Précis of *Weaving an ambiguous lexicon*

Isabelle Dautriche

Modern cognitive science of language concerns itself with (at least) two fundamental questions: *how do humans learn language?* —the learning problem —and *why do the world's languages exhibit some properties and not others?* —the typology problem. Though the relation between language acquisition and typology is not necessarily one of equivalence, there are many points of contacts between these two domains. On the one hand, children work on language through an extended period of time and their progression could plausibly reveal aspects of the cognitive blueprint for language. On the other hand, paying attention to the structural commonalities of languages can clue us into what the human learning mechanism producing these preferences must look like. These questions, although complementary, represent different approaches towards understanding the features of cognition underlying the language faculty and have often been dealt with separately by different research communities.

This dissertation attempts to link up these two questions by looking at the lexicon, the set of word-forms and their associated meanings, and ask why do lexicons look the way they do? And can the properties exhibited by the lexicon be (in part) explained by the way children learn their language? One striking observation is that the set of words in a given language is highly ambiguous. Words may have multiple senses (e.g., homophones, such as "bat" in English which means both a flying mammal and an instrument to hit baseballs) and are represented by an arrangement of a finite set of sounds that potentially increase their confusability (e.g., minimal-pairs, such as "sheep" and "ship" which differ from only one sound). Lexicons bearing this property present a problem for children learning their language who seem to have difficulty learning similar sounding words (e.g., Stager & Werker, 1997; Swingley & Aslin, 2007) and resist learning words that have multiple meanings (e.g., Casenhiser, 2005; Mazzocco, 1997).

The presence of ambiguity in languages is thus a problem for word learning. At first sight this might suggest that languages are not influenced at all by children's learning difficulties. This dissertation explores two non-mutually exclusive explanations for the presence of confusable and ambiguous words in the lexicon: 1) The presence of these words may be the consequence of other cognitive pressures not related with the acquisition of words (but related to the *use* of words) and 2) Confusable and ambiguous words that are present in the lexicon are learnable; that is, learning may exercise some more fine-grained influence on the distribution of these words in the lexicon by keeping only confusable and ambiguous words of the learnable kind. I explore these ideas by first using methods from computational linguistics to quantify the amount of word-form similarity in the lexicon. I then rely on methods from experimental psychology to explore the question of what kind of ambiguity may be learnable by children, and I finish by showing how ambiguous words challenge current word learning accounts.

Quantifying word-form similarity in the lexicons of natural languages

The second chapter of this dissertation explores whether the lexicon of natural languages is designed such that it can be reduced to properties of the cognitive system. Although the mapping between forms and meanings is arbitrary, the particular sets of form-meaning mappings chosen by any given language may be constrained by a number of competing cognitive pressures associated with language acquisition and use. For example, imagine a language that uses the word "feb" to refer to the concept CAT. If the language used the word "fep" for DOG, it would be easy to confuse with

"feb" (CAT) since the two words differ only in the final consonant and would often occur in similar contexts (i.e. when talking about pets). However, the similarity of "feb" and "fep" could make it easier for a language learner to learn that those sound sequences are both associated with animals, and the learner would not have to spend much time learning to articulate new sound sequences since "feb" and "fep" share most of their phonological structure. On the other hand, if the language used the word "sooz" for the concept DOG, it is unlikely to be confused with "feb" (CAT), but the learner might have to learn to articulate a new set of sounds and would need to remember two quite different sound sequences that refer to similar concepts.

The first part of this chapter investigated whether the structure of word-form similarity in the lexicon is the result of cognitive pressures associated with language acquisition and use. On one hand, one might expect that a well-designed lexicon should avoid confusable word-forms to satisfy communicative constraints: a *pressure for distinctiveness*. On the other hand, one might expect that a well-designed lexicon should favor word-form similarity to make the lexicon easier to produce, learn and remember, a *pressure for clumpiness*.

We proposed a new methodology to investigate whether the structure of word-form similarity in the lexicon differs from chance and in what direction. Because of phonotactics and constraints on the human articulatory system, a naive approach would quickly conclude that the lexicon is clumpy. To assess whether a pressure for distinctiveness or clumpiness drives the organization of word-form similarity in the lexicon, we must be able to compare it to some baseline that would reflect the *null hypothesis* about how language may be structured in the absence of cognitive forces. Our method follows the logic of standard statistical hypothesis testing: we created a sample of null lexicons according to a statistical baseline with no pressure for either clumpiness nor distinctiveness. We then computed several test measures (e.g., string edit distance) and assessed whether real lexicons have test measures that are statistically different from what would be expected under the null lexicons.

We studied monomorphemes of Dutch, English, German and French. To accurately capture the phonotactic processes at play in each language, we built several generative models of lexicons: ngrams over phones, ngrams over syllables, and a PCFG over syllables. After training, we evaluated each model on a held-out dataset to determine which one most accurately captured each language. The best model, a 5-phone model, was used as the statistical baseline with which real lexicons are compared. Because our baseline models capture effects of phonotactics, we were able to assess pressures for clumpiness or distinctiveness over and above phonotactic and morphological regularities. Across a variety of measures, we find that natural lexicons have the tendency to be clumpier than expected by chance. This reveals a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonetically distinct as possible.

Why might the lexicon be clumpy? A possible explanation for greater lexical clumpiness in the lexicon is the presence of form-meaning regularities. Several studies suggest that systematic form-meaning mappings may facilitate word learning (e.g., Imai & Kita, 2014; Monaghan et al., 2011). The idea is that learning similarities among referents (and hence forming semantic categories) may be facilitated if these similarities appear also at the level of the word-form. For instance, it might be easier to learn the association of *fep* and *feb* to CAT and DOG than to CAT and UMBRELLA. This advantage in learning provides an explanation for the observation of sound-symbolism in languages (see e.g., Bremner et al., 2013; Hamano, 1998) and predicts that phonologically similar words would tend to be more semantically similar.

Yet there may also be a functional disadvantage for form-meaning regularities. Another feature of semantically related words is that they are likely to occur in similar contexts. For instance, weather words like "rainy", "windy", and "sunny" are all likely to occur in the same discourse contexts. As a result, one might imagine that context makes it more difficult to distinguish between semantic similar words. If someone said, "It's ___ outside today," the missing word could plausibly have be "sunny" or "windy", but it's unlikely to be "funny" or "Cindy". Therefore, one would also predict that semantically related words should be more distant in phonological space than semantically unrelated words.

In the second part of this chapter, we looked at the interaction of word-form similarity and semantic similarity in relation with possible functional advantages. We showed that across 101 languages, similar sounding words tend also to be more semantically similar above what could be expected by chance (an extension of Monaghan et al. 2014 in English). In order to remove the contribution of morphology from this correlation, we conducted the same analysis on the set of monomorphemic lemmas of a restricted number of languages and found exactly the same pattern of results. This suggests that the pattern of clumpiness in the lexicon may be in part explained by form-meaning regularities, over and beyond morphological regularity, across a large range of typologically different languages. To date, with 101 languages in the sample, this is the largest cross-linguistic analysis that offers insight into the processes that govern language learning and use across languages.

In sum, across a large range of measures, we showed that there is more phonological similarity in the lexicon than expected by chance and that these words tend to be correlated with greater semantic similarity. This suggests that there is a pressure for the lexicon to be more clumpy. Such a pressure may be beneficial not only for speakers, as it minimizes articulatory effort and relieves memory load (as less sound sequences are used), but also for learners at it may help them to learn some aspects of their language (as word-form regularity may be helpful in segmenting words from speech, see Altvater-Mackensen & Mani 2013, and helps category formation, see Monaghan et al. 2011). However, such properties may be detrimental for some other aspects of learning as previous results suggest that children may have a hard time differentiating these forms to attribute them meanings (e.g., Casenhiser, 2005; Swingley & Aslin, 2007). One important question is thus: How do children manage to learn such a lexicon?

Learning confusable and ambiguous words

Word learning is described as a process where children *"flag "new word!" upon hearing a phonological sequence with no current lexical entry"* (Carey, 1978). Indeed, one feature of novel words is that they are often composed of unfamiliar sequences of sounds. However, a new word can be phonologically similar or even identical to a word that already exists in the child's lexicon and yet, be associated with a novel meaning. For instance, the child may already know the word "sheep" but needs to be able to identify that "ship", a minimally different word-form, is a different word despite the phonetic variability of the speech signal. Similar-sounding words present thus learners with a challenging case where they need to find the right balance between phonological tolerance, to recognize known words, and phonological sensitivity, to be able to learn these new words.

Not only must children be able to identify novel *word-forms* in the signal to consider them as candidate lexical entries, they also must be able to identify novel *meanings* even when the word-form is identical to a form they already know, as in the case of homophones. For instance the child may already know that "bat" means bat-animals and be confronted with a sentence such as "aluminum bats are easier to swing compared to wooden bats". How does the child determine that "bat" is used here to refer to a baseball-bat and not an animal-bat? Homophony thus presents learners with a unique word learning situation where they cannot rely on the signal alone to determine whether a phonological form is a candidate for a novel entry in the lexicon as a new word.

The third chapter of this dissertation investigates whether 18- to 20-month-old toddlers take into account other factors than phonology when determining what counts as a new word. To answer this question, we used methods from psycholinguistics on toddlers (preferential looking paradigm, Golinkoff et al. 1987) and statistical analysis techniques initially developed in neurosciences (cluster-based permutation analysis, Maris & Oostenveld 2007). We showed that French toddlers had no problem learning object labels that were phonological neighbors of a familiar verb (e.g., learning "kiv", a neighbor of "give") but did find it difficult to map neighbors of a familiar noun onto a novel object (e.g., learning "tog", a neighbor of "dog", a replication of Swingley & Aslin 2007). This suggests that toddlers are not confused by phonological similarity per se when learning words. In fact, even in cases where the novel word is phonologically identical to a word in toddlers' lexicons (i.e., a homophone), we showed that toddlers correctly learnt

the novel meaning, provided that the two homophones are sufficiently distant syntactically (e.g. “an eat” is a good label for a novel animal) or semantically (e.g. “a glass” for a novel animal). When the homophones were close on both dimensions (e.g. “a cat” for a novel animal), however, no learning was observed.

The experimental results obtained in this chapter suggest that the learning system of young children is equipped with constraints and mechanisms that allow them to successfully learn homophones and similar-sounding words as long as these words can be distinguished in a context that children can capitalize on (syntactic and/or semantic). For instance, the probability that the novel noun “a kiv” will be considered as a variant of its neighbor verb “give” (an action) is low because the novel word appears in a noun context and labels an object, and in this case toddlers are not overwhelmed by the phonological resemblance. Thus, children can deal with ambiguity as long as distinctiveness along other dimensions that are relevant for them is maximized. One important question is whether the structure of the lexicon reflects these constrains on learning: Is it the case that members of a homophone pair, or a minimal-pair, are more distant from one another than would be expected by chance alone?

To explore that question, we evaluated whether similar-sounding words and homophones that exhibit properties that make them learnable by children are more represented in the lexicon of natural languages than similar-sounding words and homophones that are harder to learn. In other words: Are there more minimal-pairs and homophones from different syntactic categories in lexicons? And are members of a minimal-pair or a homophone pair more likely to be semantically distant in languages? For this, we extracted the minimal-pairs and the homophone pairs of 4 languages (Dutch, English, French and German) and looked at their distributions in both the syntactic and semantic dimensions. Interestingly, our results show that despite being more learnable, there is no pressure for minimal-pairs to appear across distinct syntactic categories in the lexicon and to be semantically distinct (as we uncovered in the second chapter). In contrast, such a pressure exists for homophones, that is, there are more across-categories homophones than expected by chance (where the chance level is simulated by randomly shuffling the syntactic categories within words of the same length in each lexicon) and homophones are as semantically dissimilar as any random pair of words in the lexicon (where semantic similarity is computed using Latent Semantic Analysis, a method developed in computational linguistics ,Landauer & Dumais 1997, on Wikipedia for each language).

The present results show that there are some correspondences between what makes homophones easy to learn and how they are organized in the lexicon. Yet, how can we explain that minimal-pairs and homophones are distributed differently in the lexicon? We suggest that these differences may reflect different tradeoffs of functional pressures associated with language acquisition and language use. As we pointed earlier, greater than chance word-form similarity would be advantageous for speech production, memory and learning categories (i.e., syntactic and semantic classes). A pressure for clumpiness would thus prevail over a pressure for distinctiveness for minimal-pairs. However, the requirements to learn minimal-pairs and homophones are different. While minimal-pairs can be distinguished at the word-form level, homophones cannot: The only way for homophones to be recognized as such is to have sufficiently distinct meanings. Interestingly, Bloomfield (1962) reports that in a dialect of Southwestern France, when the Latin forms ”gallus” *rooster* and ”cattus” *cat* were in danger of merging into one form, ”gat”, another novel word acquired the meaning *rooster*, suggesting that the use of the same label for *cat* and *rooster* was unwanted and caused speakers to remap a new form onto one of these meanings. This illustrates that pairs of homophones that belong to the same semantic field tend to be eliminated during the course of language evolution. Contrary to minimal-pairs, homophones may thus show an advantage of distinctiveness over clumpiness as, because of form-identity, there is no choice but for these words to be distinctive in meaning in order to be learnable and transmitted with accuracy.

What homophones say about words

Learning the meaning of a word is not an easy task, though children make it appears this way. Certainly, after a few presentations children are able to correctly identify the appropriate referent of a novel word they just have been taught.

Yet, it is unclear what exactly they have "learned" about the word and which lexical representation they have formed. Suppose that a speaker uses "banana" to refer to the fruit the child is eating, what can the child infer about the word "banana"? Certainly "banana" could refer to the set of all bananas and only bananas but many other meanings are consistent with that one experience: the set of all fruits, the set of all yellow objects and so on (Quine, 1960).

Existing theories of word learning have stressed the importance of prior knowledge to constrain the hypothesis space of possible word meanings (e.g., Bloom, 2001; Markman, 1989). One way in which learners may reduce their hypothesis space is by assuming that those concepts which have word-forms associated with them have to be *convex*, that is, the instantiations of these concepts are contiguous in conceptual space (Gärdenfors, 2004; Murphy & Medin, 1985). For instance, a concept such as CAR OR WATER, is not a proper candidate for a concept because its members would be drawn from two disjoint sets which do not form a convex group of entities in conceptual space. Crucially, one piece of all current word learning models (associative learning accounts, e.g., Regier 2005; hypothesis elimination accounts, e.g., Siskind 1996; Bayesian accounts, e.g., Xu & Tenenbaum 2007) that make them succeed is that they translate such a convexity constraint over *concept* to a convexity constraint over *word-forms*: That is, if A and B can be labeled using the sound /kaet/, then all objects falling in between A and B in conceptual space can also be labeled with the word /kaet/. Such a constraint has specifically addressed the problem of learning unambiguous words where a single form is used to refer to a single meaning; yet, it also bans homophony from the system entirely.

The fourth chapter of this dissertation provides the first careful look at what homophones have to say about word learning, from a theoretical and an experimental standpoint both with adults and with 5-year-old children. Our point of departure was the work of (Xu & Tenenbaum, 2007) which implements the convexity constraint over word-forms in a predictive model. In this study, the conceptual space is defined over a tree-structured representation of entities by clustering a set of entities based on their similarity. Subtrees correspond to categories that words could label at different levels of granularity (e.g., cat, feline, mammal, animal). When exposed to a set of learning exemplars uniformly sampled from a category, adults extend the label to the minimal subtree including all the exemplars. For example, when presented with three "feps" labeling three Dalmatians, adults readily extend "fep" to the set of all Dalmatians, would they be presented with a Dalmatian, a Labrador and a German-shepherd they would extend the label to the set of all dogs. In other words, participants pick the smallest generalization that satisfies the convexity constraint on word-forms.

To learn a homophone, such as "bat", learners will observe several exemplars of animal-bats and several exemplars of baseball-bats. In such a case, if learners hypothesize that a word applies to a single, convex concept, as most words do, they would never discover homophony. Instead, they could consistently postulate that "bat" refers to some superordinate, coherent category encompassing both animal-bats and baseball-bats, just like a word like "thing" does. However, if "bat" was linked to such a broad category, it is likely that learners would have observed many things that are called "bat" but are neither animal-bats nor baseball-bats (a *uniform* distribution of exemplars drawn from the superordinate category of "things") rather than having observed only animal-bats and baseball-bats (a *bimodal* distribution of exemplars within the superordinate category).

We showed that both adults and children capitalize on the *sampling distribution of the learning exemplars* to postulate homophony. That is when presented with exemplars clustered at two distant positions in conceptual space, such as {two primates, two mushrooms}, adults and children did not extend the label to all objects falling within a convex category encompassing all exemplars (i.e. LIVING BEINGS), rather, they preferred to restrict the label to members of two disjoint subcategories (i.e. PRIMATE and MUSHROOM). Importantly, we found evidence that these meanings were stored *separately*, suggesting that adults and children's representation of the novel word in these conditions is very much similar to homophony (Srinivasan & Snedeker, 2011; Zwicky & Sadock, 1975). Altogether, our results suggest that adults and children by the age of 5 use information about the sampling distribution of learning exemplars to discover whether a novel word is associated with one or several meanings.

Our results suggest that children expect meanings to be convex, and are willing to postulate homophony rather

than breaking this constraint (postulating a disjoint meaning) or than enforcing that convexity constraint at all cost (postulating a broad lexical entry for problematic words). The present work thus has important implications for current word learning accounts. When a label seems to apply to a disjoint set of objects, the learner has two options: 1) postulate homophony or 2) postulate that the label is a single word that applies to a larger set of objects (the category that covers all the positive instances of the label). Most accounts predict that 2) is the default, but this has to be refined since children (and adults) eventually learn homophones. An important open issue then is to equip the learning system with the right built-in constraints. At this point it seems that word learning is guided by a) learners' expectations that concepts are convex, always; b) learners' expectation that word-forms are linked to one meaning, in general; and c) the possibility that a word-form maps onto several distinct meanings if a) is challenged. The present study contributes to point c) above: children can entertain the possibility that a word maps onto several distinct meanings to accommodate apparent violations of a) concept convexity at the detriment of b) a one-to-one mapping between forms and meanings.

Conclusion

This work investigated why lexicons are ambiguous. A prominent feature of this dissertation has been the combined use of lexical models to quantify the amount of ambiguity in the lexicon and experimental methods in toddlers and adults to investigate what exactly enables children to learn ambiguous and confusable words. This research suggests that ambiguous and confusable words, while present in the language, may be restricted in their distribution in the lexicon and that these restrictions reflect (in part) the existence of several other constraints on the lexicon (chapter 2) and (in part) how children learn languages (influence of their developing parsing system, chapter 3, and existence of built-in constraints on the learning system allowing for the existence of homophones, chapter 4). For a long time, the presence of ambiguity has challenged the view that language is designed to fit our learning and processing needs. Yet, the present data together with previous findings (e.g., Piantadosi et al., 2012; Wasow et al., 2005) suggest that the existence of ambiguity in languages may find an explanation when one is considering the competing needs of learners, listeners and speakers of the language.

References

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, n/a–n/a.

Bloom, P. (2001). Précis of How children learn the meanings of words. *Behavioral and Brain Sciences*, *24*(06), 1095–1103.

Bloomfield, L. (1962). Language. 1933. *Holt, New York*.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "bouba" and "kiki" in namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, *126*(2), 165–172.

Carey, S. (1978). The child as word learner.

Casenhiser, D. M. (2005). Children's resistance to homonymy: an experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319–343.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, *14*(01), 23–45.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Hamano, S. (1998). *The sound-symbolic system of japanese.* ERIC.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298.

Kim, K.-O. (1977). Sound symbolism in korean. *Journal of Linguistics*, *13*(01), 67–75.

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.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Mit Press.

Mazzocco, M. M. (1997). Children's interpretations of homonyms: a developmental study. *Journal of Child Language*, *24*(02), 441–467.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299–20130299.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive science*, *29*(6), 819–865.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, *62*(4), 245–272.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive psychology*, *54*(2), 99.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the Web of grammar: Essays in memory of Steven G. Lapointe. CSLI Publications*.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Zwicky, A., & Sadock, J. (1975). Ambiguity tests and how to fail them. *Syntax and semantics*, *4*(1), 1–36.