The role of semantics, pre-emption and skew in linguistic distributions: the case of the un-construction

How to cite:

Link(s) to article on publisher’s website:
http://dx.doi.org/10.3389/fpsyg.2013.00989

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
The role of semantics, pre-emption and skew in linguistic distributions: the case of the un-construction

Paul Ibbotson*
Centre for Childhood, Learning and Development, Open University, Milton Keynes, UK

We use the Google Ngram database, a corpus of 5,195,769 digitized books containing ~4% of all books ever published, to test three ideas that are hypothesized to account for linguistic generalizations: verbal semantics, pre-emption and skew. Using 828,813 tokens of un-forms as a test case for these mechanisms, we found verbal semantics was a good predictor of the frequency of un-forms in the English language over the past 200 years—both in terms of how the frequency changed over time and their frequency rank. We did not find strong evidence for the direct competition of un-forms and their top pre-emptors, however the skew of the un-construction competitors was inversely correlated with the acceptability of the un-form. We suggest a cognitive explanation for this, namely, that the more the set of relevant pre-emptors is skewed then the more easily it is retrieved from memory. This suggests that it is not just the frequency of pre-emptive forms that must be taken into account when trying to explain usage patterns but their skew as well.

Keywords: generalizations, verbal semantics, pre-emption, skew, diachronic use

INTRODUCTION

People are both creative and conventional with their language use. Linguistic theory must therefore perform something of a balancing act: granting people enough inventiveness to generate new ways of speaking while at the same time assuming they conform with the linguistic norms of their community. Describing the cognitive and social processes that give rise to this situation is one of the central goals of linguistics.

To take an example, words beginning with the prefix un- such as unlock and unbutton typically designate reversible actions. It sounds somewhat odd, however, to say unclose a door or unlift one's arms even though these are reversible actions and semantically interpretable statements. Here we make use of a relatively new tool to investigate some of the cognitive processes that might be involved in restricting the un- generalization. In 2011 Google made public its Ngram database, a corpus of 5,195,769 digitized books containing ~4% of all books ever published. Using the un-construction as a test case, we use the Ngram database to test three ideas that are hypothesized to account for linguistic generalizations: verbal semantics, pre-emption and skew.

In construction grammar the form of constructions carry meaning independent of the meaning of the items that appear in those constructions. For example, the gazer nibbed the toma is typically interpreted as a transitive not because of the compositionality of the meaning of the items (they are meaningless) but by analogy to other sentences of the form the X Yed the Z. Because this approach allows constructional form to carry meaning as well as lexical items, items can be more or less semantically compatible with the construction they appear in. The prediction is that that the more compatible an item's semantics is with that of the construction the more likely it is to appear in that construction. Semantic fit has thus been suggested as a possible mechanism that can explain linguistic patterns, such as why certain forms resist generalizations in certain directions (e.g., Pinker, 1989; Ambridge et al., 2011). With regard to the semantics of the un-construction, Whorf (1956) suggested it could be described as a kind of crypototype; a cluster of features implicit to the language user (hence crypto), that tend to co-occur—no single one of which could be considered the definitive meaning of the un-form but taken together describe the semantic core of un prefixation. These features include whether one object affects another, touches it, and distorts it for example. The semantic fit hypothesis predicts that un-forms with features that are the closest match to the un-form cryptotype will be the ones that most readily appear in the un-construction. Those forms with semantic features at odds with the semantics of the construction will be resisted and therefore rare in the language.

A further explanation to the distribution of un-forms is that the process of hearing a verb with sufficient frequency in a particular grammatical context prevents that verb from appearing in other structures unless it is also attested in those structures (Braine and Brooks, 1995; Brooks and Tomasello, 1999). The idea that speakers are conservative about their generalizations with items that are entrenched in the contexts they hear them has received some empirical support (e.g., Brooks et al., 1999; Theakston, 2004; Ambridge et al., 2008). This partly answers why some forms resist being “exported” out the construction they are entrenched in, however, it does not explain why it is these forms that show entrenchedness in the first place, rather than some other (semantically possible) forms. One answer to this could simply be linguistic drift. Historically-speaking, some forms have randomly walked into a position where they are more frequent in the language and through the principle of least effort it is these forms that remain entrenched. Despite this, we still need a mechanism that explains, where innovations could occur, and are semantically plausible, why these still might be resisted. One suggestion is...
that as well as items showing resistance to being “exported” out of
their usual grammatical context (entrenchment) there is an addi-
tional resistance to them being “imported” into a new context if
that linguistic niche is already filled with a successful competitor.
Thus, if a verb is repeatedly used in a construction, (e.g., I filled
the cup with water) that serves the same communicative func-
tion as a possible unattested generalization (e.g., I filled water into
the cup), then the attested form blocks or pre-empts the possible
generalization (Pinker, 1981; Clark, 1987; Goldberg, 1995).

In previous work, Ambridge (2013) found experimental sup-
port that adults and children are sensitive to both the semantic fit
and pre-emption when judging the acceptability of various un-
forms. In his study children aged 5–6, 9–10, and adults rated
the acceptability of un- prefixed forms of 48 verbs (and, as a
control, bare forms). The higher the frequency of potentially pre-
empting forms (e.g., open), the less willing participants were to
accept a particular un- form (e.g., *unclose). Semantic fit between
un-forms and the un-construction was also a significant posi-
tive predictor of acceptability, for all age groups. The focus of
that study was synchronic use; investigating the mechanisms by
which children retreat from overgeneralization in the process of
language acquisition. Thus, semantic features and pre-emption
were originally concepts developed to solve problems in the lan-
guage acquisition domain. Here we apply these processes in the
diachronic domain; building on Ambridge’s findings we asses to
what extent semantic fit, pre-emption and skew (which Ambridge
did not consider) can explain the historical distribution and
frequency of un-forms.

Skewed distributions have been shown to have a beneficial role
in learning linguistic and non-linguistic categories [Elio and
Anderson, 1984; Avrahaemi et al., 1997; Gentner et al., 2002;
Casenhiser and Goldberg, 2005; Year, 2009, but see Cordes
and Krajewski, in preparation; Year and Gordon, 2009 1]. Cognitive
anchoring is one mechanism that has been suggested to play a role
in this effect (e.g., Goldberg et al., 2004; Goldberg, 2006, p. 88).
According to the theory, a high frequency exemplar that emerges
early in learning provides a salient standard of comparison against
which subsequent instances are judged. For example, Casenhiser
and Goldberg (2005) investigated how children map a novel
meaning onto a new phrasal form, and how this process could
be facilitated by constructions that are centered around prototyp-
ical verbs. Fifty one 5- to 7-year-old children watched a short set
of video clips depicting objects appearing in various ways. Each
scene was described using a novel verb embedded in a novel con-
struction. The novel construction NP1 NP2 Novel Verb + O was
indicative of the meaning of “appearance.” The different verbs
indicated manner of appearance. For example, children heard
the spot the king moo-po-es, they then saw a video-clip where a
spot appears on the king’s nose, and then they heard the spot
the king moo-po-ed. There were five novel verbs and 16 ex-
amples of the construction so that in the more balanced condition

1 Year and Gordon (2009) finds a consistent null effect of skewed input on
Korean speakers’ acquisition of the English ditransitive construction. However
when the experiment’s full dataset is considered (see Year, 2009), there is
evidence that suggests that skewed input gives participants an increased
awareness of semantic constraints on the use of the ditransitive early on.

Children who watched the videos and heard the accompanying
description were able to match new descriptions that used
the novel construction with new scenes of appearance. There was
a facilitative effect for the disproportionately high frequency of
occurrence of a single verb in a particular construction such as
has been found to exist in naturalistic input to children (Goldberg
et al., 2004, p. 3). The statistical skew in the input on this view
facilitates relatively rapid category construction compared to flat-
ter distributions because a minority of high frequency types
provide a relevant anchor that serves to organize memory and
reasoning about other types.

The idea here is that a similar process may be at work in
restricting un-construction generalizations. The more the set of
relevant pre-emptors is skewed then the more easily an alternative
form for the un-construction is retrieved from memory because
it has a higher resting activation level. For example, unclose might
be pre-empted as a generalization not just because open is fre-
quent [as established by Ambridge (2013)] but because open is a
clear modal alternative to unclose in the set of forms competing
to express a similar concept, making it relatively easier to access
from memory compared with flatter distributions.

To add a diachronic perspective to this analysis we make use of
so-called “Big Data” in the form of Google’s Ngram corpus. Using
this dataset Michel et al. (2011) showed how it was possible to
provide insights about fields as diverse as lexicography, the evolu-
tion of grammar, collective memory, the adoption of technology,
the pursuit of fame, censorship, and historical epidemiology.

A corpus of words extracted from written text is clearly dif-
ferent from that of spoken language. In general the printed word
is subject to more deliberation both on behalf of the writer and
the publisher (it is more planned, proofed, edited and so on). One
might expect more conservative use of forms on this basis.
However, the written word does not rule out non-conventional
use. In creative writing genres such as fiction and poetry we
might expect more unconventional use than in spoken discourse
where authors push and play with what is considered conven-
tional (many of these texts will of course be quoting dialog also).
Second, the relative performance of Ngram to other printed word
corpora has proved particularly useful for detecting linguistic
“dark matter.” Using Google Ngram Michel et al. (2011) estimated
that 52% of the English lexicon—the majority of the words used
in English books—consists of lexical material undocumented
in standard references like the Oxford English Dictionary and
the Merriam-Webster Unabridged Dictionary. If unconventional
forms (by today’s standards), like unwont, uncame, and ungive,
were once more frequent, the increased coverage of this corpus
gives us a better chance of detecting them. Third, assuming writ-
ten text is relatively impoverished in unconventional un-forms, if
there are enough raw forms to perform the analysis, we are look-
ing at changes in use within the bounds of conservative use and
not making claims about absolute levels (which may or may not
be higher in spoken language).

Here we focus on the use of un-forms and their respective pre-
emptors from 1800–2008. We assess the extent to which the
combination of semantics, pre-emption and skew can account for
usage patterns of the un-construction.
METHODOLOGY

SEMANTIC FEATURE RATING FOR UN-FORMS

In order to see the effect of semantics on historical frequency we need some measure of the extent to which particular verbs exhibit the properties of the un-form semantic cryptotype. Li and MacWhinney (1996) asked 15 native English speakers to rate un verb forms such as unfreeze, unbuckled and unbent on the basis of whether they exemplified the 20 features thought to be relevant from Whorf’s earlier cryptotype work (Whorf, 1956):


Thus, each verb had a score of between 0 and 15 for each semantic feature, corresponding to the number of participants who judged the feature to be relevant to the verb’s meaning. Using a subset of Li and MacWhinney’s original verb list, Ambridge (2013) condensed the 20 semantic predictors using Principle Components Analysis, selecting the single component with the largest eigenvalue. The result of this is a single value for each verb showing how well (or poorly) it fits the cluster of semantic features of the un-prefixation cryptotype. We use this value to cluster our verb types well (or poorly) it fits the cluster of semantic features of the un-form semantic cryptotype.

Ambridge chose the two most commonly suggested pre-empting forms for each verb, for example for unsqueeze it was release and loosen. For each pair of pre-empting forms, Ambridge calculated the natural log of \( N + 1 \), where \( N \) was the frequency of the pre-emptor in the British National Corpus. These two values were then summed to yield the pre-emption measure. Ambridge also calculated various other measures (e.g., frequency of the single most commonly suggested form, sum frequency of the top two forms/all forms, number of suggested forms, number of participants suggesting the top form), but all were less successful as predictors of un-form acceptability. As Ambridge notes, the advantage of including only the two most commonly suggested forms—as opposed to all forms—is that it excludes the more marginal examples that were only suggested by one or two participants. However, we consider all pre-emptive forms here that were suggested by at least one participant in Ambridge (2013) as we are not only interested in the top two pre-emptive forms but the shape of the distribution of the forms (analysis 3). The number of types and tokens for a set of pre-emptors associated with each verb was recorded and the skew calculated using the adjusted Fisher-Pearson standardized moment coefficient, a common statistical measure of skew defined by Equation (1).

\[
S = \frac{n}{(n - 1)(n - 2)} \sum \left( \frac{x_j - \bar{x}}{s} \right)^3
\]

where \( n \) is the set of numbers for which we want to calculate skew, where \( x_j \) is any particular score within that set, and \( s \) is the estimated standard deviation. A potential concern with this statistical measure is that it does not completely capture the sense in which skew has been discussed in the linguistic literature. Recall, Casenhiser and Goldberg (2005) used a skewed distribution in their experiment of 8-2-2-2-2 with one item (represented by the 8 here) making up the lion’s share of the token count in comparison with a flatter distribution of 4-4-2-2-2. The analogous example in English is where the verb give makes up the lion’s share of ditransitive uses whereas the transitive construction has a flatter distribution with no one item being the clear modal verb (Figures 1A,B).

Returning to the present measure, the Fisher-Pearson skew defines any distribution that is symmetrical about the mean (e.g., \([5,5,5],[3,4,5],[1,5,9]\)) as equally non-skewed and returns a value of 0—based on the fact that the mean and median are the same for these distributions. Distributions count as skewed when they are not symmetrical about their means (e.g., \([1,5,13]\)). Now, a distribution of \([1,5,9]\) would count as skewed under most linguistic discussions of skew (Goldberg’s definition of give in the ditransitive for example; Goldberg et al., 2004) so in theory there is potentially a disconnect between the cognitive a anchoring

www.frontiersin.org
definition of skew and our definition. However, all of our samples showed some degree of skew due to the very robust finding that natural language follows something like a power-law distribution (Zipf, 1935/1965). So in practice, we were only ever measuring the degree to which the sample is skewed, and given that there are many more ways to be non-symmetrically distributed about the mean than to be symmetrically distributed this problem was avoided. None of the preemtions sets analyzed in this study were symmetrically distributed about the mean (which would be confounded with a flat distribution) so we could be confident that the measure was capturing the degree to a sample was skewed. For example, for the possible un-prefixation unstraighten the pre-emt set was as follows [12-2-2-1-1-1...] which in terms of types was respectively bend-crease-curl-curve-constrict-misalign...While the examples have been chosen here from opposite ends of the flat/skewed spectrum to illustrate a point, the measure of skew employed here is inherently continuous and so can capture the extent to which a distribution is skewed, rather than some binary distinction between flat and skewed.

CORPUS EXTRACTION PROCEDURE

We extracted the same set of 48 verbs Ambridge used in order to have independent semantic and pre-emption measures. We obtained the time series of word frequencies via Google's Ngram tool (http://books.google.com/ngrams/datasets) in the 1-gram English dataset (combining both British and American English). The corpus gives information on how many times, in a given period of time, namely, those with the best fit have been getting more frequent. The results show that natural language follows something like a power-law distribution (Zipf, 1935/1965). The results show an increasing trend (significant positive correlations) in the frequency of the best and the better fit semantic categories. The data also show a decreasing trend for average, poor and poorest fit semantic categories. On the advice of one reviewer we also analyzed the historical trends based on the raw frequency data (raw column Table 1 and Figure A1 in appendix; For those interested, item analyses are also presented in the Appendix, both for proportions, Figure A2, and raw frequencies, Figure A3). The pattern observed in the proportion analysis is largely confirmed in the raw frequency analysis. The two analyses answer slightly different questions (i) can semantics predictive the frequency of un-forms relative to other un-forms and (ii) whether semantics predicts the relative frequency of un-forms as a function of corpus frequency (in this latter analysis we divide raw scores by the frequency of “the” as outlined in the procedure). What this means is that the semantic crypotype identified and tested by Whorf (1956), Li and MacWhinney (1996), and Ambridge (2013) is a good predictor of how frequency of these forms changes over time, namely, those with the best fit have been getting more frequent those with the poorer fit have been getting less frequent. This does not tell us whether the rank order of the semantic fit categories is what we would expect from their semantics. To examine this we compared the rank order of the semantic groups against what we would expect, namely those with a better fit to the crypotype should be more frequent (and thus a higher rank) than those with a poorer fit, Table 2.

The results show the semantic rating does a good job overall of predicting the rank order of the frequencies over the historical period of English use covered by Google Ngram. From 1800 to 1850 it predicts the rank order of the poor and poorest fit categories but fails to predict the correct rank of any other categories. From 1851 to 2008 the predicted rank is a statistically
Ibbotson
Semantics, pre-emption and skew

Figure 2 | Historical trends of un-forms (N = 828, 813) as a function of semantic rating. M = mean semantic rating for the nine verbs in each group, ranges are in square parentheses, verbs in curly parentheses. Note the y-axis is relative frequency, thus at any slice on the x-axis values sum to 100%. We are interested in whether the historical proportion of un-forms forms is explainable in terms of their semantic rating. Linear lines of best fit are plotted to support the Pearson correlation (see analysis below).

<table>
<thead>
<tr>
<th>Semantic rating</th>
<th>%</th>
<th>Raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best fit</td>
<td>$r_{(209)} = 0.461, p &lt; 0.001$</td>
<td>$r_{(209)} = 0.558, p &lt; 0.001$</td>
</tr>
<tr>
<td>Better fit</td>
<td>$r_{(209)} = 0.645, p &lt; 0.001$</td>
<td>$r_{(209)} = 0.639, p &lt; 0.001$</td>
</tr>
<tr>
<td>Average fit</td>
<td>$r_{(209)} = -0.714, p &lt; 0.001$</td>
<td>$r_{(209)} = -0.307, p &lt; 0.001$</td>
</tr>
<tr>
<td>Poor fit</td>
<td>$r_{(209)} = -0.254, p &lt; 0.001$</td>
<td>$r_{(209)} = 0.03, p = 0.667$</td>
</tr>
<tr>
<td>Poorest fit</td>
<td>$r_{(209)} = -0.333, p &lt; 0.001$</td>
<td>$r_{(209)} = -0.270, p &lt; 0.001$</td>
</tr>
</tbody>
</table>

So semantics help to explain some of the overall historical trajectories of un-forms yet there is clearly some historical variation left to explain, for example, why particular un-forms of the same “fit” category still show historical variation in frequency within their category. One possibility is that importing close, for example, into the un-cons construction is blocked because open already serves this communicative purpose very well. More than that, we wanted to know (a) whether there is any historical evidence for this idea and (b) whether the skew of the preemptor set also has an effect on the frequency of the un-form.

To analyse this possibility six verbs were randomly selected from each semantic group. As this analysis involved many more data points than the previous analysis and much more data processing per analysis, we chose six verbs per category to make the task more manageable. For each un-form and its top two semantic competitors we plotted the historical frequency data. As we are also interested in shape of the distribution of the competitors—not just their frequency—we included all competitor forms. The skew plots that relate to this appear directly beneath each verbal plot, Figure 3.

In theory, the relationship between un-forms and their preemptors could be positive—both forms increase or decrease relative to each other—or negative—as one increases the other decreases. To analyse these possibilities we collapsed the individual verbs displayed in Figure 3 into the semantic

www.frontiersin.org December 2013 | Volume 4 | Article 989 | 5
Relative frequency of the item. This raises the question of what is pre-empting form may not be driven by the skew or may not have a high resting state, thus the selection of the selection threshold. The top pre-emptors in the skewed set may have a higher resting activation level. This then blocks or pre-empts the free choice of the item. This is evidence for the skew-preemption hypothesis outlined in the introduction. The cognitive means by which this is suggested to work is as follows. The frequency of competing forms. This is the frequency of use, with more frequently used words having a lower relative frequency of use.

2 "Construction" is used here to mean a form-meaning pairing, as in the Construction Grammar of Goldberg (1995). It should be noted that neither the analysis nor conclusions necessarily depend on this sense of the word construction; the findings are also consistent with framing the un-form as the more theoretically neutral concept "schema" or as a semantically-constrained lexical rule.
FIGURE 3 | The historical frequency data of un-forms and their associated top two competitors plus the histogram plots for all preemptors associated with a particular un-verb. Rows are organized into the same five groups as displayed in Figure 2—from best examples of un-forms at the top through to worst examples at the bottom.3
un-forms in the x-axis timeline rather than a sudden influx appeared over 40 times. Is likely that this explains the delayed appearance in the corpus. For the Google ngrams corpus any ngram needs to have they are less likely to reach the minimum threshold for inclusion in especially those that are poor fit with the cryptotype. Because of this the correct set over which to calculate frequency. If one assumes the pre-emptor frequency is calculated over all possible items then the top pre-emptor may or may not have a relative high resting potential. If one assumes however the frequency is calculated over a semantically relevant set of competitors then the top pre-emptor will always have a relatively higher resting activation if it is part of a skewed set. A relevant semantic set must be factored in at some point in processing as if pre-emptors were selected on the basis of raw frequency alone, then highly frequent (the most frequent) items would always be chosen regardless of whether they were semantically acceptable or not. As a reviewer pointed out, in flatter distributions, the best competitor receives more competition for selection from other competitors in their set. So this raises the possibility that it is not that items in a flatter distribution have lower resting activations but that competitors are more comparable in their activation level. In a skewed distribution, all of the poorer competitors are, by definition, poor competitors with the top pre-emptive form making it easier to select.

In the current study the effects of overall frequency and skew of pre-emptors are still somewhat confounded. Recall that Ambridge (2013) found a positive correlation between the raw frequency of the top pre-emptors and the acceptability of their associated un-forms. We find that there is collinear relationship with the skew of the set that these forms happen to part of. This is presumably something which could be experimentally disassociated—for example by controlling for overall frequency of the best competitor and varying skew (and vice versa), and measuring the effect on generalization a novel form.

In conclusion, we found evidence that the extent to which present-day raters consider an un-form as a good fit for the semantics of the un-form cryptotype predicts the historical frequency of that form. This indicates this construction’s stability and that the skew of its pre-emptors, $r_{485} = -0.0007$, $p = 0.88$. Better fit $r_{485} = 0.03$, $p = 0.509$ Average fit $r_{485} = 0.096$, $p = 0.034$ Poor fit $r_{485} = -0.006$, $p = 0.89$ Poorest fit $r_{485} = -0.27$, $p = 0.552$

This suggests that it is not just the raw frequency of the pre-emptive forms that must be taken into account when trying to explain usage patterns but the shape of the distribution as well.

ACKNOWLEDGMENTS

Thank you to Ben Ambridge who provided insightful comments on these ideas and was very helpful in sharing data; to the three reviewers of this paper who made excellent suggestions to improve the paper; and Adele Goldberg who provided helpful comments on earlier drafts of this paper.

REFERENCES


Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 August 2013; accepted: 11 December 2013; published online: 25 December 2013.


This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.

Copyright © 2013 Ibbotson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.
APPENDIX

**FIGURE A1** | Historical trends of un-forms ($N = 828,813$) as a function of semantic rating, raw frequency.

**FIGURE A2** | Historical trends of un-form items.
FIGURE A3 | Historical trends of un-forms ($N = 828,813$) item analysis on raw frequency.