meanings. The phonaestheme clearly divides into a major conceptual unit versus minor units mostly unrelated to the historic root.

Cohesiveness and semantic distances

We used independent samples t-tests to examine the semantic cohesiveness of **ghel* words within the *gl-* space and similarly for *gl-* words within the **ghel* space.

In the **ghel* space, the average semantic distance within the *gl-* cluster is lower than the average distance between the *gl-* and non-*gl-* clusters (M_{intra}=0.39, M_{inter}=1.65, $t(219)$=9.86, p<.0001). However, no significant difference was found between the non-*gl-* cluster and the overall **ghel*-set (M_{intra} =1.85, M_{inter}=1.66, $t(93)$=0.61, *n.s.*). Non-*gl-* items are therefore disparate and less cohesive than the *gl-* phonaestheme.

In the *gl-* space, words that have the same PIE root show higher cohesion than words that do not (M_{intra}=0.15, M_{inter}=1.81, $t(556)$=9.82, p<.0001). Words that are **ghel* based are more cohesive than the whole *gl-* space as the average distance between the **ghel* set and other PIE roots is lower than the internal average distance within the **ghel* set. (M_{intra}=0.13, M_{inter}=3.31, $t(187)$=2.36, p<0.05)

These results are congruent with the previous analysis, as the strongest cohesiveness is found in the set that is both *gl-* and **ghel* based.

General Discussion

In this paper we show that, in the case of *gl-*/**ghel*, historical (here PIE) and morpho-phonological (here phonaesthemes) aspects are autonomous but highly correlated and that both have a tangible impact on word meaning. More specifically, we showed that phonaesthemic sets have a higher cohesiveness within historical sets and historical root sets have a higher cohesiveness within phonaesthemic sets.

These results suggest that the lineage of a language plays an important role in the distribution of linguistic meaning. In particular, the phonaestheme *gl-* seems to be based on the PIE root **ghel*. It therefore seems clear that, at least in some cases, historical information influences the distribution of word meaning in non-trivial ways. One reason for this could be that lexical items are linked to conceptual networks that are rooted in history. By incorporating historical and etymological information into statistical models such as word-space vectors or clique-based synonym sets we might improve their performance.

The conceptual networks visible for *gl-* words keep traces of older semantic content, notably the fact that verbs starting with *gl-* and related to light or vision can have two arguments, an animate one (as in *glance*) or an inanimate one (as in *glow*). This particular aspect relates vision to light emission and participates in creating a semantic unity contrary to modern beliefs that clearly separates emitting light from perceiving it (cf. Philps, 2008a). However, at this point it is unclear what the cognitive value of these semantic traces is and how it relates to the role of language as a means for decoding the world.

Interestingly, some words of obscure origin have high productivity although they cannot be traced back to PIE. One example of this is the word *globe* which seems related to the visual modality, though there is no historical evidence for such a connection. This gives rise to a new question – How do newly formed words find their place within an existing conceptual network? It may be that new additions to the vocabulary are likely to be patterned after existing words in a manner that makes them compatible with the rest of the set. New words which contain an existing phonaestheme are likely to fit its conceptual pattern as well.

In this paper we focused on examining the role that language history and sound/form relationships might play in conceptual organization in the case of **ghel*/*gl-*. Our results suggest that such analyses can provide important insights into the inter-relation of semantic concepts. In particular, it seems some aspects of meaning may be more stable than others. However, at this point it is not clear whether this stability is attributable to some fundamental characteristics of human cognition or to the broader social contexts in which language is used.

Moreover, using this information and integrating it with current distributional models is not a trivial task, and several different routes seem to present themselves. A possible route might involve defining a new, etymological, index that could be used to enrich current models of conceptual organization and semantic similarity. Finally, it seems that a better understanding of how languages change and evolve might lead to a better understanding of the interrelation between language, culture, and cognition.

References

Bergen, B. (2004). The Psychological Reality of Phonaesthemes. *Language*, 80(2), 291-311.

Bolinger, D. (1950) Rime, assonance, and morpheme analysis. *Word* (6), 117-136.

Buckley, C., Singhal, A., Mitra M., & Salton, G. (1996) New retrieval approaches using SMART:TREC4. *Proceedings of the Fourth Text Retrieval Conference (NIST Special Publication 500-236)*, 25-48.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25, 211–257.

Fellbaum, C. (1998) *WordNet, an electronic lexical database*. MIT Press, Cambridge, MA.

Firth, J. (1930) *Speech*. London: Oxford University Press.

Jackendoff, Ray, 1983. Semantics and Cognition, Cambridge, Massachusetts: MIT Press.

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.

Levin, E., Sharifi, M., & Ball, J. (2006) Evaluation of utility of LSA for word sense discrimination. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City. 77-80.

Louwerse, M. M., Hu, X., Cai, Z., Ventura, M., & Jeuniaux, P. (2005). The embodiment of amodal symbolic knowledge representations. In I. Russell & Z. Markov (Eds.), *Proceedings of the 18th International Florida Artificial Intelligence Research Society*, pp. 542–547. Menlo Park, CA: AAAI Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28*, 203-208.

Otis, K. & Sagi E. (2008) Phonaesthemes: A corpora-based analysis. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society.*

Marchand H. (1960) *The categories and types of present-day English word-formation.* AL: University of Alabama Press.

Marcu, D (2003) Automatic Abstracting, In Drake, M. A., (ed), *Encyclopedia of Library and Information Science*, pp. 245-256.

Philps, D. (2008a) Sons et lumières: le marqueur sub-lexical <*gl-*>. In G. Girard-Gillet (ed), *L'envers du décor, Études de linguistique anglaise*. Avignon, Publication des Presses de l'Université d'Avignon, pp. 24-43.

Philps, D. (2008b) From mouth to eye. in A. Smith, K. Smith & R. Ferreri (eds.), *The Evolution of Language*, Singapore: World Scientific Publishing, pp. 251-258.

Ploux, S. (1997) Modélisation et traitement informatique de la synonymie. *Linguisticae Investigationes*, 21(1):1-28.

Ploux, S. & Ji, H. (2003) A Model for Matching Semantic Maps between Languages (French/English, English/French), *Computational Linguistics*. 29(2):155-178.

Ploux, S., Victorri, B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *TAL*, 39, n°1.

Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In Basili R., and Pennacchiotti M. (Eds.), *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*. Athens, Greece.

Schütze, H. (1996) *Ambiguity in language learning: computational and cognitive models*. CA: Stanford.

Takayama, Y., Flournoy, R., Kaufmann, S. & Peters, S. (1999). Information retrieval based on domain-specific word associations. In Cercone, N. and Naruedomkul K. (eds.), *Proceedings of the Pacific Association for Computational Linguistics (PACLING'99)*, Waterloo, Canada. 155-161.

Watkins Calvert. (2000). *The American Heritage Dictionary of Indo-European Roots.* Second Edition. Houghton Mifflin Harcourt Compagny.

Appendix A – Sample words used in this study

PIE Root	Words
*ghel	yellow, melancholy, gulden, guilder, gowan, gold, glow, gloss, gloat, gloam, glitter, glister, glisten, glissade, glint, glimpse, glimmer, glide, glib, gleg, gleeman, gleed, glee, glede, gleam, glaze, glass, glare, glance, glad, gill, gild, gall, felon, cholera, choler, chloroform
*Dļk-u-	glucose, glycerine
*gel-^{2}	Glace
*gladh-	glabrous
*glei-	glue, gluten, glutinous
*glôgh-	glossa, glottis
*gwelə-^{2}	gland, glans
*kel-^{1}	gladiator , gladiolus
*kelə-^{2}	Glairy
*lep-^{2}	Glove
Unknown root	glacier, glade, glam, glamour, glaucoma, glean, glebe, glen, gloaming, globe, gloom, gloriosa, glory, glout, glucinum, glum