Studying the semantics of loanwords which have near synonyms in the host language: Psycholinguistic and multidimensional corpus representation methods

Joanna Rączaszek-Leonardi
(Faculty of Psychology, University of Warsaw)

Introduction

The issue of why certain linguistic forms are selected in the process of cultural evolution and persist in a language, while others perish, has fascinated researchers for many decades (e.g., Traugott & Dasher, 2002). With language being a multisystem and multiscale phenomenon (Rączaszek-Leonardi, 2009; Rączaszek-Leonardi, 2013), one should not expect the causes for the above selective processes to be simple. They will include pressures from many systems and processes unfolding on many time-scales. Some of them will have to do with the ratio of perceptual saliency/production costs (Lindblom et al., 1984; Oudeyer, 2006), others with learnability criteria (Deacon, 1997), and still others will be due to the systemic level of linguistic structures: fulfilling a distinct, well-defined, and necessary function in the system. The latter pertains both to the level of interaction in dyads (how well a term establishes common ground in communication), interaction in groups (if a term serves to define and distinguish a group, Roberts, 2010) and to the individual level (if a term fills some semantic need or stabilizes a cognitively motivated categorization).

Thus forms will be more likely to stabilize within a language if they: i) are easier to produce, ii) are less confusable with other forms, iii) are easier for children to learn, iv) are better suited to fulfill the communicative needs of dyads, v) match better the lexical representation, or conceptual structures, vi) fulfill a sociolinguistic role, such as group identification or group contrasting. It is important to note that these constraints act simultaneously on linguistic structures. Sometimes a given structure might not be optimal on one level if it serves an important enough function on another. For example, an expression that is relatively difficult to pronounce and learn will still be retained if it serves an important semantic, interactional, or sociolinguistic function.

The processes listed pertain both to the native forms of a language and to loanwords. One significant difference, obviously, lies in the differing word origins and in a "recognition of being foreign" in the case of loanwords, which in itself might be a feature that is perceptually, semantically, or sociolinguistically important in learning, designating, or group identification.

In this project we approach the question of form selection from a novel angle. We make use of an interesting phenomenon: the persistent co-existence of synonymous (or near synonymous pairs of words, one of them borrowed and the other native. This phenomenon (which is quite ubiquitous in languages, contrary to the commonly held belief that a foreign word is borrowed only when there is no suitable native term), motivates the following questions: Why, if there is a native word for a concept, is a loanword adopted at all? Is it just that laziness prompts speakers to adopt what they frequently hear, instead of searching for a native word? If we were to assume that this is so, then why is such "laziness" greater in the case of some words and not others? And what does this "laziness" consist in, what are its psychological mechanisms? Finally and crucially: why, then, does a loanword begin to be used in the first place, before it ever becomes frequent in a language? Why, when facing a certain semantic "need", due for instance to the emergence of a novel concept, will a speaker use a loanword rather than invent a "new" native word? Might there be something in the foreign form itself that facilitates its adoption (and then perhaps adaptation) as a loanword?

As we have noted, all the factors listed above in i) to vi) will be at work also in cases of loanword / native word near synonymy. Accounting for all of the factors might not be easy, as some of them – as mentioned above – might counter the influence of others. Yet in our work we will try to approach as many of the factors as possible, using a broad array of methods: linguistic-descriptive case-studies of loan-native pairs, psychological experiments using various psycholinguistic methods, and qualitative and quantitative analysis of multidimensional semantic representations constructed from large text corpora.

The analyses will focus mostly on the first five factors, leaving out the sociolinguistic causes of using loanwords. There will be a particular emphasis on factor v), the match between the form of a word and the representative structures of its meaning – which, as we will see in the discussion on sound symbolism, may also impinge on (and be stabilized by) iii) child learning processes. The causal role of this factor rests on the assumption that features of a form of a word may have associative semantic relations with its meaning.

This assumption that some linguistic forms may better match the semantic functions of words goes against the well-established view of the arbitrary relation between form and meaning in language. Before we move on to describing the methods to be used to study the presumed impact of form on interpretation, we will briefly address the reasons that justify some qualification of the arbitrariness assumption. Next we describe the planned psycholinguistic studies and the planned multidimensional semantic representation analyses. Finally, the relation between the three types of investigations within this project is explained.

Non-arbitrariness of linguistic forms

The claim about arbitrariness of linguistic forms, i.e., that the choice of a word is not determined by its meaning, has been an important element of natural language theories (e.g., de Saussure, 1916/1983). Hockett (1960) considered it to be one of the defining features of language. The idea of this independence is in an intuitive way supported by the often quoted fact that the same things have different names in different languages, or by the existence of polysemy and homonymy in language.

De Saussure himself recognized that arbitrariness is limited, claiming that a sign has to be phonologically plausible in a given language, and that the free variation of signs is *checked by historical and social factors* (Chandler, 1995). These historical and social factors are emphasized by others: "As Lévi-Strauss noted, the sign is arbitrary a priori but ceases to be arbitrary a posteriori – after the sign has come into historical existence it cannot be arbitrarily changed (Lévi-Strauss 1972, 91). "Thus (…) every sign acquires a history and connotations of its own which are familiar to members of the sign-users' culture" (Chandler, 1995). The latter process is often seen as "conventionalization".

In other words, the very use of a structure in linguistic interactions stabilizes it. The latest psychological research strengthens earlier stipulations that the selection of structures in this conventionalization-by-use process may be guided, in part, by the match between the physical properties of words and the properties of their referents, i.e. by semantic factors. This phenomenon has been termed "sound symbolism", "phonetic symbolism" (Brown, 1958), "natural expressiveness of the sound" or "iconicity of form-meaning pairings" (see e.g., Perniss et al., 2010). The form-meaning arbitrariness concerns different levels of language structure, not just the lexical one: sequential order of words may symbolize temporal ordering (*veni, vidi, vici*); proximity of enumeration may reflect semantic closeness (Bybee, 1985, after Perniss et al., 2010). A very interesting line of research shows the reliable use of voice amplitude and speaking rate consistently with the described event (going up vs. down, or going fast vs. slow) (Shintel et al., 2006). Prosodic contours, whose function traditionally has been limited to syntactic marking or emotional expression, may also carry semantic information (e.g., fundamental frequency of words referring to bigger objects was lower than for words referring to smaller ones, Ohala, 1994). In the present work, however, we need to focus on the lexical level. Studying near-synonymous loan- and native words requires us to scrutinize their morphological and phonological structure and its possible influences on interpretation.

Studies of these phenomena were quite frequent in the early years of scientific experimental psychology. In the now classic study, words *kiki* and *bouba* (Köhler, 1929) or *takete* and *maluma* (Köhler, 1947) were shown to match, respectively, jagged and rounded forms across cultures. Later studies, which aimed at proving the universality of sound symbolism, consisted in presenting the participants with pairs of words corresponding to opposite pairs of concepts, such as "love–hate", in an unknown language with their native language translations. Participants were to guess which of the presented pair of foreign labels was the name for which concept. The hypothesis was that matching beyond chance would mean that there are some cues to the semantics of a word in the form of the word itself. This study was first done on English speakers with Japanese words (Tsuru & Fries, 1933), and later with Hungarian (Allport, 1935), Polish (Rich, 1953), and French (Peterfalvi, 1970) words (for details see Allott, 1995).

Intuitively obvious cases of form-meaning correspondence can be found in onomatopoeic words, whose number has usually been judged to be rather small in language. However, another sound-symbolic phenomenon quite similar to onomatopoeia, known as phonaesthesia, seems much more frequent.

Phonaesthesia is the "attribution of common elements of meaning or connotation to certain sound sequences, especially consonant clusters, for example initial *sl-*, as in *slow*, *sleep*, *slush*, *slide*, *slip*" (Oxford Dictionaries Online). Such sound sequences are obviously different from morphemes. Phonaestemes are judged to be rather frequent in languages, although the data on them are rather anecdotal and not always convincing.[1] Inventories have been made for many languages, see, e.g., Abelin (1999) for Swedish, or an inventory if iconic words in Japanese (Atoda & Hoshino, 1995, after: Perniss et al., 2010).

[1] For example, Perniss et al. (2010) lists the ending *-ack* as associated with meaning "forceful, punctuated contact". While some support for this claim can be found (*smack*, *whack*) there are dozens of words ending with *-ack*, which do not have such connotation (*quack*, *pack*, *track*). The specificity and reliability of the form cues to meaning should also be analyzed.

Phonaesthesia is explained by the existence of similarity (or “natural association”) of experiences evoked by a form and its meaning. Experiences can be either evoked by perceptual properties of the produced sound or by the properties of the gestures that produce the form. It is claimed that some of these systematic relations hold across languages (such as back vowels and voiced consonants being associated with larger size, front vowels and voiceless consonants with smaller size, see Sapir 1929, Thorndike, 1945, Ultan, 1978), which by some researchers was later explained through the association of shape of vocal organs and shape of objects and then generalization to the sounds produced by vocal organs moving in a given way (e.g., Ramachandran and Hubbard, 2001).

Research on iconicity in language seems to have grown less popular in the second half of the last century, especially when judged in terms of English language publications. This may have been due to the dominance of generativist approaches to language which advanced theories of language processing, in which lexical, syntactic, and phonetic levels of language representation were imagined to be stored separately and the processes within them thought as informationally encapsulated (e.g., Forster, 1979, Fodor, 1983).

The last 20 years, however, have seen a definite trend towards demodularizing language components (MacDonald et al., 1994) and allowing different kinds of information to impinge on each other in language processing (see e.g., Vigliocco and Kita, 2006). Perhaps this is one of the reasons why the topic of form iconicity and sound symbolism is seeing a resurgence in research. A re-examination of form iconicity across languages shows that this phenomenon has been underappreciated in the past, possibly because research was more frequently conducted on languages in which onomatopoeic, phonaesthetic, or otherwise iconic words are relatively infrequent. Perniss et al. (2010) claim that in some languages they are quite widespread – e.g., a Japanese dictionary of sound-symbolic forms lists over 1700 words (Atoda & Hoshino, 1995, after: Perniss et al., 2010).

New sound symbolic phenomena are being discovered and others reconfirmed. The sound symbolic properties of back and front vowels being conductive of size information have been confirmed, for instance, by Shinohara et al. (2010), while some novel relations have also been proposed, e.g., a relation between voicing and dirtiness in Japanese (Kawahara & Shinohara, 2011). Several other studies have linked the manner (difficulty, effortfulness) of pronunciation of an entire word to the evaluative feelings and the experience of novelty. Reber et al. (2004) showed that high processing fluency is experienced as positive, while Song and Schwartz (2009) demonstrated that objects with names that are difficult to pronounce tend to be associated with higher risk, both desirable and undesirable, and that this influence seems to be mediated by the experience of novelty of the item.

With newer methods available, researchers are beginning to uncover sound-symbolic effects also in on-line language processing. Perniss et al. (2010) mention several studies in which it was demonstrated that the form of a word actually affects lexical access on-line, facilitaing processing for words whose form matches the experiences of their meanings. There also seems to be a growing agreement that the form of words facilitates learning processes (Imai et al., 2008; Nygaard et al., 2009), even for foreign language forms, which seems to point to universal tendencies in sound-symbolism, congruently with what Köhler (1929) and Sapir (1929) claimed almost a century ago. Learning effects are detected for adults, but more importantly, sound symbolism or iconicity may also help children in the acquisition of language. Using an artificial verb learning task, Yoshida (2012) demonstrated that children benefit from sound–meaning correspondences for verb learning, both in Japanese and in English. This might be a clue to one important mechanism through which sound symbolism may stabilize in language evolution.

The development of brain imaging techniques has made it possible to gather another type of evidence for the psychological reality of sound symbolism (Kovic et al., 2010). Event related potentials (ERPs) were found to show a characteristic negative peak (N200) for congruent matches of a word form and the picture of an object (which, the authors claim, shows that participants were sensitive to sound-symbolic contingencies) and a slightly increased (albeit localized) mismatch negativity (N400) for incongruent label-shape stimuli.

Since knowledge about sound symbolism has application in marketing (designing product and brand names, e.g., Lowrey & Shrum, 2007) we will probably soon see a surge in research on this topic using sophisticated methods and experimental designs. As an example of the fomer we can point to Aramaki et al. (2012), an abstract published after the CogSci 2012 conference, in which machine learning methods were used to detect and learn matching between phonetic and semantic features. A generator of new sound symbolic forms was created and when tested showed an impressive 88% congruence with human judgments.

Summarizing the above, we can risk the conclusion that, contrary to the arbitrariness assumption, forms are not in fact fully independent of meanings. To be sure, a fair degree of arbitrariness is obviously present in language and thus one should think about sound symbolism as constraints on possible forms rather than as determinants of forms. Our study of factors influencing the adoption and adaptation of foreign words assumes that some of these factors are due to the form itself. The research results briefly described above help ground this assumption and suggest psycholinguistic methods that can be employed to identify the specific factors.

Planned research

Our project is truly interdisciplinary and studies will be based on a "triangulation" of methods: linguistic analysis, psycholinguistic experiments, and quantitative corpus analysis. All the methods will seek to discover semantic factors that determine loanwords fate and characteristics, while at the same time serving as validations for the other methods. The methods of strictly linguistic analysis have been described elsewhere. Below we describe the planned psycholinguistic experiments and quantitative corpus analyses.

Psycholinguistic experiments

Despite quite a number of psycholinguistic studies having been conducted on sound symbolism, surprisingly few studies on near synonyms can be found within the psycholinguistic methodology. More often we find linguistic, qualitative, case based methods applied to assess how the semantic prosody of words is utilized (e.g., Nagórko, Łaziński, Burkhard 2004, Xiao & Mcenery, 2006). As mentioned before, we are abstracting away from some of the factors which may be important for preserving a certain form in language, which are of a more sociolinguistic nature (such as status marking by using loanwords, etc.). They are interesting in themselves, as a word's origin may influence speakers' perceptions in differing ways (e.g., Latin origin may suggest sophistication, aptly or not, while English origin may signal someone "up on the latest trends"; both features may be perceived as positive or negative, depending on the social context). However, the studies below may uncover certain semantic features evoked by foreign forms, which can serve as hypotheses for later "speaker perception" research.

With the psycholinguistic methods we aim to gain information at two levels:

i) having data from groups of people for each word, we will have an idiosyncratic (but population-generalizable) characterization for each pair of words,

ii) having data for a number of words we can attempt generalizations about certain features of loanwords or about the contrast between loanwords and their native synonyms.

In the first case the psycholinguistic data will help linguists trace better the nuances in near synonymous loan–native word pairs. In the second case we may begin to uncover the ‘meaning of being a loanword’, i.e. how ‘being foreign’ biases a form's interpretation. We expect that these biases may be specific to the perceived origins of the words. These generalizations should be performed also within categories of words distinguished on the basis of their origin.

Some of the the very few experiments that may be useful in designing the present research are studies of sound symbolism that tried to assess semantic traits that are evoked by forms that are easy and difficult to pronounce (Reber et al., 2004, Song & Schwarz, 2009). Before passing on to our proposed research, let us briefly describe these latter studies, summarizing the findings and pointing to the problems in the methods used.

Song and Schwarz (2009) had participants make decisions about objects based on made-up non-word names. The names were either easy to pronounce or difficult to pronounce. Participants rated food additives as more/less hazardous for one's health and as old/new, just on the basis of their made-up names. Only the difficulty of pronunciation was pre-tested. Subjects judged additives that had easy-to-pronounce names (such as *Magnalroxate*) as less harmful than those with difficult-to-pronounce names (such as *Hnegripitrom*). In a follow-up study, easy- or difficult-to-pronounce names of amusement-park rides were presented (again without any additional information about the rides), asking some of the participants to identify rides that are adventurous and exciting and others to identify rides that are too risky, and hence likely to make them feel sick (manipulating the instruction).

The authors claim their results show that low processing fluency increased perceptions of desirable as well as undesirable risks. They also introduce the factor of novelty as a mediating factor. Thus the pattern of results is interpreted as showing that “fluency influences risk perception through its effects on the perceived novelty of the stimuli”, rather than fluency-elicited affect influencing the results directly.

This study has serious shortcomings which should be avoided in planning new research on the topic of form influence over meaning interpretation. First, no other features of the words were controlled for, apart from their “difficulty to pronounce” and length. Thus the possible effects of other features, such as strongest associations, the presence of round vowels, or voiced vs. unvoiced consonants, which the previous literature has shown to be sound-symbolic (i.e. be correlated with semantic interpretations, see above) were entirely neglected. In the example given (*Magnalroxate* vs. *Hnegripitrom*) we can easily see that the words differ also in terms of such factors (which do not have to be directly connected to facility of pronunciation). Besides, *Magna* may be associated with magnesium, a known, and healthy, food supplement, or *Magnalroxate* with *Maalox* (a brand name antacid, know to bring relief in excessive stomach acidity). These, of course, are only stipulations, but criticisms of his sort could be avoided by conducting more careful pretests of the stimuli. The case of the amusement park rides is similar. The "easy" names were: *Chunta*, *Ohanzee*, and *Tihkoosue* and the "hard-to-pronounce" names were *Vaiveahtoishi*, *Tsiischili*, and *Heammawihio*. Again, these names are full of sound-symbolic traits (rounded vowels vs. front vowels, etc.) that were already mentioned by Köhler (1929) in his *kiki* and *bouba* study, as suggesting pointy and round things respectively. One can only hope that the rest of the pairs were more carefully chosen (unfortunately, the reminder of the pairs used were not provided). This critique undermines the exclusiveness of the "difficulty of pronunciation" factor, but, obviously, does not undermine the sound-symbolism principle (in which difficulty of pronunciation may be a factor).

Another problem with Song and Schwartz's (2009) study is that even though the difficulty of pronunciation was the independent variable, stimuli were presented in written form and the instructions did not ask participants to read them aloud. There are studies which show automatic activation of phonological form by graphic word representation but it is surely less strong than in the case of actually pronouncing a word.

Finally, it also has to be noted that the methods applied by Song and Schwartz (2009) can be easily criticized on the grounds of the ecological validity of the task: after all it is not a natural situation to rate the physical traits of objects solely on the basis of their (novel, invented) names – in real life we usually know more about the options we have to choose from than just the names of those options. Subjects may thus have easily seen through the experimenter's intentions and exaggerated the sound-symbolic features of the labels.

Song and Schwartz's study thus leads to several important conclusions: 1) studies need to be more ecologically valid, 2) if claims about the link between the sound or manner of pronunciation and semantic interpretations are to be made it would be useful to have participants actually hear or/and pronounce the words, and 3) the influence of form over meaning (or rather the linking of form to meaning) should be sought not only at the level of motor effort when pronouncing the words, but also in more conventionalized sound symbolism in language. As usual, there are many factors that might be at play and in our studies we will try to take this into account.

Study 1: Free associations

This study will obtain free associations for both loanwords and native words from the pairs selected in linguistic analyses (hereafter called "target words"). Next, the sets of free associations will be analyzed qualitatively and quantitatively. The quantitative analysis will consist in comparing frequency distributions of free associations.

We will ask 2 groups of participants to mention the 3 words that first come to mind in connection with the target words. The target words will be presented in booklets, one per page, with randomized order of presentation. Each group will have half loanwords and half native forms, crossed between groups (i.e., no participant will generate associations to the two forms (borrowed and native) referring to the same (or similar) concept.

Since it is probable that the same associations will recur across participants, the first analysis will consist in comparing the relative frequencies of the particular associations. The relative frequency and the semantics of associations will be informative in themselves, but also could serve as validation data for corpus studies performed by the linguists in this project and for the outcomes of the analyses with multidimensional semantic representations, described below.

A variant of the free association study would be to obtain the descriptors for the typical objects or features described by the chose words. This means giving the instruction: “Please describe a typical *helikopter*” and “Please describe a typical *śmigłowiec*” (both words denoting the same object in Polish, i.e. a helicopter), or “When someone *partycypuje* in something it means that…” and “When someone *uczestniczy* in something it means that…” (both words referring to the same activity, i.e., participating in something), instead of “Say the first three words that come to mind…”.

Study 2: Assessment of semantic prosody

In this study concepts evoked by loanwords and native words will be evaluated while presented on their own. Words will be presented in the same fashion as in Study 1, also to two groups of participants. Under each word Likert scales will be presented with the names of dimensions relevant for a given word.

Some dimensions will come from Osgood semantic differential (Osgood, Suci, Tannenbaum, 1957, Snider & Osgood, 1969, Osgood, 1971), while others will be specified in the course of the linguistic analysis of the pairs of words. The latter will generate hypotheses as to the differences in connotation between loanwords and native forms and these hypotheses can be tested in this study. Other hypotheses will come from the literature on phonesthemes and other types of sound symbolism, and from the analysis of the results of Study 1, where the frequency of association may suggest other important dimensions on which the forms differ.

Ratings will be compared using non-parametric tests. They will serve, as mentioned earlier, as a further validation of the hypotheses advanced in linguistic analyses. Interesting results will be obtained not only for the particular items (pairs) compared, but also for the whole set of loanwords compared against the whole set of native words, on the dimensions that were common for all of them. This will make it possible to detect the semantic connotations of "being foreign", if there are any. The latter analysis may have to be performed for subgroups of words (classified by their structure, or language of origin).

Both Study 1 and 2 may be modified to have participants hear or pronounce the words (rather than just reading them silently). This is in order to check if the phonological representation evoked in silent reading has similar effects as a written representation. This, obviously, will make no sense in the case of pure variants (differing only in graphic form).

Study 3: Stories and questions

Stories will be composed around possible dimensions that are discovered in the literature, in the linguistic analyses of this project, and in Studies 1 and 2 above. Two versions of each story will be created, which will be identical but for the target words. In one version of the story a loanword will be present, in the second version a native word for the same (similar) concept will be used. Questions will be then asked about the objects in the stories, and concerning the preferences and affective attitudes towards them. This is a version of the study by Song & Schwartz (2009), but with real objects and with additional information provided about them. It is also inspired by studies on gender and race discrimination, where versions of (otherwise identical) CVs are composed with names of the persons that differ in terms of gender and ethnicity (Steinpreis, Anders & Ritzke, 1999), then study participants are asked to evaluate the propensity for the given person to be hired. This should be a more ecologically valid task than judging names in isolation (taking such an approach could have rendered Song & Schwarz's 2009 study more ecologically valid).

Only certain loan–native pairs are suitable for this type of task, mainly nouns, such as *helikopter*/*śmigłowiec* 'helicopter' (e.g., judging the speed, size, comfort of traveling and whichever features will prove relevant to the distinction), *symptom*/*objaw* 'symptom' (e.g., judging graveness), but also verbs (e.g., *partycypować*/*uczestniczyć* ‘participate’). Variants are obviously also suitable for this kind of study (*kemping* vs. *camping* ‘camping site’; *jazz* vs. *dżez* ‘jazz’) indicating perhaps influences of the graphic form on images associated with meaning – i.e., a form of “graphic symbolism”.

Analyses of multidimensional representations of meaning

The third method to be employed, besides the linguistic case-based research and the empirical studies geared towards generalizations (see above), will be a very interesting and increasingly popular method of analyzing multidimensional representations of meaning based on co-occurrences of words.

Basing on co-occurrences of words in large corpora, the method enables assessment of the relative similarity of the contexts in which loanwords and native words from each pair are used (which, it is assumed, is directly related to their meaning). Again, this will be informative for:

i) each pair of words, allowing:
- the assessment of their semantic distance relative to other pairs, and
- finding the ‘clouds’ of closest semantic neighbors for each word from each pair,

ii) generalizing properties of semantic representations of loanwords as one category vs. native words as the other.

Previous research (Kruszyński & Rączaszek-Leonardi, 2006) has shown that even for small corpora (0.5 mln) and for languages with rich inflexion and thus relatively free word order (Polish), it is possible to construct a multidimensional meaning representation based on the co-occurrences of words that 1) reflects the intuitive closeness in meaning in the form of distance between the vectors representing the meaning, 2) performs semantic classifications, 3) performs syntactic classifications. The method employed then (Hyperspace Analogue to Language, Lund & Burgess, 1996, henceforth HAL) used a rather simplistic meaning representation: the meaning of a word was characterized as the mean distance of this word to other words in a corpus. All the words present in a corpus were listed, and arranged as a first row and a first column of a matrix. Vectors of distances between the words were calculated by moving a window of a specific size (experiments showed that windows of the size 8-10 gave the best results) in a large corpus of texts. The closer the two words appear in a text, the larger the value written in a cell, which is at a crossing between the appropriate row and column. One may imagine such a matrix of co-occurrences of words as akin to the matrix of distances between the cities often found in road atlases (only the values will be the inverses of distances: the closer the cities the bigger the value). The position of a particular city with respect to the others can then be inferred from a vector of distances to the other cities. Cities with similar values in all elements of the vector will be positioned close to each other.

Recently, however, methods of multidimensional meaning representation based on averaged distances to other words in corpora have been greatly improved. The best known successor to HAL is COALS, the Correlated Occurrence Analogue to Lexical Semantics (Rohde et al., submitted). In HAL the co-occurrences with words of high frequency (such as function words) biased the assessment of differences between words. The COALS method factors out the frequency. The "distances" in HAL are substituted with a conditional rate of co-occurrence of two words and not the raw rate of word-word co-occurrence. It is equal to 0 if a word *b* occurs equally often in the vicinity of a word *a* as in the vicinity of any other random word. That is, the score in the matrix at the crossing of the column for word *b* with a row for word *a* tells us if the word *b* occurs more or less often in the vicinity of word *a* than it does in general (Rohde et al., submitted).

The COALS model predicts human semantic similarity judgments better than HAL, and in most cases better than LSA or WordNet based models. Impressively, it scored above 88% on the TOEFL subtest consisting in choosing the word that is most similar in meaning to a given word out of 4 possibilities. College applicants taking the test score, on average, 65% (Rohde et al., submitted).

The present project offers an opportunity to develop a fully functional tool for building multidimensional semantic representations on the basis of various corpora that can be later used by linguists and other researchers. This will involve writing an interface to the existing, freely available procedures, that were made accessible by the Airhead Research Group as the S-Space Package (Jurgens and Stevens, 2010) (http://code.google.com/p/airhead-research/). The S-Space Package is hosted on GitHub at https://github.com/fozziethebeat/S-Space. Once the interface is in place, we will construct a COALS matrix for the National Corpus of Polish Language in order to conduct the following studies:

Study 4: Semantic closeness of pairs

Pairs of synonymous (or near-synonymous) words will be found in the multidimensional semantic representation (COALS) and the distance between them will be computed. In COALS, the semantic similarity between two words is given by the correlation of their vectors. In this way we will obtain an independent measure of semantic closeness of the words in each pair (a measure independent of linguistic intuitions, linguistic case studies, association frequencies from the Study 1, and evaluations from other studies).

Study 5: Semantic surroundings

Pairs of synonymous (or near-synonymous) words will be analyzed with the COALS model created on the basis of the largest corpus of the Polish language. Nearest neighbors of each word (borrowed, native) in each pair will be found and compared in terms of their relative distances. Sets of neighboring words will be compared with the associations generated by participants in the Study 1, as well as with the relations found through linguistic studies.

Study 6: Diachrony of loanwords

If time and resources allow, we will construct corpora from different time-periods to see the trajectories within semantic space of the loanwords (especially in relation to their native counterparts). Perhaps this could help generate hypotheses about possible different factors at work in the development of words which retained the foreign form vs. ones which received a native orthographic form. Such a study could also justify potential claims about the semantic split between once-synonymous forms.

Summary: A cross-validation of methods

As mentioned above, the project is based on a sort of 'triangulation' of the problem of comparing near-synonymous native and borrowed words using different methods. All of them may serve as sources of hypotheses, but the results coming from the application of one method can also be verified by the other methods.

Features detected in collocations, through corpus studies that will specify the differences in semantic prosody of the words, will be tested in experiments. Thus the psycholinguistic experiments will provide validation of linguistic case-analyses.

Similarly, the quantitative methods of corpora analysis that are based on multidimensional representations of meaning will help to verify the case-based, qualitative linguistic methods of collocation analyses that lead to conclusions about semantic prosody of particular words. At the same time, the quantitative methods based on multidimensional representations of meaning will allow for validation of the data on the experimentally induced meaning descriptions. The methods will also serve as a source of new hypotheses for linguistic analyses, detecting dependencies that could have been overlooked by analyses of individual, or even frequency-based characterization of usage.

The multidimensional semantic representation method (COALS) will itself come under interesting scrutiny, by being compared against linguistic analyses and experimentally obtained data. To our knowledge this would be one of the first validations of the semantic differentiation powers of the multidimensional meaning representation, which utilizes human data gathered with Osgood semantic differential. This could be a valuable topic of analyses from the standpoint of the creators of such models.

Bibliography

Abelin, Å. (1999). Phonesthemes in Swedish, in: *Proceedings of XIV International Conference of Phonetic Sciences 99,* University of California, Berkeley, 1333– 1336.

Allott, R. (1995). Sound Symbolism. In: *Language in the Würm Glaciation*, ed. by Udo L. Figge, 15-38. Bochum: Brockmeyer. Retrieved from: http://www.percepp.com/soundsmb.htm.

Allport, G.W. (1935) Phonetic Symbolism in Hungarian words. Harvard University MS Thesis.

Aramaki, E. Miura, S., Yasuda, S. Miyabe, M., Murata, M. (2012). Which is Stronger? Discriminative Learning of Sound Symbolism, CogSci (Poster Presentation).

Atoda, T., and Hoshino, K. (1995). *Giongo Gitaigo Tsukaikata Jiten* [Usage Dictionary of Sound/Manner Mimetics]. Tokyo: Sotakusha.

Brown, R.W. (1958). *Words and Things*. New York: The Free Press.

Bybee, J. L. (1985). *Morphology: A Study of the Relation Between Meaning and Form.* Amsterdam: John Benjamins.

Chandler, D. (1995). *Semiotics for Beginners*. Consulted on 01.12.2012. http://www.aber.ac.uk/media/Documents/S4B/sem0a.html.

Deacon, T. (1997). *The Symbolic Species*. Harmondsworth; Penguin.

Fodor, J.A. (1983). *The Modularity of Mind.* Cambridge, MA: MIT Press.

Forster, K.I. (1979). Levels of processing and the structure of the language processor. In W.E. Cooper & E. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, N.J.: Erlbaum.

Hockett, Ch. F. (1960). The Origin of Speech, *Scientific American, 203*, p. 89-96.

Imai, M. Kita, S., Nagumo, M., and Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition 109*, 54–65.

Jurgens, D. & Stevens, K. (2010). The S-Space Package: An Open Source Package for Word Space Models. In *System Papers of the Association of Computational Linguistics.* Retrieved from: http://aclweb.org/anthology-new/P/P10/P10-4006.pdf

Kawahara, S., Shinohara, K. (2011). Poster presentation available at: http://www.rci.rutgers.edu/~kawahara/pdf/ICC_KawaShino.pdf

Köhler, W. (1929). *Gestalt Psychology.* New York: Liveright.

Köhler, W. (1947). *Gestalt psychology, an introduction to new concepts in modern psychology* (2nd ed.). NY: Liveright.

Kovic, V., Plunkett K., and Westermann, G. (2010). The shape of words in the brain. *Cognition 114*, 19–28.

Kruszyński, B; Rączaszek-Leonardi, J. (2006). Między strukturalistyczną a psychologiczną reprezentacją znaczenia: wielowymiarowa przestrzeń semantyczna HAL. W: P. Stalmaszczyk (red.) *Metodologie językoznawstwa*. Łódź: Wydawnictwa Uniwersytetu Łódzkiego.

Lévi-Strauss, Claude (1972). *Structural Anthropology*. Harmondsworth: Penguin.

Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (1984). "Self-organizing processes and the explanation of phonological universals". In B. Butterworth, B. Comrie, and O. Dahl (eds), *Explanations for Language Universals*. New York: Mouton, 181–203.

Lowrey, Tina M. & L. J. Shrum (2007). "Phonetic Symbolism and Brand Name Preference," *Journal of Consumer Research*, 34 (3), 406-414.

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

MacDonald, M., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676-703.

Nagórko A., Łaziński M., Burkhardt H. (2004). *Dystynktywny słownik synonimów*, Kraków: Universitas.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science, 33*, 127-146.

Ohala, J. J. (1994). The Frequency Codes Underlies the Sound Symbolic Use of Voice Pitch. In L. Hinton, J. Nichols, and J. J. Ohala, eds., *Sound Symbolism,* 325-347. Cambridge: Cambridge University Press.

Ohala, J. J. (1997). Sound symbolism. *Proc. 4th Seoul International Conference on Linguistics [SICOL]* 11-15 Aug 1997. 98-103.

Osgood, C.E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Osgood, C. E. (1971). Exploration in semantic space: A personal diary. *Journal of Social Issues,* 27, 5-64.

Oudeyer, P-Y. (2006). *Self-Organization in the Evolution of Speech, Studies in the Evolution of Language*. Oxford: Oxford University Press.

Oxford Dictionaries Online, consulted ofn23.03.2013, http://oxforddictionaries.com/.

Peterfalvi, J-M. (1970) *Recherches Expérimentales sur le Symbolisme Phonétique*. Paris: C.N.R.S.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language; Evidence from spoken and signed languages. *Frontiers in Psychology, 1*, 1-17.

Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia: a window into perception, thought and language. J. Conscious. Studies 8, 3 –34.

Rączaszek-Leonardi, J., & Scott Kelso, J. A. (2008). Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. *New Ideas in Psychology*, vol. 26 (2), pp. 193-207.

Rączaszek–Leonardi, J. (2009). Symbols as constraints: the structuring role of dynamics and self-organization in natural language. *Pragmatics and Cognition,* 17:3, pp. 653-676.

Rączaszek-Leonardi, J. (2013). Language as a system of replicable constraints. In: Pattee, H.H. & Rączaszek-Leonardi, J. *Laws, Language and Life: Howard Pattee's Physics of Symbols*. Springer.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*, 364–382.

Rich, S. (1953). The Perception of Emotion. Cambridge MA: Radcliffe College thesis.

Roberts, G. (2010). Divergence through cooperation: an experimental simulation. Paper made available for the participants of LaSC, Warsaw 16-18 September 2010.

Rohde, D.L.T., Gonnerman, L., and Plaut, D.C. An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*. Available at: http://tedlab.mit.edu/~dr/Papers/RohdeGonnermanPlaut-COALS.pdf

Sapir, E. (1929). A study in phonetic symbolism, in: *Journal of Experimental Psychology* 12 225-239.

Saussure, de, F. (1916/1983). *Course in General Linguistics*. London: Duckworth.

Shinohara, K. and Kawahara, S. (2010). A cross-linguistic study of sound symbolism: The images of size. Proceedings of Berkeley Linguistic Symposium 36.

Shintel, H., Nusbaum, H. C., and Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language, 55*, 165–177.

Snider, J. G., and Osgood, C. E. (1969). *Semantic Differential Technique: A Sourcebook*. Chicago: Aldine.

Song, H. & Schwarz, N. (2009). If it's difficult to pronounce, it must be risky: Fluency, familiarity, and risk perception. *Psychological Science*, 20(2), 135–138.

Steinpreis, R.E., Anders, K.A. & Ritzke, D. (1999). The impact of gender on the review of curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles, 41*, 7/8, 509-528.

Thorndike, E. L. (1945). “On Orr’s hypothesis concerning the front and back vowels,” *British Journal of Psychiatry* 36, 10-14.

Traugott, E.C. & Dasher, R.B. (2002). *Regularity in Semantic Change*. Cambridge, UK: Cambridge University Press.

Tsuru, S. and H.S. Fries. (1933). Sound and meaning. *Journal of General Psychology.* 8: 281-284.

Ultan, R. (1978). “Size-sound symbolism”, In J. H. Greenberg, C. A. Ferguson, and E. A. Moravcsik (eds.), *Universals of human language, Vol. 2: Phonology.* Stanford, CA: Stanford University Press. 527-568.

Vigliocco, G., and Kita, S. (2006). Language specific effects of meaning, sound and syntax: implications for models of lexical retrieval in production. *Lang. Cogn. Process*. 21, 790–816.

Xiao, R. & Mcenery, T. (2006). Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27.1:103-129.

Yoshida, H. (2012). A Cross-Linguistic Study of Sound Symbolism in Children’s Verb Learning. Journal Of Cognition And Development, 13(2): 232–265.

This study was carried out within a Polish-Czech project: *Adaptation, perception and reception of verbal loans in Polish and Czech – linguistic, psychological and historic-cultural factors*, supported with a grant from the National Science Centre, Poland, no DEC-2011/03/B/HS2/02279.