Antonym adjective pairs and prosodic iconicity: Evidence from letter replications in a blogger corpus

Susanne Fuchs¹, Egor Savin¹², Stephanie Solt¹, Cornelia Ebert¹, Manfred Krifka¹²

¹Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany
²Humboldt-Universität zu Berlin, Germany

fuchs|savin|solt|krifka|ebert@leibniz-zas.de

Abstract

While the general assumption has long been that natural languages exhibit an arbitrary pairing of form and meaning, there is increasing empirical evidence that iconicity in language is not uncommon. One example from spoken language involves iconic prosodic modulation, i.e. the changing of prosodic features such as duration and fundamental frequency to express meanings such as size and speed. In this paper, we use data from an English social media corpus (140 million words written by 19320 bloggers) to investigate a counterpart to iconic prosodic modulation in written language, namely letter replications (e.g. looooong). We examine pairs of gradable adjectives such as short/long, tiny/huge and fast/slow, finding a higher frequency of letter replications for adjectives associated with greater size or spatial/temporal extent. We did not find an iconic effect of the number of letter replications, though younger bloggers were found to make greater use of letter replication than older bloggers. Our results show evidence for iconic prosody in written language, and further demonstrate that social media databases offer an excellent opportunity to investigate naturalistic written language.

Index Terms: Iconic lengthening, prosody, social media, blogger corpus, English

1. Introduction

1.1. Iconicity, prosody and semantics

It is frequently assumed that one “design feature” of language is that there is no relation between sound and meaning [16], which makes languages so different and flexible. However, there is accumulating empirical evidence that this statement is only partially true. Blasi and colleagues [9] analyzed 100 basic vocabulary items of 4298 different languages and 359 lineages. They found evidence that certain phones are favored or avoided in these 100 words, e.g. /v/ is favored in words meaning ‘small’. Thus, iconicity in spoken language may be less exceptional than expected.

Perniss and colleagues [6] argue that “iconicity is a fundamental property of language, representing an adaptation to a critical constraint on the phylogenesis, ontogenesis and use of language, namely the need to map linguistic form to human (sensory, motor and affective) experience.” (p.2). They also mention that prosody might contribute to iconicity in spoken language, for example by an extra lengthening of loooong in referring to the experience of an event taking a long time. Schlenker [4] has recently proposed that such lengthening is an iconic enrichment of arbitrary encoding of linguistic meaning.

Iconic prosodic modulation consists of changing certain prosodic features such as duration, fundamental frequency (F0), or amplitude to express additional meaning. In this line, Nygaard et al. [7] investigated the relation between prosody and meaning in two experiments. Three speakers had to read novel words in an infant-directed speech style to stimulate a situation of engagement. The novel words, embedded in a frame sentence, were first read with a relatively neutral prosody as a baseline. Hereafter, the words were presented together with pictures that referred to the meaning of an adjective from the pairs: happy/sad, hot/cold, big/small, tall/short, yummy/yucky, and strong/weak. Speakers had to read the words again. Differences in acoustic properties were found (mean F0, F0 variation, duration, and amplitude) depending on the adjective. The authors also investigated whether listeners could reliably infer the meaning of these novel words. For this purpose, listeners saw two pictures representing an antonym pair, heard one of the previously recorded sentences, and had to choose the picture that would correspond to the perceived novel word. Listeners were significantly better in choosing the right picture when listening to the speaker’s meaningful prosody than to the speaker’s neutral prosody as well as when prosody matched than when it mismatched. The authors suggest that prosody could augment, disambiguate, or reinforce meaning (p.142).

Perlman [14] showed video clips involving fast and slow events to the participants of the experiment. They had to recall the different events and they did so by generally talking faster for the faster events and slower for the slower ones, without being instructed to do so. In a more recent study, Perlman and colleagues [8] extended these findings from the manner of motion to the size of an entity and additionally from concrete to abstract meanings (e.g. concrete: a fast drive; abstract: slow career progress). Speakers had to read short stories involving one of these semantic dimensions to a partner. The authors predicted that stories with different manners of motion would go hand in hand with prosodic variation in duration, but not fundamental frequency, while the reverse should be the case for stories varying adjectives corresponding to the dimension of size. Their findings show that stories in the small condition were read with higher F0 than stories in the large condition, for both abstract and concrete meanings. Moreover, stories in the fast condition were read within a shorter duration than stories in the slow condition and no differences in F0 were found. Thus, different acoustic parameters may be used to mark different semantic dimensions.

In three experiments, Shintel and colleagues [13] recorded speakers who described either a dot moving in upward and
downward direction or dots with moving with different speed. They were able to show that motions in the vertical dimension go hand in hand with changes in F0 in a similar direction and changes in speed coincided with changes in speech rate. In some follow up studies, Shintel [11] and colleagues could also provide evidence that the speed of recorded instructions had an effect on time of listener’s response. In [10, 12] the findings were extended to novel word learning, showing that congruent prosody has a positive effect on memory consolidation.

1.2. New prospects on prosody with recent technological development

With the progress in technology (e.g. computer, smartphone, tablets, fibre optic cables, satellites), digital communication has had an enormous impact on our daily life and our communication tools and styles. Social media platforms have been developing (e.g. twitter, instagram, facebook, chats, whatsapp, blogs) in parallel. They have many common features with spoken language, since writers do not follow all the formal rules of traditional written norms, and they are further enriched with icons of emotional expressions (emojis). These platforms provide a great opportunity to investigate the dynamics and creativity in the use of written language beyond prescriptive rules [5]. Using social network databases has the advantage of getting a vast amount of data (in our case ca. 140 Mio words from more than 19000 bloggers) in completely natural settings, with participants that would not feel constraint like in laboratory experiments. It has, however, the disadvantage that the data might be confounded by many unknown factors. Perlman and colleagues [8] write that “[r]emarkably little is known about how speakers use iconic prosody in the wild” (p.1349). We hope to contribute in this respect by using a social media corpus.

Among others, the idea of using social media for a better understanding of prosody goes back to Brody and Diakopoulos [2], who considered word lengthening by letter replications as an expression of prosodic emphasis. These letter replications are a way of signaling the writer’s sentiment and emotion in written text where some properties of spoken language, like intonation, are partially absent. In our previous work [3] we provided evidence that letter replications are a frequent phenomenon that occurs for all letters, that bloggers replicate most often between 3-10 letters, and that younger writers (<20 years) replicate letters more often than older writers (writers in their 30s): for example so0000000000 instead of so0000.

1.3. Research questions and expectations

On the basis of the literature on prosody and iconicity we ask the following research questions:

- In which adjectives does lengthening (letter replications) occur?

Since we assume that prosody and letter replications have some degree of iconicity, we predict that letter replications are a phenomenon in adjectives that express the larger size, e.g. in long rather than in short.

- If replications occur, of how many letters do they consist?

We would expect longer words for the larger size dimension. However, in previous work [3] we have also shown that the number of letters varies in a nonlinear way with age, with younger bloggers using more replications than older ones. This age-related effect may result in some confounds.

2. Methodology

2.1. The blogger corpus

We used an English social media corpus that is freely available for non-commercial use. It consists of approximately 140 million words written by 19320 bloggers in August 2004 [1]. The age of the bloggers ranges from 13-47 (in three age groups) with an equal number of males and females.

2.2. Selected adjectival antonym pairs

The following antonymic adjectives are comparable to the different studies described in the introduction, in particular the ones on size and motion, because they have shown some effect on duration [8,14] and there are measures of F0 possible. The following adjective pairs have been selected (Table 1).

<table>
<thead>
<tr>
<th>Smaller dimension</th>
<th>Larger dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>long</td>
</tr>
<tr>
<td>small</td>
<td>big</td>
</tr>
<tr>
<td>tiny</td>
<td>huge</td>
</tr>
<tr>
<td>near</td>
<td>far</td>
</tr>
<tr>
<td>fast</td>
<td>slow</td>
</tr>
<tr>
<td>thin</td>
<td>fat</td>
</tr>
<tr>
<td>narrow</td>
<td>wide</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>teeny, itty-bitty</td>
<td>giant, gigantic</td>
</tr>
<tr>
<td>slight, microscopic</td>
<td>gargantuan, humongous, massive, vast</td>
</tr>
</tbody>
</table>

In contrast to nouns, adjectives typically map entities to a one-dimensional dimensional scale. For this reason, most adjectives have comparative forms, as in x is longer than y. For many adjectival dimensions, we can distinguish a “small” direction, often with a zero point, and a “large” direction, expressed by antonym pairs like short and long. Typically, the adjective that expresses the “large” direction is the unmarked pair to the vertical axis, while the narrow-fat pair to the horizontal axis may correspond to the high-low pair to the vertical axis and small-big. tiny-huge do not make specific reference to the axis and are rather general size properties.
2.3. Data extraction and preprocessing
The NLTK toolkit was used as a natural language processing environment to tokenize the corpus (http://www.nltk.org/api/nltk.tokenize.html). All lower and upper case tokens were considered together. To further process the data, we used Python 2.7. and R ([15], version 3.4.1).

In a first step all replications of letters were removed from the selected adjectives in Table 1. This also included repetitions which are the orthographic norm (e.g. double l in small) and resulted in an order of strings. Hereafter, we searched for these string orders including repetitions of the same strings. From the resulting corpus the following words were eliminated: all words that do not involve all the letters of the original word (e.g. narrow instead of narrow), all words that differed in just one letter and could potentially be typos (e.g. thin instead of thin), and finally all words that might have a different meaning (e.g. tiny instead of tiny).

Furthermore we calculated the overall number of cases for each adjective with and without letter replication. The number of cases without letter replications served as a baseline to calculate how often bloggers wrote the specific word in the orthographic norm. The number of cases including additionally letter replications was set to 100 percent for each adjective in order to calculate the frequency at which bloggers wrote the selected adjectives with letter replications.

In addition, we calculated the length of the words as the number of all letters as well as the number of replicated letters.

3. Results

3.1. Percentage of adjectives with letter replications
The adjectives in the last row of Table 1 were excluded, because almost all of them did not occur very frequently and even if they occurred, they did so without letter replications.

Table 2: Adjectival antonym of dimensional pairs

<table>
<thead>
<tr>
<th>Antonym pairs</th>
<th>Percentage of word with letter replications from all</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=No. of occurrences of the adjective with &amp; without replications</td>
<td>Smaller dimension</td>
<td>Larger dimension</td>
</tr>
<tr>
<td>short (n=10848) long (n=144819)</td>
<td>0.018% 1.348%</td>
<td>1.33 &lt;</td>
</tr>
<tr>
<td>small (n=12487) big (n=29007)</td>
<td>0.024% 0.155%</td>
<td>0.131 &lt;</td>
</tr>
<tr>
<td>tiny (n=2317) huge (n=8556)</td>
<td>0.13% 0.725%</td>
<td>0.595 &lt;</td>
</tr>
<tr>
<td>near (n=6557) far (n=19424)</td>
<td>0.015% 0.0721%</td>
<td>0.057 &lt;</td>
</tr>
<tr>
<td>fast (n=7827) slow (n=4696)</td>
<td>0.102% 0.958%</td>
<td>0.856 &lt;</td>
</tr>
<tr>
<td>thin (n=1427) fat (n=5886)</td>
<td>0.070% 0.068%</td>
<td>-0.002 =</td>
</tr>
<tr>
<td>narrow (n=458) wide (n=1915)</td>
<td>0% 0.209%</td>
<td>0.209 &lt;</td>
</tr>
<tr>
<td>low (n=5022) high (n=16159)</td>
<td>0.060% 0.037%</td>
<td>-0.023 &gt;</td>
</tr>
</tbody>
</table>

Almost all antonym pairs show a higher frequency of letter replications in the adjectives with larger dimensionalities. Letter replications are particularly used in the adjectives long, slow and huge. For example, out of all occurrences of long in the corpus (n=44819) in 1.35% of the words were spelled with letter replications, i.e. different from the orthographic norm. These replications were very unlikely a typo, since we excluded all data with just one additional letter from the norm.

The percentage of occurrence of letter replications can be seen in highly frequent word in the corpus (e.g. long with n=44819), but also in less frequent words (e.g. in slow with n=4696). The group of adjectives with larger dimensions is however, realized much more frequently than the group of adjectives with smaller dimensions (see Table 2).

3.2. Number of replications
In a next step we investigated whether the number of letter replications would also differ among the adjectival antonym pairs. Figure 2 shows boxplots summarizing the length of words (as number of letters) in antonym pairs (x-axis). Note that boxplots corresponding to adjectives with smaller dimensions include fewer data points than adjectives with larger dimensions. The pairs short-long, small-big, tiny-huge, near-far and fast-slow consist mostly between 3-5 replications apart from the default number of letters of the respective word. All other antonym pairs, i.e. thin-fat, wide (but not narrow) and low-high consist of a larger number.
Strikingly, the replication length is longer in some adjectives with smaller dimensionality (e.g. small > big, tiny > huge, thin (consisting of one value only = 57 > fat, low > high)). It is however not the case for every antonym pair, because narrow (never with replication) < wide, fast < slow, and overall the difference in letter replications is rather small (mostly 1-2 letters). We should also keep in mind that letter replications for adjectives with small dimensions are rather rare and to some extent exceptional, while there are considerably more cases in adjectives with larger dimensions.

On the basis of these findings we can exclude the possibility that writers would on average lengthen all words with larger dimensions in comparison to words with smaller dimensions. Antonym pairs were however, not obtained in a contrastive context like: fast versus slow. It might well be possible that difference occur when used as antonym pairs.

Since in our previous work on letter replications [3] we found an effect of age on the number of replications, we did a further analysis where blogger’s age (x-axis) was plotted against the overall word length (Figure 3). Similarly to our previous work, the longest words are realized from younger bloggers while older bloggers realize a shorter word length or do not produce letter replications at all.

4. Discussion and conclusion

With this work, we contribute to the growing literature that shows that natural language is to some extent iconic. Based on the assumption that word lengthening in written language corresponds to prosodic emphasis in spoken language [3], our results show that lengthening is iconic indeed. Considering the three pairs that show a significant difference of the involved items in the percentage of occurrences with letter replications (short-long, tiny-huge, fast-slow), it is always the larger dimension that occurs more often in a lengthened version (long, huge, slow). We did not, however, find an iconic effect of the number of letter replications.

We would like to stress that it is arguably the adjective with the larger dimension that is lengthened and the relevant distinction is not between positive vs. negative or unmarked vs. marked adjective. The pair fast-slow is particularly insightful here. While semantically fast is the unmarked adjective of the two (i.e. not presupposing anything when used in the positive), it is slow that occurs more often with letter replications, corresponding to the iconic effect that we expect.

On a more speculative note, our results suggest that iconicity might even be sensitive to the orientation of the dimension (cf. also [13]). The only pair that makes clear reference to the vertical axis (low-high) and hence does not suggest itself for an iconic mapping with word lengthening (but possibly rather F0 variation) is the pair that sticks out for the fact that the adjective with the smaller dimension is lengthened in more of the cases.

Although we do not want to exclude the possibility that lengthening via letter replication might have other effects than the iconic mapping effect we argue for, such as e.g. emphasis in general, we are convinced that our data show that iconicity is one guiding factor. The clear statistic effects we find are even more remarkable considering the fact that our investigations are based on a naturally occurring non-elicited data set with no artificial contrasts or other contexts that could make the effects even stronger.
5. Acknowledgements

This work was supported by a grant (FU791/6-1) from the German Research Council within the priority program XPRAG.

6. References


