Semantic growth of morphological families in English

Regina Henry and Victor Kuperman

Department of Linguistics and Languages, McMaster University, Canada

This paper explores the question of when and how morphological families are formed in one’s mental lexicon, by analyzing age-of-acquisition norms to morphological families (e.g., booking, bookshelf, check book) and their shared morphemes (book). We demonstrate that the speed of growth and the size of the family depend on how early the shared morpheme is acquired and how many connections the family has at the time a new concept is incorporated in the family. These findings dovetail perfectly with the Semantic Growth model of connectivity in semantic networks by Steyvers and Tenenbaum (2005). We discuss implications of our findings for theories of vocabulary acquisition.

Keywords: morphological family, age-of-acquisition, semantic network, vocabulary acquisition

INTRODUCTION

Research of the last decades has uncovered a plethora of evidence that the visual recognition of printed words is influenced by (unseen) lexical items morphologically related to those words. Earliest inquiries into effects of morphological connectivity reported – across multiple languages – shorter recognition latencies to words that belonged to larger morphological families and thus were connected to a larger number of orthographically and semantically related words (for visual tasks see e.g., Taft, 1979; Kostić, 1995; Schreuder & Baayen, 1997; for auditory comprehension see e.g., Balling & Baayen, 2008). Further cross-linguistic research has replicated these findings for an ever-increasing number of morphological phenomena. These include inflectional case paradigms and inflectional nominal classes, as well as morphological families that share either a free (book: booking, bookstore, bookshelf, check book) or a bound morpheme (un-: unlucky, unhappy, unnerving, etc.) and are made up of derived words, compound words, or the combination of the two (e.g., De Jong et al., 2000, 2002; Milin et al., 2009a, 2009b; Moscoso del Prado Martín et al.,

Corresponding author: henryr@mcmaster.ca
Modulation of the acoustic signal during speech production has also been shown to co-vary with the size of the family of the articulated word (Ernestus & Baayen, 2003; Kemps et al., 2005; Kuperman et al., 2007; Pluymaekers et al., 2010).

This expansion in the body of empirical evidence has gone hand in hand with the increasing repertoire of lexical-statistical measures of morphological connectivity. Most commonly such measures are based on probability distributions in inflectional, derivational, compound-based or mixed morphological families: e.g., the number of word types in the family (family size) or the summed token frequency of family members (family frequency). Other information-theoretic measures have been proposed to quantify either the amount of information in a paradigm (information content or surprisal: Balling & Baayen, 2012; entropy: see e.g., Moscoso del Prado Martín et al., 2004a; Baayen, Wurm, & Aycock, 2007), the distance between probability distributions of different morphological paradigms (relative entropy and cross-entropy: see e.g., Balling & Baayen, 2012; Kuperman, Bertram, & Baayen, 2010; Milin et al., 2009a, 2009b), secondary family size (Baayen, 2010), properties of directed graphs representing a compound space (Baayen, 2010), or other related metrics (cf. Kostić et al., 2003). Across languages, these measures show reliable effects on speed and accuracy of responses to chronometric tasks requiring printed and spoken word recognition. Neural activity and eye-movements registered during printed word recognition, and the acoustic signal registered during speech production (cf. among others, Baayen, Feldman, & Schreuder, 2006; Bien, Levelt, & Baayen, 2005; Kostić et al., 2003; Kuperman, Bertram, & Baayen, 2010; Milin et al., 2009a, 2009b; Pylkkänen et al., 2004) have also demonstrated reliable effects of these measures. To sum up, the research field can be argued to currently possess detailed knowledge of both the effects of morphological connectivity on language processing and the characteristics of morphological families that appear to underlie these effects (Baayen, 2010).

A much less worked-on aspect of this inquiry is the developmental one. It regards the question of when and how morphological families – as well as connections between family members and between families – are formed in one’s mental lexicon. Behavioral data suggest that children are sensitive to morphological connectivity early: e.g., Krott and Nicoladis (2005) and Nicoladis and Krott (2007) report facilitatory effects of family size on performance in the compound explanation task in children aged 3 to 5. More direct tests of word recognition also fall in line with this finding. A facilitatory effect of family size on lexical decision latencies was found in Dutch second-graders (Perdijk, Schreuder, & Verhoeven, 2005; Perdijk, Schreuder, Verhoeven, & Baayen, 2011) and a similar effect on word reading accuracy was observed in 10 and 12 year old English speakers (Carlisle & Katz, 2006). Yet these findings only testify to the awareness of morphological connectivity at early age (Bloom, 2000), they do not shed light on how or when this awareness is acquired.
In answering the “how” part of this question, the present paper adopts the model of Semantic Growth proposed by Steyvers and Tenenbaum (2005). In this model words are represented as nodes, and connections between them indicate semantic relationships. The semantic nodes representing words acquired early, serve as hubs in the semantic network such that newly learned words and concepts preferentially attach to these hubs rather than form new hubs of their own. As a result, words tend to have more connections the earlier they are acquired. The mechanism enabling this preferential attachment is the assumption that the probability of attracting another connection is proportional to the number of existing connections to the node. The process responsible for the growth of semantic networks is akin to “semantic differentiation, in which new concepts that correspond to more specific variations on existing concepts and highly complex concepts (those with many connections) are more likely to be differentiated than simpler ones” (Steyvers & Tenenbaum, 2005). The way semantic nodes are connected in the network is predicted to influence the variables characterizing the learning trajectory, namely, the word’s age of acquisition (AoA) and its frequency of occurrence. For a detailed exposition of the model, readers are referred to Steyvers and Tenenbaum (2005) and Griffith, Steyvers, and Tenenbaum (2007). Computational simulations revealed that the model of Semantic Growth can account for the patterns of semantic organization attested in both natural and artificial semantic networks, including the World Wide Web, internet-based communities, thesauri, free association norms, ontologies, and others (Griffith, et al., 2007; Steyvers & Tenenbaum, 2005). The model also showed the predicted correlation between an earlier acquisition of a hub, estimated as its lower AoA rating, and the greater number of connections it attracts in such semantic networks as Wordnet, Word Association Network and Roget’s Thesaurus (Steyvers & Tenenbaum, 2005). Yet to the best of our knowledge, predictions of the model of Semantic Growth have not been tested against morphological families. The model predicts family growth that follows a scale-free organization, in that a few early acquired nodes have multiple connections, whereas the majority of nodes have few connections.

It is plausible that the model of Semantic Growth is meaningfully applicable to such small scale networks as morphological families. The effects of family size and frequency have been repeatedly argued to be of semantic nature and to arise due to the semantic resonance between members of morphological families (De Jong et al., 2000; 2002). For instance, morphological family size and family-wide entropy clustered with semantic variables as predictors of lexical decision latency (Baayen et al., 2006). Likewise, removal of family members whose meanings were not related to the meaning shared by the majority of family members (e.g., removing “hogwash” from the family of “car wash”, “mouth wash”, “body wash”) strengthened the effect of family size and related measures on lexical decision latencies in Moscoso del Prado Martín et al. (2004b). The remainder of this paper tests the
predictions of the Semantic Growth model against AoA ratings for derivational and compounding morphological families in English.

This model makes explicit predictions for the age-of-acquisition of the word/concept that serves as the hub for a family, and the speed of acquisition of subsequently acquired family members. First, the distribution of family sizes in a language will follow a scale-free pattern resulting in a Zipfian distribution, with a small number of large families and a large number of small families and singleton words. Second, words acquired earlier will have larger families due to preferential attachment, yielding a correlation between the family size and the AoA of the target word that is shared by the family (see a related finding in Baayen et al., 2006). Steyvers and Tenenbaum’s (2005) model also predicts that larger families will attract more connections, and they will attract them earlier. With the same amount of exposure to words A and B, it is more likely that word A will be learned and retained in the mental lexicon if it can be attached to a richer semantic neighborhood than a semantic singleton B. Thus we expect words belonging to a larger family to show, on average, earlier AoA ratings.

We test the set of predictions above by analyzing correlations between distributional characteristics of morphological families and age-of-acquisition ratings to monomorphemic target words and morphological families that share target words. The current analysis uses adult age-of-acquisition (AoA) ratings obtained in a norming study (Kuperman, Stadthagen-Gonzales, & Brysbaert, 2013), where participants were instructed to enter the age at which they thought they encountered the word for the first time. Although these ratings are subjective, there is evidence of strong correlations between adult ratings and both children’s ratings and objective measures of AoA in English, French and Dutch (see an overview in Juhasz, 2005). Using multiple regression analysis, Gilhooly and Gilhooly (1980) demonstrated that adult ratings of AoA were the only significant predictor of objective AoA data. Additionally, they demonstrated a strong correlation between AoA ratings and the age-standardized Crichton/Mill Hill vocabulary norms (r = 0.93; Gilhooly & Gilhooly, 1980).

Morrison et al. (1997) found a strong correlation (r = .76) between adult AoA ratings of English words and objective AoA ratings obtained from 220 children, aged 2 years 6 months to 10 years 11 months. The objective AoA rating was based on 75% of the children in a specific age group correctly naming an object. Charlard et al. (2003) found that the objective AoA ratings of French words obtained from children aged from 30 to 131 months, strongly

---

1 Another prediction that we do not pursue here is that larger families will tend to show a greater semantic differentiation. That is, derived and compound words that are semantically related to their shared constituent (book) will offer narrower meanings than the one that the target concept has, e.g., check book, notebook, workbook, yearbook, scrapbook, sketchbook are specific types of “book”.
correlated with the adult ratings of the same words ($r = .90$). De Moor, Ghyselinck and Brysbaert (2000) found that Dutch words rated by adults as having been acquired early were known to 6 year olds but words with late AoA ratings were unknown to the 6 year old participants and also to many of the 12 year old participants. Additional evidence that adult ratings are valid indices of AoA come from correlations between AoA ratings and children’s word frequency counts ($r = 0.85$; Carroll & White, 1973). While objective AoA estimators for morphologically complex words are a desideratum, available corpora of child or child-directed speech or age-locked corpora of written or read-to-children texts are too small for the present purposes. Given prior research, we conclude – along with Juhasz (2005) – that subjective AoA ratings such as ones used in this study are an adequate approximation to actual ages at which words are learned.

**Method**

**Stimuli.** The data pool for the present study was a list of 30,000 words for which both AoA ratings and frequency counts were available. The mean AoA ratings were obtained from a recent norming study of Kuperman, Stadthagen-Gonzales and Brysbaert (2013), while frequency counts were obtained from SUBTLEX, a 51 million-token corpus of subtitles to US films and media (Brysbaert & New, 2009). All target words were nouns productively used as constituents in families of complex words. Each selected noun (e.g. *body*) was shared as the initial constituent in a complex word family (e.g., *bodyguard*) and as the final (right) constituent by another family of complex words (e.g., *busybody*). These positional families were formed by both compound and derived words with two or more morphemes in total: no inflected forms (*bodyguards*) were considered. Cases of homonymy (cf. *bank* in *bankbook* and *sandbank*) were removed from the list of families. For practical reasons, we did not consider the degree of semantic relatedness either between family members or between family members and the shared constituent (see General Discussion).

The resulting pool contained 42 target nouns (Table 1, column 1), each associated with a left and a right constituent family. The nouns represented a broad range of AoA ratings (Table 1, column 2) – from age 2.37 (*water*) to 12.4 (*electron*) – and of lexical frequencies (Table 1, column 3) – from 0.1 per million (*electron*) to 2 per million (*talk*). An average left constituent family comprised 10.1 words (range = 1–45, $sd = 8.8$; Table 1, column 4), while an average right constituent family comprised 14.7 members (range = 1–65, $sd = 17.0$; Table 1, column 5). Considering the position of the embedded word as either a left or right family constituent was motivated by two factors. First, because English is a right-headed language, the meaning of the right constituent, i.e. the compound’s head, is on average closer to the meaning of the entire compound. This may imply a stronger correlation between AoA ratings of right family constituent members and the AoA of the target word, as compared to left constituent family members, which are typically compounds’ modifiers. Second, Libben et al. (2003) demonstrated that the transparency of an English compound’s constituent differentially affects word recognition accuracy and response times based on its position as either a left or right constituent: for effects of headedness in English and Dutch see also Inhoff, Starr, Solomon, and Placke (2008) and De Jong et al. (2000; 2002).
**Table 1.** Distributional characteristics of target words and their families. 1: target word; 2: mean AoA ratings of target words; 3: log frequency of occurrence of target words; 4 and 5: left/right family size; 6 and 7: percentage of left/right constituent family members with negative residuals (i.e. percentage of family members that had earlier AoA ratings then would be predicted from word frequency); 8: residual of the target word (i.e. difference between target word AoA rating and the AoA rating predicted from word frequency); 9 and 10: sums of residuals of left/right constituent family.

<table>
<thead>
<tr>
<th>Target Word</th>
<th>Word AOA</th>
<th>Frequency</th>
<th>Left FamSize</th>
<th>Right FamSize</th>
<th>Neg Left</th>
<th>Neg Right</th>
<th>Residual -target</th>
<th>Left SumResid</th>
<th>Right SumResid</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>3.16</td>
<td>4.42</td>
<td>28</td>
<td>65</td>
<td>0.82</td>
<td>0.68</td>
<td>-0.67</td>
<td>-61.66</td>
<td>-60.47</td>
</tr>
<tr>
<td>water</td>
<td>2.37</td>
<td>4.06</td>
<td>23</td>
<td>19</td>
<td>0.83</td>
<td>0.79</td>
<td>-2.35</td>
<td>-52</td>
<td>-37.87</td>
</tr>
<tr>
<td>light</td>
<td>4.05</td>
<td>3.93</td>
<td>13</td>
<td>33</td>
<td>0.85</td>
<td>0.73</td>
<td>-1</td>
<td>-27.5</td>
<td>-63.55</td>
</tr>
<tr>
<td>work</td>
<td>5.86</td>
<td>4.61</td>
<td>27</td>
<td>49</td>
<td>0.81</td>
<td>0.73</td>
<td>2.5</td>
<td>-44.69</td>
<td>-66.32</td>
</tr>
<tr>
<td>life</td>
<td>5.89</td>
<td>4.61</td>
<td>16</td>
<td>6</td>
<td>0.69</td>
<td>0.5</td>
<td>2.53</td>
<td>-17.93</td>
<td>-4.23</td>
</tr>
<tr>
<td>blood</td>
<td>4.89</td>
<td>3.98</td>
<td>17</td>
<td>4</td>
<td>0.53</td>
<td>0.75</td>
<td>-0.04</td>
<td>-6.75</td>
<td>-1.4</td>
</tr>
<tr>
<td>birth</td>
<td>5.21</td>
<td>3.15</td>
<td>7</td>
<td>3</td>
<td>0.86</td>
<td>0.67</td>
<td>-1.77</td>
<td>-12.52</td>
<td>-12.52</td>
</tr>
<tr>
<td>shoe</td>
<td>2.6</td>
<td>3.19</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-4.27</td>
<td>-24.35</td>
<td>-10.05</td>
</tr>
<tr>
<td>play</td>
<td>4.1</td>
<td>4.26</td>
<td>23</td>
<td>16</td>
<td>0.7</td>
<td>0.62</td>
<td>-0.13</td>
<td>-51.97</td>
<td>-5.4</td>
</tr>
<tr>
<td>game</td>
<td>4.26</td>
<td>4.08</td>
<td>5</td>
<td>3</td>
<td>0.6</td>
<td>0.67</td>
<td>-0.42</td>
<td>-0.17</td>
<td>-5.28</td>
</tr>
<tr>
<td>sex</td>
<td>9.79</td>
<td>3.89</td>
<td>14</td>
<td>3</td>
<td>0.07</td>
<td>0.33</td>
<td>4.65</td>
<td>29.49</td>
<td>3.53</td>
</tr>
<tr>
<td>boat</td>
<td>3.84</td>
<td>3.69</td>
<td>7</td>
<td>19</td>
<td>0.86</td>
<td>0.89</td>
<td>-1.8</td>
<td>-10.65</td>
<td>-48.44</td>
</tr>
<tr>
<td>piece</td>
<td>4.89</td>
<td>3.8</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0.69</td>
<td>-0.47</td>
<td>-0.1</td>
</tr>
<tr>
<td>body</td>
<td>4.28</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>0.6</td>
<td>0.8</td>
<td>-0.59</td>
<td>-3.15</td>
<td>-2.89</td>
</tr>
<tr>
<td>woman</td>
<td>4.95</td>
<td>4.35</td>
<td>7</td>
<td>33</td>
<td>0.71</td>
<td>0.85</td>
<td>0.94</td>
<td>-4.83</td>
<td>-67.63</td>
</tr>
<tr>
<td>head</td>
<td>3.42</td>
<td>4.28</td>
<td>45</td>
<td>53</td>
<td>0.71</td>
<td>0.57</td>
<td>-0.76</td>
<td>-58.88</td>
<td>-25.88</td>
</tr>
<tr>
<td>fish</td>
<td>4.05</td>
<td>3.63</td>
<td>15</td>
<td>40</td>
<td>0.87</td>
<td>0.68</td>
<td>-1.74</td>
<td>-46.59</td>
<td>-54.47</td>
</tr>
<tr>
<td>value</td>
<td>6.78</td>
<td>3.04</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>-0.47</td>
<td>-2.21</td>
<td>-0.62</td>
</tr>
<tr>
<td>baby</td>
<td>3.84</td>
<td>4.41</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-17.73</td>
<td>-6.95</td>
</tr>
<tr>
<td>face</td>
<td>3.75</td>
<td>4.17</td>
<td>8</td>
<td>12</td>
<td>0.5</td>
<td>0.5</td>
<td>-0.7</td>
<td>-0.76</td>
<td>3.07</td>
</tr>
<tr>
<td>talk</td>
<td>3.68</td>
<td>4.64</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>0.83</td>
<td>0.39</td>
<td>-14.5</td>
<td>-17.9</td>
</tr>
<tr>
<td>color</td>
<td>4</td>
<td>3.3</td>
<td>11</td>
<td>2</td>
<td>0.73</td>
<td>1</td>
<td>-2.59</td>
<td>-23.39</td>
<td>-5.33</td>
</tr>
<tr>
<td>point</td>
<td>4.55</td>
<td>4.08</td>
<td>5</td>
<td>15</td>
<td>0.8</td>
<td>0.67</td>
<td>-0.12</td>
<td>-6.1</td>
<td>-6.62</td>
</tr>
<tr>
<td>room</td>
<td>4.22</td>
<td>4.35</td>
<td>6</td>
<td>44</td>
<td>0.67</td>
<td>0.7</td>
<td>0.22</td>
<td>-7.77</td>
<td>-68.97</td>
</tr>
<tr>
<td>master</td>
<td>6.5</td>
<td>3.65</td>
<td>8</td>
<td>19</td>
<td>0.38</td>
<td>0.58</td>
<td>0.76</td>
<td>0.21</td>
<td>0.68</td>
</tr>
<tr>
<td>leader</td>
<td>6.9</td>
<td>3.2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
<td>-0.96</td>
<td>-7.11</td>
</tr>
<tr>
<td>material</td>
<td>8.06</td>
<td>3.05</td>
<td>6</td>
<td>1</td>
<td>0.33</td>
<td>0</td>
<td>0.85</td>
<td>3.34</td>
<td>1.24</td>
</tr>
<tr>
<td>dream</td>
<td>4.88</td>
<td>3.83</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-0.41</td>
<td>-18.31</td>
<td>-2.77</td>
</tr>
<tr>
<td>electron</td>
<td>12.4</td>
<td>1.57</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.63</td>
<td>-1.42</td>
<td>1.48</td>
</tr>
<tr>
<td>brain</td>
<td>5.76</td>
<td>3.59</td>
<td>12</td>
<td>7</td>
<td>0.75</td>
<td>0.57</td>
<td>-0.11</td>
<td>-12.47</td>
<td>-5.84</td>
</tr>
<tr>
<td>neck</td>
<td>3</td>
<td>3.48</td>
<td>6</td>
<td>9</td>
<td>1</td>
<td>0.44</td>
<td>-3.15</td>
<td>-15.65</td>
<td>-4.37</td>
</tr>
<tr>
<td>board</td>
<td>6.37</td>
<td>3.51</td>
<td>5</td>
<td>51</td>
<td>0.6</td>
<td>0.78</td>
<td>0.3</td>
<td>-2.68</td>
<td>-81.08</td>
</tr>
<tr>
<td>sell</td>
<td>7.05</td>
<td>3.67</td>
<td>3</td>
<td>4</td>
<td>0.67</td>
<td>0.5</td>
<td>1.37</td>
<td>-5.31</td>
<td>-2.21</td>
</tr>
<tr>
<td>sale</td>
<td>6.76</td>
<td>3.11</td>
<td>7</td>
<td>2</td>
<td>0.71</td>
<td>0.5</td>
<td>-0.31</td>
<td>-16.37</td>
<td>-1.43</td>
</tr>
<tr>
<td>sail</td>
<td>6.47</td>
<td>2.85</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>0.38</td>
<td>-1.26</td>
<td>-17.21</td>
<td>10.1</td>
</tr>
<tr>
<td>angel</td>
<td>4</td>
<td>3.6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1.86</td>
<td>-9.39</td>
<td>0.3</td>
</tr>
<tr>
<td>bird</td>
<td>3.52</td>
<td>3.37</td>
<td>9</td>
<td>17</td>
<td>1</td>
<td>0.88</td>
<td>-2.92</td>
<td>-31.63</td>
<td>-41.17</td>
</tr>
<tr>
<td>song</td>
<td>4.26</td>
<td>3.68</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-1.4</td>
<td>-16.5</td>
<td>-12.26</td>
</tr>
<tr>
<td>place</td>
<td>4.95</td>
<td>4.49</td>
<td>7</td>
<td>10</td>
<td>0.57</td>
<td>0.9</td>
<td>1.29</td>
<td>-1.66</td>
<td>-14.81</td>
</tr>
<tr>
<td>town</td>
<td>5.11</td>
<td>4.1</td>
<td>7</td>
<td>4</td>
<td>0.57</td>
<td>0.5</td>
<td>0.49</td>
<td>-8.91</td>
<td>-0.01</td>
</tr>
<tr>
<td>bath</td>
<td>3.23</td>
<td>3.2</td>
<td>10</td>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>-3.62</td>
<td>-34.23</td>
<td>-12.22</td>
</tr>
<tr>
<td>water</td>
<td>2.37</td>
<td>4.06</td>
<td>23</td>
<td>19</td>
<td>0.83</td>
<td>0.79</td>
<td>-2.35</td>
<td>-52</td>
<td>-37.87</td>
</tr>
</tbody>
</table>
Variables. Our investigation of the link between family size and the AoA of target words and family members needs to account for a set of confounds. Higher-frequency words are acquired earlier (see review in Juhasz, 2005), and words belonging to larger families tend to occur more frequently in language (Schreuder & Baayen, 1997). Thus earlier AoAs to words from large morphological families may be a reflection of their relatively high frequency of occurrence rather than of increased morphological connectivity in the family: as our interest is in the latter, frequency needs to partialed out. Furthermore, there is a debate as to whether AoA explains unique variance over and above word frequency (Zevin & Seidenberg, 2002; Brysbaert & Ghyselink, 2006). To disentangle the collinearity between AoA, family size and frequency of the word, we base our analyses on the differences (residuals) between the observed AoA for a word and the AoA expected for that word based on its frequency of occurrence. If the observed and expected AoA ratings are identical, the residual AoA is zero, and the word is judged as learned at the same time as other words of the same frequency (morphologically complex or not) are expected to be learned. That is, the residual AoA of zero implies that there is no impact of morphological connectivity on the word’s age-of-acquisition over and above the well-established influence of word frequency. A negative residual AoA suggests that the word was learned earlier than expected based on its frequency, and a positive residual implies the opposite. Our consideration of residual rather than raw AoA values also facilitates the comparison between target words and families that differ widely in their frequencies and thus in the amount of exposure that an average individual supposedly has to those words.

Results and Discussion

Self-organization of semantic networks gives rise to a Zipfian distribution in the numbers of connections to the network nodes (Steyvers & Tenenbaum, 2005). We tested this prediction by considering the distribution of sizes in English morphological families using the morphological parsing of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). We identified about 9400 unique left constituent families and 2800 unique right constituent families that included both compound and derived words. Family sizes of both left and right families showed power-law distributions, such that there were a small number of large families and a large number of small families and singleton words. Figure 1 plots ranks of family sizes against their frequencies on the log-log (base e) scale: as expected, families that had the maximum number of members and ranked the highest had the lowest frequency of occurrence. For both positional distributions of family sizes, power-law functions provided an excellent fit ($R^2 = 0.95$ and 0.91 respectively) with the exponents equal to −0.41 and −0.55. We conclude that one prediction of the Semantic Growth model holds for morphological families: the distributions of family sizes (i.e. number of connections to the family-forming target words) are Zipfian as predicted by the scale-free organization of semantic networks. The quantile analysis of the family size distributions additionally reveals that the families we selected for consideration were large indeed (the average of 10.1 and 14.7 members in the left and right families respectively), and were in the top 10% of both distributions.
Frequencies of occurrence and AoA ratings to words were strongly correlated in the 30,000-type data set of Kuperman et al. (2013): more frequent words tended to be learned earlier ($r = -0.62$, $t(30100) = -136.98$, $p < 0.0001$), see Figure 2.

A statistical multiple regression model fitted to AoA ratings with the nonlinear (restricted cubic splines with 3 knots) function of log word frequency as the predictor, provided the best fit to the data ($R^2 = 0.39$): the lm() function in the R statistical software was used. The consideration of residuals of the model – i.e., differences between the observed and the model-estimated AoA values – revealed that a slightly larger number of words in the entire dataset had positive residuals (15587, or 52%) and were located above the regression line than those that had negative residuals (14515, or 48%) and were found below the regression line. As will become important later, this suggests that overall people rate more words as having been learned later than predicted based on their log frequency.
To investigate effects of paradigmatic connectivity on AoA ratings, we calculated the number of negative and positive residuals from the regression line (based on the entire data set of 30,000 words), for each target word and each respective left or right family. Percentages of family members with negative residuals are reported in Table 1 as columns 6 and 7 respectively. Figure 3 illustrates this for the target word *light* by plotting (a) the functional relationship between AoA ratings and log frequency (solid line) based on the 30,000-word data set of Kuperman et al. (in press), (b) AoAs and frequencies of all members of the left (circle) and right (square) constituent family, and (c) the AoA and frequency of the target word (grey triangle) that is shared by the respective families as either the initial or the final constituent.

*Figure 3. AoA ratings plotted against log frequency of light and members of respective left and right families. The solid line represents the functional relationship between AoA ratings log frequency based on the entire data set. The target word light was left-embedded in 13 complex words, indicated with circles, and right-embedded in 33 complex words, indicated with squares. The triangle shape represents the target word light.*

Figure 3 demonstrates the pattern presented by the majority of the families we considered. Most members of both families are located below the regression line, i.e. have negative residuals. This implies that their AoA ratings are earlier than the ones expected on the basis of their frequencies of occurrence. In fact, for the vast majority of target words – 39 and 40 out of 42 for the left and right constituent families respectively – the percentage of family members with negative residuals was equal to or larger than that of family members with positive residuals. Across target words, the average percent of family members with negative residuals was 77% for left constituent families and 66% for right constituent families: this prevalence was significant for both families, as indicated by the chi-squared test (both $p < 0.001$). These percents were also significantly ($p < 0.01$) higher than the percent of words with negative residuals that were observed across the entire lexicon, 48%. The prevalence of families where the majority of members are acquired earlier than expected, points to an advantage
in the age-of-acquisition that morphologically complex words organized into families have over simplex words or complex words outside of families.

Percentages of members with negative vs positive residuals do not reflect the magnitudes of the deviations between the observed and expected AoAs, only their directions. To address this, we calculated sums of residuals of left and, separately, right constituent family members for each of 42 target words, reported as columns 9 and 10 in Table 1. For each target word, deviations from the regression line were added together for the right constituent and also separately for left constituent families to compute respective sums of residuals. If the resulting sum was negative, family members were overall rated as having been learned earlier than expected based on word frequency.

Figure 4 demonstrates a common pattern by example of constituent families of “game”. There are one right and two left constituent family members above the regression line. Moreover, the one member of the right constituent family with the positive residual is relatively close to the regression line, unlike the circled complex word which represents one of right constituent family members with a large negative residual. Another member of the right constituent family is just below the line, and thus has a small negative residual. When the positive deviations and the negative deviations from the regression line are added for the right constituent family the result is a negative number, i.e. the family is said to have the negative sum of residuals (−5.28). The sum of residuals for the left constituent family is also negative (−0.17).

![Figure 4](image-url)

*Figure 4. AoA ratings plotted against log frequency of game and members of respective left and right families. The solid line represents the functional relationship between AoA ratings and log frequency based on the entire data set. The target word game was left-embedded in 5 complex words, indicated with circles, and right-embedded in 3 complex words, indicated with squares. The triangle shape represents the target word game. The vertical dashed lines represent positive residuals and the vertical solid lines represent negative residuals. The arrows point to the shortest (positive) and longest (negative) residual lines. Left sum of residuals = −0.17, right sum of residuals = −5.28.*
Again, a statistically reliable majority of families (40 and 35 out of 42 for left and right constituent families respectively, both ps < 0.001 in the chi-squared tests) showed negative sums of residuals. Not only were the early-acquired family members in majority, but also the relative AoA advantage of early members outweighed the AoA lag of the late members. This confirms the tendency to associate complex words from productive families with earlier AoA, arguably by virtue of their paradigmatic connections within sets of structurally similar words.

Correlational analyses of the target families revealed a strong negative correlation between family size and the family sum of residuals (left: $\rho = -0.60$, p < 0.001; right $\rho = -0.70$, p < 0.001): here and in the remainder of the paper we use the non-parametric Spearman regression due to the skewness of data distributions. Similarly strong negative correlations were observed between family size and the percentage of negative residuals in the family (all ps < 0.001). Thus, the larger the family, the more members it had below the regression line and the higher the likelihood that when the residuals of all member words were summed (sum of residuals), the result would be a negative number. This indicates that raters judged members of larger families to be learned earlier than expected on the basis of their frequency. Taken together, our findings so far can be interpreted as indices of the preferential attachment principle. Complex words are more likely to connect to the network earlier as they typically specify the meaning of the shared constituent and thus attach to an existing semantic hub (a node with existing connections e.g.: football, basketball, baseball as more specific types of balls) rather than form their own hub in the network. As the likelihood of attaching to a hub is proportional to the number of existing connections to the hub (Steyvers & Tenenbaum, 2005), it is more likely that the families of early acquired words will be larger than the families of words learned later.

The model of Semantic Growth also predicts that if the target word (the hub of the family network) is learned early, family members (connections to the hub) will also be learned earlier. This is indeed what we found. For each positional family (left and right), we calculated the mean residual (i.e. the mean distance between the observed AoA and the AoA predicted from word frequency) as the ratio of the sum of residuals and the family size. The mean residual is a measure of how much advantage or lag in AoA an average member of family has: this measure is independent on how large the family is. A negative mean residual points to an earlier acquisition of an average family member, a positive one points to a lag. The mean residuals are reported in Table 1 as columns 9 and 10. Calculating mean residuals enabled us to tap into the relationship between the AoA of the family and the AoA of the target word. Further analyses revealed that the mean residual for right constituent families was positively correlated with the target word’s AoA ($\rho = 0.51$, p < 0.001), as were the mean residuals for left constituent families ($\rho = 0.53$, p < 0.001) (Table 1, column 8). These

2 The correlations held even if the residual AoA, and not the raw AoA, was considered for target word: there was a positive correlation between the residual AoA of the target word and the average residual of the left and the right constituent family (left: $r = 0.59$, p < 0.001; right: $r = 0.30$, p = 0.05).
correlations match the prediction of the Semantic Growth model on the privileged status of an early hub, i.e. an early-acquired target word, as a preferential and early attractor of connections.

![Figure 5A: Left](image1.png) ![Figure 5B: Right](image2.png)

*Figure 5A: Left* and *Figure 5B: Right*

*Figure 5. Correlations between mean residuals of the left and right constituent family and AoA ratings of target words.*

It is a logical possibility that the observed tendency to associate complex words from large families with an earlier AoA is not due to formal and semantic connections between family members. Rather it may be a reflection of the rater’s familiarity with a frequently recurring string of characters embedded in a large number of other words. This increased familiarity of an embedded word segment (not necessarily a morpheme) may lead to a boost in word identification and affect the evaluation of AoA (Bowers, Davis, & Henley, 2005; see Baayen et al., 2006 for related criticism). We tested this possibility by considering 100 pseudo-complex words from the stimuli list of Rastle et al. (2004). All these words embedded simple words but not as morphemes (e.g. arse-arsenal; bone-trombone). That is, there is a full orthographic overlap between a word embedded in the pseudo-complex word and the pseudo-complex word, just like there is a full orthographic overlap between a target word (a shared compound’s constituent) and all members of the respective family (e.g., house-doghouse). If it is the familiarity of word segments that is at stake, we would see an advantage in AoA ratings for pseudo-complex words, just like we did for genuinely complex words. The log frequency of the pseudo-complex words ranged from 0.69 – 9.42, which is comparable to the 0.69 – 12.66 range for the complex words. Our results ran counter to this hypothesis. Out of 93 pseudo-complex words that overlapped with our dataset, 58% were rated with a later AoA than suggested by their log frequency (i.e. had positive residuals), and only 42% were judged to be learned earlier than expected (i.e. had negative residuals).
This proportion did not differ significantly (p >0.05 in the chi-squared test) from that in the overall dataset of 30,000 words (48% words with negative, 52% with positive residuals). However, the prevalence of later-than-expected AoA ratings in pseudo-complex words was significantly different from the opposite trend observed in the left and right constituent families of our 42 target words (ps in chi-squared tests <0.01). We conclude that the advantage in the age-of-acquisition that is observed in English morphological families cannot be explained by the mere orthographic overlap of the embedding word and the lexical string it embeds, as seen for instance in pseudo-complex words, but is due to the morpho-semantic overlap.

GENERAL DISCUSSION

Correlational analyses of AoA ratings to English morphologically complex words, pitted against a much broader variety of words, indicated several regularities that provide insight into the formation of morphological families in the mental lexicon. First, we observed that morphologically complex words that belonged to a morphological family tended to be learned earlier than words that had the same frequency of occurrence but were not part of a morphological family. This observation stemmed from a comparison of a subset of words that shared morphological constituents to the remainder of a 30,000 set of words that were either simplex or did not form morphological families. Second, the larger the family, the earlier the AoA ratings to words in those families were, as evidenced by the strong negative correlation between AoA ratings to members of a morphological family and family size. Third, words in the families were acquired earlier if the constituent they shared was acquired earlier. That is, the apparent overall advantage in learning complex words, compared to other frequency-matched words, is independently strengthened by two factors: the strength of paradigmatic support for the word determined by the size of the morphological family it belongs to, and the age at which the kernel of the family, the shared constituent, is acquired. We also showed that the observed correlations were not due to a purely formal overlap: most pseudo-complex words (trombone vs bone) were rated as acquired later, in line with an overall trend in the large set of English words, but contrary to the observed advantage of AoA in words from morphological families.

Our findings dovetail perfectly with the nomenclature and predictions of the model of Semantic Growth (Steyvers & Tenenbaum, 2005), on the assumptions that (A) a morphological family is a semantic network with a single hub, i.e. a semantic node yielding multiple connections, and (B) morphological families are connected into larger semantic networks via shared nodes (e.g., the compound postman is shared by both the right constituent family of man, which also includes fireman and policeman, and the left constituent family of post, which also includes post office and postcard). As shown here and in De Jong (2002), morphological families in languages like English and Dutch
show the critical property of semantic networks, namely the scale-free power-law distribution of the number of connections (i.e., family sizes). Our data also corroborate the prediction of Steyvers and Tenenbaum (2005) that the speed of growth and the size of the network depends on how early the hub is acquired (AoA of the target constituent) and how many connections (family size) it has at the time a new concept joins the network. To sum up, the convergence between present findings and architectural premises of Steyvers and Tenenbaum’s model suggests that the model can be fruitfully applied for simulating the formation of morphological families in the mental lexicon, and adds to the known inventory of semantic networks that operate on English words (e.g., synonyms and semantic associates, see Griffith et al., 2007).

It is important to realize that this model makes a strong commitment to a specific networking structure of semantic memory and to semantic differentiation as a mechanism of its growth in the developing mental lexicon. This commitment runs counter to conceptual premises of several widely received models including Latent Semantic Analysis (henceforth, LSA; Landauer & Dumais, 1997) that advocates statistics of co-occurrences between words and concepts as the organizing principle of the mental semantic space. The learning mechanism associated with LSA is such that a child establishes the semantic similarity between words by encountering those words in similar contexts. Steyvers and Tenenbaum (2005) demonstrate that LSA cannot account for the observed properties of several semantic networks. For instance, the architecture of LSA does not differentiate between simple nodes and hubs, i.e. privileged semantic nodes that attract an increasing number of connections, nor does it replicate the empirically observed power-law distribution of connections in the network. We remind the reader that power-law distributions of family sizes are observed in our data, contrary to predictions of the LSA account of lexical learning. Whether semantic differentiation, or mental abstraction over co-occurrences of concepts, or some other learning mechanism, is at the core of the language learning experience of a child is then a question for the future research: our present data provides evidence in favor of the former.

**Future directions.** Several aspects of the present study require further research. We identify the need to quantify the strength of semantic relationships between members of the family and between family members and the target words. Here we made a simplifying assumption that all family members are associated with the shared constituent to the same degree. Yet it is known that some members of large families may be semantically different from the target word and other family members in the present-day English (e.g. *hogwash* in the right constituent family of *wash* used to mean the liquid food for pigs when registered in the mid–15th century, then was extended to mean “cheap liquor”, but now has the meaning of “nonsense”). As demonstrated in Moscoso del Prado Martín et al. (2004b), family members with unrelated meanings attenuate the facilitatory effect of the family size on complex word recognition (see also a discussion in Balling and Baayen, 2012). It stands to reason that variations in the strength
of semantic relatedness between individual family members and target words may also affect AoA ratings and the formation of families as semantic networks. There may also be systemic differences between left and right constituent families. Since the right constituent is the head in most English compounds, left constituent family members may not be as closely related as right constituent family members, e.g., birdcall and birdhouse are not as semantically close as birdhouse and doghouse (e.g., Libben et al., 2003). The degree of semantic uniformity within morphological families may thus be an additional factor predicting the likelihood that the family attracts more members, and the speed at which such attachments are made.

Acknowledgments

We thank two anonymous reviewers for their feedback on the earlier draft of this paper. We also benefitted from the feedback of the audience at the Mental Lexicon conference, Montreal, 2012. This research is supported by the NSERC Discovery grant 402395–2012 to VK.

REFERENCES


