Revisions for A case for systematic sound symbolism in pragmatics

We thank the reviewers for their excellent reviews, and respond to their points below.

The major revision in this paper is that we re-collected the word data and re-ran all tests. This is because we discovered that the WOLD and IDS databases released new versions between the start of our investigation and this review. The versions differed quite substantially for some languages, and added many more new languages (increasing the number of languages from 172 to 226), so we decided to re-run everything.

The results are qualitatively the same, except for the comparison between initial and non-initial languages (test 4). Previously, there was a significant effect when controlling for language family, but now that significance has disappeared. This does not impact the general message of the paper. Indeed, reviewer 2 suggested that we downplay this result in light of some related tests not working out, and agreed that the paper was sound and valuable even without this result. In fact, in the original submission we do not include this result in the final conclusion, so the new results actually cause few changes to the main body of the text.

Below we respond to the individual points of the reviewers.

Reviewers comments appear in this font

**Our responses appear in this font**

-\*Reviewer 1\*  
  
 - This paper is excellently written and clearly argued. Especially the   
statistical analyses are very clearly explained. The conclusions are a bit too   
positive for my taste, in that the medium statistical effects are a bit   
overblown (see running commentary below). I still think that the paper is worthy   
of publication with just a few minor corrections.  
  
Running Commentary  
  
p.10, midpage: “homophones”: probably better “polysemes”, but that is nitpicking

**We have changed this to “polysemous words”.**

p.14, first paragraph: another problem, not mentioned here, is that in the   
wordlist data the same word can occur multiple times in case of polysemy. This   
is surprisingly common, e.g.<http://ids.clld.org/contributions/288> HOW=WHAT or   
<http://ids.clld.org/contributions/272> WHO=WHICH. Such multiple entries of course   
inflate any effects of identical first sounds. I have no intuition whether it is   
common enough to influence the analyses in the paper, but it seems frequent   
enough to warrant checking.

**We thought about this problem, but it’s quite difficult to address. In an extreme reading, our hypothesis would predict that a language would have identical forms for all wh-words. Therefore, removing the duplicates removes part of the effect we are trying to detect, which seems a little unfair to us.**

**Also, we made the following assumptions about the data: Empty cells indicate that the language has no lexicalised form for the concept. Duplicated forms mean that they use the same form for both concepts. So, if the language really only had one *concept* for ‘how’ and ‘what’, then it would receive only one entry. Duplicate entries suggest that speakers have separate concepts, but identical forms. In this case, we think it’s fair to count them as separate entries.**

**Actually implementing this check is also difficult. A lot of time was put into cleaning and simplifying the representations of words, often on a language-by-language basis. Forms that look identical in the final data may actually be phonemically different in their raw form. It is also not clear, if a language has a duplicate form, which concept to exclude.**

**Still, we created a new dataset with duplicates within languages removed (conservatively based on the cleaned forms) and performed some of the main tests. These can be found in the supporting information (S3). The tests did not differ greatly from the original results. For example, the original mean entropy score for wh-words was 0.43, and when removing duplicates this increases to 0.46. This is still well below the same entropy measure for pronouns (0.64, duplicates in pronouns were not removed).**

p.14, considering using initial segments: note that there are are various   
languages in which the typical question-word-repitition seems to occur at the   
end, e.g.<http://ids.clld.org/contributions/288> final “rõ” or   
<http://ids.clld.org/contributions/401> final “aj.1”. This might strengthen the   
claims made in the paper

**We have added these two references into the discussion (section 9)**  
  
p. 24, “some examples”: there is only one example given.

**replaced “some examples” with “an example below”**  
  
p. 31, line 4: maybe insert “also” to make “study 1 and 2 also held”

**Added “also”**  
  
p. 31, “This contrasts with previous assumptions that formal regularities within   
question words are not present (Cysouw 2004)”: This reference to my 2004-paper   
is not correct. I did not say that there are no regularities, but (as quoted in   
the present paper on p.8) that they are “not nearly as universal as often   
thought”. I don’t think that the results of the current paper contradict that   
Statement.

**We have changed this to read “Although previous analyses suggested that there are no strong universal patterns in wh-words (Cysouw, 2004), the present”.**  
  
p. 32, start of last paragraph: “perhaps the most surprising find…” This   
conclusion is not acceptable as it is formulated right now. This is strongly   
overreaching the results of the paper. The result is basically not significant,   
and in one case just significant, so it cannot be said to be the most surprising   
result of the paper.

**This is a fair interpretation. We have adjusted the paragraph to read:**

**“Our strongest prediction, derived from theories of interaction, was that question words would be more similar in languages that placed them at the beginning of turns. The results of these tests were mixed. Although the trend was in the predicted direction, the test of initial segments stratified by language family was only just significant (p = 0.047; p = 0.01 for just first consonants, see S3), the result was not significant when controlling for geographic area. Both tests were significant when using only unanalyzable words, but this was only run on the 20 languages for which the data was available. Furthermore, body concepts showed the same pattern, which was not predicted. Therefore, the trend is tantalising, but ultimately not robustly supported by this study. Even the general trend may be explained by other mechanisms.” [... continues to explain the confound with word order ]**

**We believe the result is fairly characterised elsewhere in the text.**

-\*Reviewer 2\*  
  
 -  
  
Review of A case for systematic sound symbolism in pragmatics: Universals in   
wh-words (PRAGMA\_2016\_432)  
  
GENERAL COMMENTS  
This is a nice paper, presenting a hypothesis well grounded in pragmatic theory   
and observations, checking it using appropriate statistical techniques, and   
presenting finding in support of the hypothesis (but also, healthily, little   
wrinkles where not everything is totally confirmed). In the comments below I mix   
smal and larger things. The latter are few in number. In general I don't see   
causes for a major revision, a minor one will do.  
  
SPECIFIC COMMENTS  
p. 8: "we note that there is no relation between degree of resonance and   
cognitive complexity". This requires a bit more explanation. Is the observation   
due to the authors or to Mackenzie? How can (eidemic) resonance (which is   
illustrated by sl- in words like slime, slippery, etc.) be measured in degrees?   
How would a relation between degree of resonancy and cognitive complexity   
manisfest itself? If all this is not so important an option is to leave out the   
Statement.

**This was supposed to draw the reader’s attention to one of Mackenzie’s results, not our own interpretation, but the reviewer is right that this is ambiguous. We have removed this part of the sentence. (Regarding the degrees of resonance, we suppose that resonance may be present in more or less words in a set)**

Tables 1 & 2: Only a generic reference is given to Haspelmath and Tadmor (2009),   
which is not fair to the authors who compiled the vocabularies used. There   
should be individual references for each language.

**We have added references for each language.**

p. 14: Spraakbanken -> Språkbanken

**Corrected**

p. 14: About 7% of possible entries had no data -> About 7% of possible entries   
were empty

**Corrected**

p. 15: The link<https://github.com/seannyD/UniversalsInWHWords> is not active yet.

**The link will be active after acceptance.**

p. 16: The left side of the equal sign in the equation falls below the right   
side. Same for an equation later on. While the f in Ef is written with a   
subscript in the equation it is not in the text.

**We have corrected the equations and Ef is corrected in text to f in subscript.**

Fig. 3, caption: (left, bottom) -> (right, bottom)  
p. 17: within Indo-European language family -> within the Indo-European language   
family. Also, give a page number for the Cysouw 2004 statement.

**Corrected**

Section 3.3: From Fig. 3 one gets the idea that words are swapped between pairs   
of languages. But I suspect that the illustration shows a simplification of the   
procedure. How is the swapping concretely done? Given a set of languages belong   
to a group (family, area or whatever), I imagine an algorithm something like   
this: each question word meaning is a vector and the positions in that vector   
are words in each language. The vector indices are now randomized, presumably   
using R's sample function, sampling as many indices as there are in the vector.   
Presumably sampling without replacement is used. This is done for each vector   
and then the vectors are put together through rbind and the entropy is   
calculated for each column. This is repeated 10,000 times. This is how I guess   
it is done, but I could be wrong, so the authors need to be a bit more explicit.   
They also need to state how empty cells are handled, this is often a thorny   
problem in these kinds of exercises.

**Yes, this is essentially correct. We have added an extra few sentences explaining how the premutation worked. We have also included all R scripts which explicitly show how the permutation was done.**

p. 19, text below Table 3: it is found here that words belonging to some other   
domains are also more similar within than across languages. This is attributed   
to differences in phonologies between languages. But there could be real effects   
of sound-meaning associations here. See Dautriche, I., Mahowald, K., Gibson, E.   
and Piantadosi, S. T. (2016), Wordform Similarity Increases With Semantic   
Similarity: An Analysis of 100 Languages. Cogn Sci. doi:10.1111/cogs.12453.

**We have included this citation, as well as Blasi et all, 2016 which shows a similar effect.**

Table 4: THe p-value written "< 0.000005" could just as well be written "< 0.00001"

**Corrected**

p. 29: with a random intercepts -> with random intercepts

**corrected**

p. 30: When considering at all segments -> When considering all segments (?)

**corrected**

p. 30: "when considering at first consonants (mean difference = 0.15, z = 2.37, p = 0.01) but this was marginal when considering first vowels". I don't   
understand what is meant by "at first" and "first". My guess is that the authors   
would be looking for a word like "separately".

**Corrected**

p. 31: First, there is undeniable impact of the contact -> First, there is an   
undeniable impact of contact

**Corrected**

p. 33: "The derivation of forms from common ancestors is likely a prominent   
mechanism by which question words become similar, but we argue that the ultimate   
reason that it applies for question words is for action ascription." This is not   
an argument but a postulate, because there is not any check for the impact of   
shared morphology. Earlier on in the paper this problem was also swept under the   
rug (p. 11): "A second confound is that a set of question words within a   
language often derive from a common ancestor word, making them similar by   
descent. However, the fact that the question words that had undergone changes   
have maintained the root at the beginning of the words fits well with the   
proposed hypothesis. In the jargon of evolutionary theory (Mayr, 1961;   
Scott-Phillips, Dickins & West, 2011), common derivation would be a proximate   
mechanism by which question words come to be similar, but the ultimate reason   
that this mechanism applies is because of a pressure from turn taking." It has   
never, to my knowledge, been shown that words retain their roots because of   
selective pressure. Phonological change is still overwhelmingly best explained   
by regularities that are independent of meaning as per the comparative method.   
Where sound-meaning associations arise, in my view, is when forms are created.   
As far as we know subsequent changes are not influenced by the kinds of   
cognitive forces or selective pressures that give rise to submorphemic   
recurrency. The possibility that shared morphology (whether fossilized or simply   
not recognized by the authors for lack of expertise) may contribute a great deal   
to the findings is a weak point of the paper, and this should be recognized by   
making it a third point which is given the same prominence as the first one   
about contact and the second one about diverse phonologies.

**The reviewer argues that there is no evidence that morphological change is affected by selective pressures. Our aim here was to present the confound that words are derived from common forms. We now make it clearer what this confound is, using an example and link it to the tests of other word sets. We also separate it from the argument about proximate and ultimate research questions.**

**We have changed the latter section to this:**

**“Words within a language may also be similar due to historical processes. For example, in English the interrogatives *when, where, which, who* and *why* are all derived from a Proto-Indo-European interrogative pronoun stem *\*kwo-* (Harper, 2016a). This process of derivation therefore preserves a similarity between forms (compared to, for example, if they were all derived from different sources). This is difficult to address directly because the common elements may be fossilised beyond straightforward recognition and there is little cross-linguistic data on common derivations (we note that the DiACL database is a promising source, Carling, 2017). Our solution is to compare the results for interrogative words with results for other sets of related words, such as pronouns, which also often have common derivations (Bickel & Nichols, 2007: 209). If similarities in interrogatives are purely driven by a general process of common derivation, then, everything else being equal, they should show similar measures to pronouns.**

**It is important to note that common derivations represent a confound for our analysis, but not actually an opposing explanation. In the jargon of evolutionary theory (Mayr, 1961; Scott-Phillips, Dickins & West, 2011), common derivation is a *proximate mechanism* which preserves similarity while adaptation to turn taking would be an *ultimate reason* for similarity. That is, we don’t make a claim about *how* interrogative words come to be similar, we are interested in *why* they are similar. It is possible that selective pressures cause the mechanisms of historical change to apply differently to different parts of the lexicon (e.g. pressures from turn taking preserve similarities in interrogatives, so common derivation is one of the mechanisms that brings about the ultimate adaptation to turn-taking), though we know of no evidence for this.”**

**(As a side note, some of our other work suggests that sound-meaning associations emerge over time rather than at the moment of formation.)**

p. 33: Whichmann & Holman, 2016 -> Wichmann et al., 2016 [funny typo in the   
context]. In the References this could be supplied with an URL:   
<http://asjp.clld.org/>.

**Corrected, URL added**

p. 33: South-America -> South America

**Corrected**

p. 33: similarities of also -> similarities also of

**Corrected**

Table 6 should give the source and the variants in question should be indicated   
in the caption. Quechua and Aymara are whole language families.

**We realized table 6 was not necessary, so we have moved the examples into the main text, referring to specific languages and citations as necessary.**

References:

Berlin: 76 –93 -> 76–93 (close up space)

**Corrected**

Discrepancies in the use of capitalization in book titles

Blasi et al are in the references but doesn't seem to be mentioned in the text,   
although it is relevant. Check for other such extraneous or lacking references.

**This publication now appears in the text.**

Blevins: place of publication lacking.

**Corrected**

I stop these observations at Blevins, the authors should be professional enough   
to make a the list of references consistent by themselves.

**Corrected**