

## Manuscript Details

<b>Manuscript number</b>	PRAGMA_2016_432
<b>Title</b>	A case for systematic sound symbolism in pragmatics: Universals in wh-words
<b>Article type</b>	Full length article

### Abstract

This study investigates whether there is a universal tendency for content interrogative words (wh-words) within a language to sound similar in order to facilitate pragmatic inference in conversation. Gaps between turns in conversation are very short, meaning that listeners must begin planning their turn as soon as possible. While previous research has shown that paralinguistic features such as prosody and eye gaze provide cues to the pragmatic function of upcoming turns, we hypothesise that a systematic phonetic cue that marks interrogative words would also help early recognition of questions (allowing early preparation of answers), for instance wh-words sounding similar within a language. We analyzed 226 languages from 66 different language families by means of permutation tests. We found that initial segments of wh-words were more similar within a language than between languages, also when controlling for language family, geographic area (stratified permutation) and analyzability (compound phrases excluded). Random samples tests revealed that initial segments of wh-words were more similar than initial segments of randomly selected word sets and conceptually related word sets (e.g., body parts, actions, pronouns). Finally, we hypothesized that this cue would be more useful at the beginning of a turn, so the similarity of the initial segment of wh-words should be greater in languages that place them at the beginning of a clause. We gathered typological data on 110 languages, and found the predicted trend, although statistical significance was not attained. While there may be several mechanisms that bring about this pattern (e.g., common derivation), we suggest that the ultimate explanation of the similarity of interrogative words is to facilitate early speech-act recognition. Importantly, this hypothesis can be tested empirically, and the current results provide a sound basis for future experimental tests.

<b>Keywords</b>	interrogatives; wh-words; turn taking; cross-linguistic
<b>Corresponding Author</b>	Sean Roberts
<b>Corresponding Author's Institution</b>	Max Planck Institute for Psycholinguistics
<b>Order of Authors</b>	Anita Slonimska, Sean Roberts
<b>Suggested reviewers</b>	JP de Ruiter, Michael Cysouw, Terry Regier
<b>Opposed reviewers</b>	Stephen Levinson, Damian Blasi, Marcus Perlman, Fiona Jordan, Bodo Winter

## Submission Files Included in this PDF

### File Name [File Type]

RevisionsforAcaseforsystematicsoundsymbolisminpragmatics.docx [Response to Reviewers]

Pragmatics\_Revision1.docx [Manuscript File]

Vitae.docx [Author Biography]

Highlights.docx [Highlights]

Title page.docx [Title Page (with Author Details)]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

## 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-repetition seems to occur at the end, e.g. <http://ids.cld.org/contributions/288> final "rõ" or <http://ids.cld.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 small 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 manifest 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 permutation 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 al, 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: Wichmann & Holman, 2016 -> Wichmann et al., 2016 [funny typo in the context]. In the References this could be supplied with an URL: <http://asjp.cld.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**

## **A case for systematic sound symbolism in pragmatics:**

### **Universals in *wh*-words**

**Abstract.** This study investigates whether there is a universal tendency for content interrogative words (*wh*-words) within a language to sound similar in order to facilitate pragmatic inference in conversation. Gaps between turns in conversation are very short, meaning that listeners must begin planning their turn as soon as possible. While previous research has shown that paralinguistic features such as prosody and eye gaze provide cues to the pragmatic function of upcoming turns, we hypothesise that a systematic phonetic cue that marks interrogative words would also help early recognition of questions (allowing early preparation of answers), for instance *wh*-words sounding similar within a language. We analyzed 226 languages from 66 different language families by means of permutation tests. We found that initial segments of *wh*-words were more similar within a language than between languages, also when controlling for language family, geographic area (stratified permutation) and analyzability (compound phrases excluded). Random samples tests revealed that initial segments of *wh*-words were more similar than initial segments of randomly selected word sets and conceptually related word sets (e.g., body parts, actions, pronouns). Finally, we hypothesized that this cue would be more useful at the beginning of a turn, so the similarity of the initial segment of *wh*-words should be greater in languages that place them at the beginning of a clause. We gathered typological data on 110 languages, and found the predicted trend, although statistical significance was not attained. While there may be several mechanisms that bring about this pattern (e.g., common derivation), we suggest that the ultimate explanation of the similarity of interrogative words is to facilitate early speech-act recognition. Importantly, this hypothesis can be tested empirically, and the current results provide a sound basis for future experimental tests.

Keywords: interrogatives; *wh*-words; turn taking; cross-linguistic

## **1. Introduction**

One of the key insights of an evolutionary approach to language variation and change is that different linguistic structures may be more or less effective at fulfilling a particular function, and that this effectiveness influences how likely a given structure is to be used (e.g. Croft, 2000). That is, just as for biological species, the most effective linguistic structures are selected for reproduction, while the less effective ones fall out of use, leading to cultural evolution. The end product is that languages should appear to be adapted to their cultural ecology. When looking at biological species it is often easy to identify the ecology to which individuals must adapt. Deserts apply a selective pressure for water retention, cold climates apply a selective pressure for heat retention and so on. However, when looking at language, identifying the primary ecology - the most important constraints - is more difficult.

Many studies have shown that languages and linguistic structures are adapted to many different functions and domains. For example: the brain is an ecology that exerts a pressure for effective storage and processing, and studies have shown that word order rules often align to make processing more effective (Hawkins, 1994; Ferrer-i Cancho, 2008); a pressure for effective communication can lead to frequent words being short, serving efficient production (Zipf, 1949) or to dispersed phoneme inventories which maximise intelligibility (de Boer, 2000); the physical constraints of articulation and perception can influence phonological rules or changes (Blevins, 2004), or even the fundamental inventory of phonemes (Moisik & Dediu, 2016); languages also need to be repeatedly learned and transmitted, which can lead to the emergence of compositionality (Kirby, Cornish & Smith, 2009).

One often neglected domain when trying to explain the cultural evolution of language is pragmatics, and in particular interactive conversation. This is surprising since conversation is an indispensable part of human life. It enables us to exist in society, express ourselves, expand our knowledge, influence others, and attain our goals. Conversation is the most frequent use of language and provides the raw data for language learning (Levinson, 2006). It has been estimated that on average humans spend 2-3 hours a day speaking and producing up to 1200 turns (Levinson, 2016). Therefore, just as languages are shaped by cognitive demands on processing or physical demands on articulation, the constraints of conversation should also affect the cultural evolution of language. That is, we argue that conversation is the primary ecology of language (Levinson, 2006), and we should expect languages to show signs of adaptation for conversation (see also Micklos, 2014; Roberts & Mills, 2016).

Indeed, there is a growing body of literature offering evidence for such adaptations. For example, the repair-initiating word “huh?” is ubiquitous in the world’s languages, with its form being well adapted to be used as a salient, rapid interjection (Dingemanse, Torreira & Enfield, 2013). There also appears to be a universal set of interaction sequences which support social actions (Kendrick et al., 2014). [Another study by the authors XXXX] links pressures from the timing of turn taking to the emergence of basic word order patterns. Studies using experimental semiotic paradigms such as iterated learning also demonstrate the role that interaction has in shaping fundamental properties of language such as systematicity (Tamariz et al., 2012; Macuch Silva & Roberts, 2016), iconicity (Verhoef, Roberts & Dingemanse, 2015; [Another study by the authors, in prep]) and predictable variation (Feher et al., 2016).

One of the key differences between domains like cognition or processing and conversation is interaction. In conversation, multiple interlocutors produce turns at talk in contingent sequences in real time, with the content and function of one turn relying on the previous turns (Sacks, Schegloff & Jefferson, 1974). Considering that the sequences are not entirely predictable, turns are exchanged between interlocutors with very precise timing. For example, answers are produced around 200ms after the end of a polar question (Stivers et al., 2009), considerably quicker than the average time to plan and produce even one word (600ms, Levelt, Roelofs & Meyer, 1999). This implies there is a

point at which a listener is trying to comprehend what their interlocutor is saying at the same time as they are planning their response. This cognitive burden would not be present except for the pragmatic norms for responding quickly in real-time conversation. That is, the constraints of turn taking create a harsh ecology in which linguistic structures must be effective in order to “survive” and be reproduced at a later stage.

Perhaps the context that puts greatest strain on processing, and therefore the context where we should expect to find adaptation, is answering content questions. This involves understanding what information the questioner is asking for and also retrieving the answer from a potentially massive number of options, all while the usual norms of the timing of turn-taking apply. Any clue to help the answerer respond quickly would be advantageous.

The present study looks for systematic cues that interlocutors can use to predict whether a turn is a content question, so called *action ascription*. There are many studies which demonstrate the use of paralinguistic cues such as prosody or eye gaze for action ascription (see section 2), and of course there are clear semantic and structural aspects to questions. More generally, there are phonological and prosodic cues to major syntactic classes, which is hypothesised to help acquisition (Cassidy & Kelly, 1991; Berlin 1994; Monaghan, Christiansen & Chater., 2007). Here, we investigate whether languages exhibit systematic phonetic cues to aid conversational turn taking. Specifically, whether interrogative words (e.g. *what, when, where, which, who, why* in English) sound similar within a language to provide a low-level cue for questionhood. That is, speakers of English can use the [w] sound as a cue that a content question is about to be asked. Similarly, in Latvian one can listen for a [k] (*kas, kad, kur, kurš, kas, kāpēc*).

The similarity of interrogative words is hardly news to linguists - indeed, they are often referred to as “wh-words”, reflecting the tendency of many to start with “wh” in English. Furthermore, many interrogative words often have common derivations. However, this is just the proximate mechanism by which they come to be similar. We hypothesise that the ultimate reason that they do sound similar is to aid action ascription. To be clear, we are not proposing a universal iconic link between the “wh” sound and interrogation - it is quite clear that the “wh” pattern is not a universal across languages. Rather, we try to detect a statistical tendency for interrogative words to sound similar to one another within languages. More specifically, we predict that interrogative words will be (1) more similar within languages than between languages; (2) more similar within languages than a random selection of words or conceptually related sets of words from those languages (e.g. pronouns); and (3) be composed of sounds that are particularly salient or detectable within a language. Furthermore, since cues to action ascription are most useful the earlier they appear in the turn, we predict that (4) interrogative words will be more similar in languages which place them at the start of clauses. The last prediction is particularly important, since it attempts to explain differences between languages due to the interaction between the constraints of conversation and the structure of language, not just universal tendencies in all languages (see Lupyan & Dale, 2016).

The paper is organised as follows. First we review the literature on turn taking in conversation, cues to action ascription and the form of interrogative words. We then conduct four quantitative studies on a worldwide sample of data to address each of the predictions made above. We end with a discussion of the implications of our findings and directions for future research.

## **2. Background**

### **2.1 Turn-taking**

Conversation happens by exchanging turns in sequences. One speaker produces a speech act to which the next speaker responds with another (preferably) appropriate speech act, to which the first speaker responds and so on. It would be difficult to imagine how feasible communication would be if people performed their speech acts simultaneously, since conversation is usually contingent - the pragmatic action that one speaker performs depends on the previous pragmatic action.

Although conversations may involve periods of talk by two people simultaneously, or periods of silence, on the whole interlocutors strive to minimise gaps and overlaps (Sacks, Schegloff & Jefferson, 1974; Levinson, 2016). The timing of turn transition is very precise. The most frequent gap between a polar question and a (yes/no) answer in many languages is roughly 200ms (Stivers et al., 2009), which is much shorter than the time it takes to retrieve, plan and begin producing a single word (600ms, see Levelt, Roelofs & Meyer, 1999). This implies that for the second speaker an overlap of comprehension and production must occur (Levinson & Torreira, 2015, see figure 1).

This precise timing is mandated because of the role of conversation in social action. In conversations with many people, the opportunity to take the floor rapidly disappears. Furthermore, delayed turns can be interpreted as unwillingness to respond, especially in transitions between questions and answers (Roberts & Francis, 2013; Kendrick & Torreira, 2015). Even young infants exhibit sensitivity to unusual timing in turn-taking (Casillas, 2014; Stephens & Matthews, 2014). While the main content of turns can be delayed by using turn-preserving placeholders (e.g. hesitations, um, er, Clark & Fox Tree, 2002; Strömbergsson et al., 2013), this can only mediate the process to some extent.

Such rapid reactions are, of course, possible because speakers recognise speech acts and predict the content and timing of turns before the previous speaker has finished speaking (Levinson, 2013). Apart from the usual processing of semantic and structural aspects of turns, several studies have demonstrated that speakers use lower-level information as cues, which we cover in the next section.

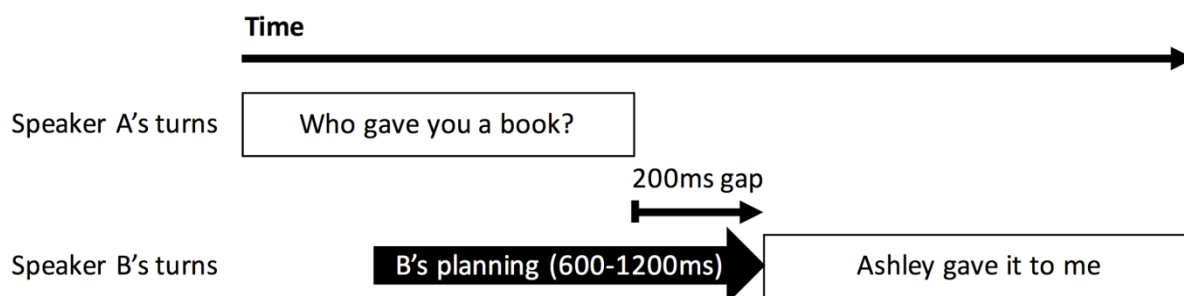


Figure 1. Overlap of comprehension and production in conversation (Levinson 2013, p.104). The typical gap between two turns is of the order of 200 milliseconds, but production planning takes at least 600 milliseconds, meaning that B must start planning their turn in the middle of comprehending the previous turn.

### 2.3 Action ascription

Turn-taking in conversation is not merely well timed, but built on contingent sequences. Namely, some speech acts require very particular responses. For example, a greeting normatively requires a greeting in return, and a question makes it relevant to provide an answer. Two turns, where the following turn is normatively dependent on the previous turn are called *adjacency pairs* (Sacks, Schegloff & Jefferson 1974, Schegloff 2007). Providing that adjacency pairs are based on regularities, it becomes easier to predict what would be the next most appropriate turn as a response. It is possible that humans make use of this characteristic of adjacency pairs in order to ascribe the speech act and to start planning the answer in advance (Roberts, Torreira & Levinson, 2015). For example, Gisladdottir, Chwilla, & Levinson (2015) show that listeners recognize speech acts before the end of the sequence if they are in a highly constraining context, namely if the context constitutes the initial turn of adjacency pair like a question for an answer or offer for a declination. On the other hand, given that pre-offers are less predictable, they require additional processing and listeners make use of the entire utterance. Based on these findings Gisladdottir, Chwilla, & Levinson (2015) conclude that previously available context allows next speaker to project the action-underspecified turn that has not yet finished and start planning the response.

While intuitively it seems less surprising that humans are capable of greeting each other in a timing-wise fluent manner due to the social context, it is extraordinary that the same fluency is achieved with answers to questions. Greetings, indeed, require a very particular *responding action*, and are limited to a limited set of possible responses (e.g. *Hi, hey, hello, good morning*, etc. for English speakers). Similarly, polar questions require a response from a closed class (yes/no). Content questions, in contrast, require the provision of new information, which involves both the comprehension of the question, retrieval of the answer and the planning of a possibly complex response that fits the pragmatic intentions of the answerer. Indeed, production planning starts as

soon as the information for an answer can be retrieved (Bogels, Magyari & Levinson, 2015; Barthel, Meyer & Levinson, 2017). Therefore, content questions followed by answers can be seen as one of the harshest ecologies for language in conversation. In addition, since questions and answers are very frequent (Levinson, 2013, p.112) and represent the prototypical adjacency pair, if languages adapt to the constraints of conversation, it is here that we might expect the greatest amount of adaptation.

Indeed, there are a number of studies which demonstrate a range of cues which help interlocutors recognize questions. Intonation can play a prominent role in action ascription. Rising intonation is a cue for questions in many language, but cues also exist at the start of the turn. Sicoli et al. (2014) argue that initial pitch functions as phonetic cue for ascribing social action type of questions. They show that people tend to use higher initial pitch for questions that have an evaluative action (i.e., indirect speech act) rather than a request for information (i.e., direct speech act). Thus, they argue that deviation from average pitch at the beginning of questions helps an addressee recognize that a question is not to be perceived directly. Similarly, eye gaze is used both as a cue to questions (Rossano, Brown & Levinson, 2009) and as a tool for the management of turn timing such as holding the floor or giving a cue to turn boundaries (Rossano 2013).

Overall, there is increasing evidence that speakers take advantage of front-loading of cues in order to facilitate early question recognition, although most of the previous research has concentrated mainly on paralinguistic cues. Without doubt, action ascription seems to be achieved by means of an interplay of auditory and visual communicative tools at speaker's disposal. Surprisingly, however, there is little work on linguistic cues and what their systematicity might contribute to question recognition.

## **2.4 Interrogative words**

Many languages exhibit lexical or morphological cues for recognising questions, for example question particles and interrogative morphology. Question particles in particular offer a clear cue for questionhood, but it is unclear whether these evolved for *rapid* action ascription, and whether they have a wider effect on the language (though see Thompson, 1998 and [another study by the authors]). Another clear candidate for cues are question words (table 1). Content question words (also called interrogative words or wh-words) target a specific piece of information. For example in the sentence “who gave you a book?”, *who* targets information about a person, while in “what did they give you?” *what* targets an (inanimate) object. The distinction between human and non-human question words is very common in the world's languages (Ultan 1978, Lindström 1995), though many also have dedicated forms targeting other categories. Cysouw (2004) identifies 4 major types - person (who), thing (what), selection (which), and place (where), 3 minor types - quantity (how much), manner (how) and time (when) and various less frequent incidental types, including reason (why) and quantity (how much/ how many).



Many languages have question words that are at least partially transparent and analyzable. For example, the French phrase “pourquoi” (targeting a reason) is derived from the word “quoi” (targeting a thing). This pattern of derivation is common (in fact, the English word ‘why’ is a rare example of an unanalyzable form targeting reason, Cysouw, 2004), and many other question words are derived from other question words. For example, the question word for manner often (synchronically) derives from the question word for thing (e.g. Everett 1986: 239-245, Foley 1991: 114-115; see Cysouw, 2004). Diachronic derivation is also common, such as many question words in English deriving from a single form in Proto-Germanic *\*hwa* (see Harper, 2016). Mackenzie (2009) suggests that there is a semantic hierarchy of complexity in question words, increasing in cognitive complexity for person, location, time, manner and quantity. Furthermore, there is an iconic link between cognitive complexity and form complexity.

Many languages obligatorily place the interrogative words at the beginning of clauses. This “front-loading” could provide a cue for rapid question ascription (Levinson, 2013). Even in languages where formal grammar rules do not require this, often the colloquial variety will place interrogative words at the beginning of turns (e.g. in Japanese, Levinson 2013, p.112, though some claim that interrogative phrases in content questions are avoided in Japanese, see Hinds, 1986, 32). Placing easily recognisable words at the beginning of a turn would provide an optimal cue, especially if the words shared some clear phonetic similarity. For example, the majority of interrogative words in English share the same initial phoneme - /w/. We note that this is also a visually salient phoneme, due to lip rounding. Similar regularities can be observed in many other languages, though Cysouw suggests that these are “not nearly as universal as often thought” (Cysouw, 2004, p.3). Indeed, it is clear from table 1 that languages span the range of possible diversity in initial segments of question words.

Mackenzie (2009) also notes systematic similarities in question words in many languages, suggesting that they are a form of submorphemic relation (Lehman, 1993) which show ‘eidemic resonance’, the same phenomena as the sound symbolism in word sets like *slime*, *slippery*, *slither*, *slug* etc. and links this to the suggestion by Bickel & Nichols (2007: 209) that these similarities could be used as “psycholinguistic cues”. In an analysis of 50 languages, Mackenzie finds resonance in the question words of 33 languages, though most cases only cover a minority of the forms within a language. Mackenzie’s study investigates the cognitive complexity of the question word semantics, which we do not explore here.

Wh-word (English)	Hindi Indo-European	Telugu Dravidian	Yaqui Uto-Aztecan	Bulgarian Indo-European	Dehong Tai-Kadai	Vietnamese Austroasiatic
how	k <sup>h</sup> se	<i>elaa</i>	<i>jachinia</i>	<i>kák</i>	<i>com.2 sə.2</i>	<i>sao</i>
how many	kit <sub>n</sub> an <sub>n</sub> a: ; kit <sub>n</sub> an <sub>n</sub> e	<i>enni</i>	<i>jaikim</i>	<i>kólko</i>	<i>xo.1</i>	<i>máy</i>
how much	kit <sub>n</sub> an <sub>n</sub> a:	<i>enta</i>	<i>jaiki</i>		<i>la.3 lai.6;</i> <i>jom.4 lai.6</i>	<i>bao nhiêu</i>
what	kja:	<i>eem;eemi[Ti]</i>	<i>jita</i>	<i>kakvó; štó</i>	<i>ʔen.3</i>	<i>gì</i>
when	kab	<i>eppuDu</i>	<i>jakko</i>	<i>kogá</i>	<i>hak.8; kek.8</i>	<i>khi nào</i>
where	kaha:n <sub>n</sub>	<i>eTa; eedi;</i> <i>ekkaDa</i>	<i>jaksa</i>	<i>kədə</i>	<i>cup.7</i>	<i>đâu</i>
which	kɔn <sub>n</sub> sa:	<i>eevi</i>	<i>jita</i>	<i>kój</i>	<i>hɔp.9 ʔum.3</i>	<i>nào</i>
who	kɔn <sub>n</sub>	<i>ewaru</i>	<i>jabesa</i>		<i>xʔn.3 cep.9</i>	<i>ai</i>
why	kjo:n <sub>n</sub>	<i>en[du]ceeta;</i> <i>enduku</i>	<i>jaisakai</i>	<i>zaštó</i>	<i>mai.3 caʔ.6</i>	<i>tại sao</i>
E <sub>f</sub> initial segment	0.0	0.0	0.0	0.28	0.84	1.0

Table 1. Examples of question words in different languages. Question words in Hindi, Telugu, Yaqui and Bulgarian show systematic similarities, although the patterns are not universal across languages. In contrast, Dehong and Vietnamese show no systematic similarities in the initial segments of its question words. The final row shows the entropy efficiency of the initial segment of each word within languages (see section 3.2). Low values indicates consistency and high values indicate inconsistency. Data from the World Loanword Database and Intercontinental Dictionary Series (Haspelmath & Tadmor, 2009; Key & Comrie, 2015; Hindi: Saxena, 2015; Telugu: Saxena, 2015;

*Bulgarian: Vakareliyska & Horissian, 2015; Yaqui: Estrada-Fernández, 2015; Dehong: Peiros, 2015; Vietnamese: Alves, 2009).*

While it's clear that some form of similarity in question words is common, we know of no systematic, quantitative study which investigates a statistical bias for systematicity in form for question words, and in particular with a hypothesis motivated by the needs of rapid turn taking in conversation. Therefore, we proceed by implementing quantitative tests on a large set of languages from different parts of the world. The aim is to explore whether front-loading applies to the *wh-words* themselves. Namely, whether the first segments of a word tend to match within the set of question words and whether this occurrence is present across languages above chance.

## **2.5 Cultural evolution**

In this section we formalise a theory of the cultural evolution of question words under a pressures from turn taking in conversation. Croft (2000) suggests that words and phrases evolve according to Darwinian evolution. In every turn produced by a speaker, they must select words and phrases from a set of possible alternatives. From turn to turn, these elements replicate and appear again. In order to survive through time and from generation to generation, elements must replicate at a certain frequency, creating potential competition. Elements that are more successful in replicating have higher fitness. In cases where a certain pressure promotes the replication of one element over another, for example shorter forms being more efficient to produce, we can talk about *selection*. For example, when recognising words, the context will provide some constraint on possible interpretations, but it is beneficial for the listener if semantically similar words have distinct forms (arbitrariness, see Gasser, 2004). This should impose a pressure against semantically similar concepts having similar forms. Indeed, in current lexicons, polysemous words often belong to distinct contexts (e.g. a money bank and a river bank).

In contrast, a pressure for rapid action ascription could be facilitated by a phonetic cue to content questions, such as a systematic similarity in question words. If this benefitted rapid action ascription, then systematic similarities would be selected over non-systematic alternatives (or alternatively there would be less pressure for question words to diversify), leading to an increase in the systematicity of question words. Of course, other pressures and the current state of the language as a whole will affect how the precise systematic similarities are manifested.

It is worth comparing this hypothesis with another case of adaptation to conversation. Dingemanse, Torreira & Enfield (2013) showed that the word “huh?”, which is used to initiate repair, can be found in very many languages, and suggest that it is salient and quick to produce, which perfectly suits its purpose as an interjection to signal a problem in real-time. They suggest that this pattern arose due to convergent evolution (many languages arriving at the same solution

independently, as opposed to an ancient conserved word). In a similar way, we argue that question words in different languages might have undergone common selective pressures and changed to better serve effective conversation. We do not expect the same phonetic form to exist across all languages, since the constraints are weaker (they need to be salient for the listener but not quick to produce). Indeed, we don't expect to find the ideal pattern in all languages. However, we do expect that languages are likely to converge on the same kind of strategy to provide cues to action ascription, namely question words with front-loaded phonetic similarities. Another parallel with Dingemanse et al. is that they found that the exact pronunciation was tuned to the phonology of the language in which it was used (e.g. the vowel was appropriate for the phonology of the language). Of course, we expect the cues to respect the phonological rules of the language, but we also expect variation between languages according to whether the question word is front-loaded. However, in our case, front-loading increases the strength of the general selection pressure for salient cues, while in the case of "huh?" the particular phonology of the language changes the ideal target.

Note that many cultural evolution mechanisms identify an advantage to a single individual (speaker or listener, whose preferences are often presented as opposed), while in this case the benefit is to all participants in the conversation. This makes sense if we see conversation as fundamentally a cooperative activity (Hutchins, 2006; Dingemanse et al., 2015) where all participants have a preference for the conversation to progress (Stivers & Robinson, 2006). Indeed, many pressures can be seen as deriving from a general preference for progressivity, for example clear recognition of words avoids the need to spend time repairing misunderstandings.

## 2.6 Potential confounds

It is clear that there are many complications to this study, including differences between phonological inventories, the common derivation of many question words, compounds and analyzable forms. The first confounding factor is that phonological inventories of languages limit the amount of variation within a language. On average, it's likely that any set of words would look more similar within a language than between languages, simply because the phonological inventories differ. In order to address this we simplify the phonological representations of words in our sample, and also compare the results for question words with other sets of words (randomly sampled words, words from the same semantic domain and words within tightly related semantic domains). Another problem is that inheritance and borrowing between languages can inflate apparent cross-cultural patterns (Roberts & Winters, 2013). Indeed, many Germanic languages have a word for 'what' inherited from Proto Germanic (\**hwat*, compare with German *was*, Dutch *wat*, Danish *hvad*, Icelandic *hvað*, see Harper, 2016b). More generally, related languages may have similar phonotactic restrictions on word-initial segments by descent, meaning that they are not independent observations. We use stratified permutation and random independent sampling to control for historical and areal contact (see below).

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.

A third problem is that some languages have interrogative words that are composed of common sub-elements. For example in English “how many” and “how much” are used to ask about countable and uncountable quantities, but can be analysed as a phrase composed of two elements with independent meanings, and so the systematicity is due to the compounding. Similarly, Japanese has many analyzable question words (see table 2). Six words start with /d/, but 3 are analyzable as deriving from the same word. To ensure that results are not influenced by compound phrases, we run additional analyses using only unanalyzable words. This procedure also removes some words which are historically derived.

If question words themselves can serve as an indicator of the incoming speech act, it would be plausible to assume that matching phonemic onset could trigger the addressee to detect a possibility of incoming question. Before assessing the plausibility of such pragmatic benefits, first we have to assess whether any systematicity within interrogative words can be detected, and whether this is independent of historical contact between languages.

Meaning	Word	Gloss	Analyzability
how many?	ikutsu	iku-tsu <i>some-CLASS</i>	analyzable derived

how much?	ikura	iku-ra <i>some-PL</i>	analyzable derived
how?	dō		unanalyzable
what?	nani		unanalyzable
when?	itsu		unanalyzable
when?	nanji	nan-ji <i>what-hour</i>	analyzable derived
where?	doko	do-ko <i>Q-place</i>	semi-analyzable
which?	dono	do-no <i>Q-ATTR</i>	semi-analyzable
which?	dore		unanalyzable
who?	dare		unanalyzable
why?	dōshite	dō-s-ite <i>how-do-CONV</i>	analyzable phrasal
why?	naze	nani-semu-ni <i>what-do-ADV</i>	semi-analyzable

*Table 2: A list of question words in Japanese with analyzability. Data from Data from the World Loanword Database (Schmidt, 2009).*

### 3. Study 1 - Similarity of interrogative words

#### 3.1 Material

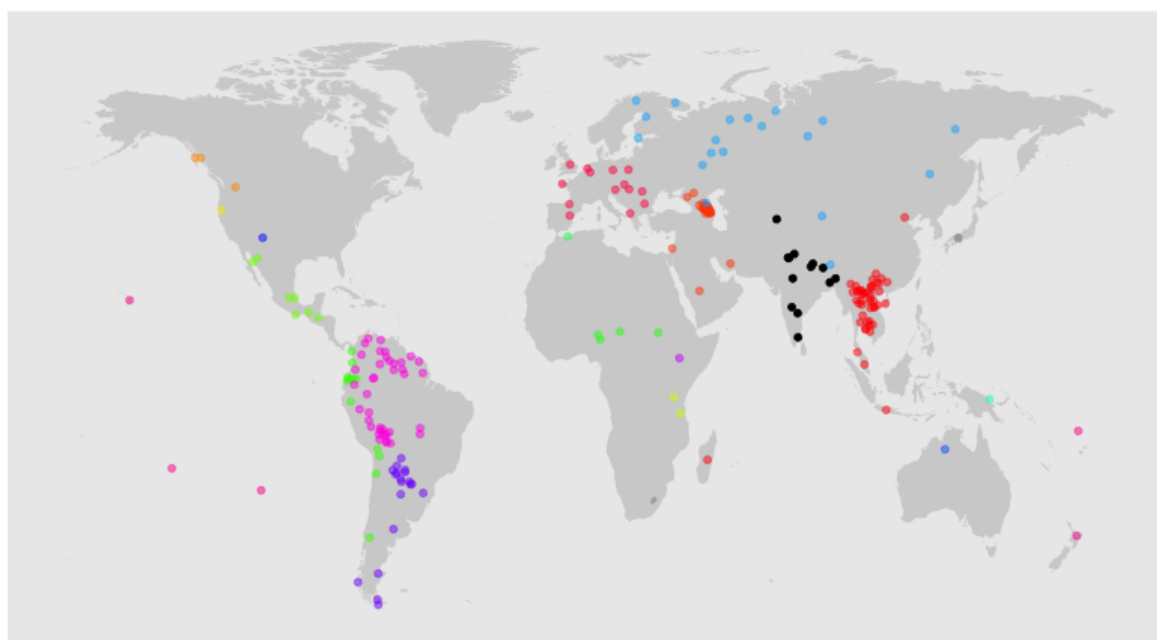
Lexical data was collected from the Intercontinental Dictionary Series (IDS) corpus (Key & Comrie, 2015), The World Loan Word Database (WOLD) (Haspelmath & Tadmor, 2009) and the Språkbanken word list database (Borin, Comrie & Saxena, 2013). The languages were chosen according to whether phonemic transcription was available. Phonemic transcriptions were added for English, Dutch and German from the CELEX database (Baayen, Pipenbrock & Gulikers, 1995). If a particular language had two or more words referring to one of these question words, they were all taken into account. The final dataset included only languages with entries for at least 5 out of 9 interrogative concepts (*how*, *how many*, *how much*, *what*, *when*, *where*, *which*, *who*, *why*, these are all separate concepts according to the IDS concept list). We excluded creole or reconstructed (proto-) languages. The IDS includes many closely related dialects, and we excluded languages with duplicate glottolog codes or which were explicitly marked as alternative varieties of other languages.

The final data set for analyses consisted of 226 languages which come from 66 different language families (including language isolates, according to Glottolog, Hammarstrom et al., 2016) and from 20 geographic areas (defined according to Autotyp regions, which capture known language contact areas, Nichols, Witzlack-Makarevich & Bickel, 2013, see fig 2). About 4% of possible entries were empty, either due to missing data or more frequently because a language did not have a given word. A full list of languages in the database can be found in the supporting information S1.

Later tests require sets of concepts with which to compare question words. The first set consisted of all other concepts in the corpus that had entries in at least 75% of the final languages (990 concepts). The IDS (and by inheritance WOLD and Språkbanken) divides concepts into 24 semantic fields (e.g. animals, religion and belief, sense perception), 20 of which had enough data for the languages considered. The concepts were matched across all three databases and each of these fields was used as a set of conceptually related concepts (some manual correction of the semantic field codes was carried out, see SI). Because question words are very similar in their pragmatic function, three sets of more closely related concepts were also used as a baseline, including: nine nouns from the domain of *body*, all relating to the head (*head*, *face*, *forehead*, *cheek*, *chin*, *eye*, *ear*, *nose*, *mouth*); nine verbs from the domain of *basic actions* (*do/make*, *fold*, *work*, *break*, *pull*, *press*, *wash*, *pour*, *build*); and a set of pronoun concepts (*I*, *you (singular)*, *he/she/it*, *we*, *you (plural)* and *they*). The pronouns in particular were meant to mirror the closed-class, tight semantic links between question words. In addition, pronouns are often derived from the same words or share phonological similarities (Bickel & Nichols, 2007: 209).

The initial segments of all words in the raw data were composed of 188 different segments, including specifications of aspiration, breathiness, palatalization and vowel length. Transcriptions came from different sources with different standards and conventions, and varied in the level of detail or range of features coded. In order to make the test more conservative and reduce impact of phonemic diversity across languages that could confound the results, the phonology was simplified by taking into account only voicing, place and manner of articulation of the phoneme. As a result, the simplified phonology consisted of 51 different initial segments. The R code for applying the simplifications and running analyses is available online (<https://github.com/seannyD/UniversalsInWHWords>) and full database of words is included in the supporting information S2<sup>1</sup>.

A subset of this data was extracted based on the analyzability of the words. The WOLD database codes words as “unanalyzable” (“the form cannot be analyzed into two or more constituents”), or varying degrees of analyzable (semi-analyzable, derived, compound or phrasal). The subset included only unanalyzable words and only languages with 5 or more unanalyzable wh-words. Note that this is a conservative measure, since WOLD only lists that the words are analyzable, not that they are composed of elements of other question words. This restricted the subset of words to 34 languages.



*Figure 2. Distribution of languages in the study, coloured according to geographic contact areas.*

---

<sup>1</sup> Note that the simplified phonology was designed specifically for looking at initial segments of words for this study. Other studies are advised to use the original sources for data.



### 3.2 Measuring similarity

Entropy efficiency ( $E_f$ ) was used to measure the similarity of a set of words. Entropy measures the amount of disorder in a set. A set of 9 identical segments would have a low entropy (low disorder, high similarity) and a set of 9 entirely different segments would have a high entropy (high disorder, low similarity). The exact value of entropy changes with the size of the set. Since different languages have different numbers of entries, we use a normalised entropy measure called entropy efficiency. This is a value between 0 and 1 which measures the amount of disorder as a proportion of the maximum possible disorder given the size of the set (0 = all segments are the same, 1 = all segments are different). Formally, if we have  $n$  different segment types in a set, named  $x_1, x_2, \dots, x_n$ , the probability of observing a given segment type  $x_i$  is  $p(x_i)$ , and the entropy efficiency is calculated as:

$$E_f = - \sum_{i=1}^n \frac{p(x_i) \log(p(x_i))}{\log(n)}$$

The following examples are provided to clarify the notion of entropy (see Table 1). In Hindi, all interrogative words start with a phoneme /k/. In this case there is no uncertainty within this set of words, therefore  $E_f$  is equal to 0. On the other hand, in Bulgarian all but two interrogative words start with /k/ (the /k/ phoneme is coincidental). In this case  $E_f = 0.28$ , still low but higher than for Hindi. In Dehong, all 9 interrogative words start with a different phoneme except two words, which start with /t/. Accordingly,  $E_f = 0.93$ . A language where all interrogative words start with a different phoneme would have  $E_f = 1.0$ .

### 3.3 Method: Permutation

All analyses were carried out in R (R core team, 2016). Random permutation tests were used to assess the significance of the similarity of the interrogative words (figure 3). The principle of random permutation is that if there are patterns in the data, then permuting the data - randomly swapping the membership of data points to languages - should destroy this pattern. This test has the advantages of not requiring a normal distribution and allowing unbalanced designs.

The data was organised into a matrix where rows were concepts and columns were languages, with each cell representing an entry (zero, one or more words). The mean entropy efficiency scores for each column (each language) was calculated, and the mean of these values is the *true mean entropy score*. The permutation involved randomly swapping words between languages within the same concept. The entries within a row were swapped randomly, and this was done for each row independently. This broke the links between words in the same language, so for example the Spanish entry for “who” might be swapped with the Dutch entry for “who”. The entropy score for each column was recalculated, and the mean of these values was a *permuted mean entropy score*. If words are more similar to each other within a language than between languages, then the permutation should increase the amount of variation and therefore increase the entropy score. If words are not similar within languages (the null hypothesis), then we would expect a random

permutation to result in a similar score. We carry out many permutations (e.g. 10,000) in order to obtain a distribution of permuted mean entropy scores. We can compare this distribution with the true mean entropy. We expect the true mean entropy to be lower than the majority of the permutations. Of course, some random permutations might result in a lower score than the true mean, but if 95% of permutations align with the prediction, then we reject the null hypothesis.

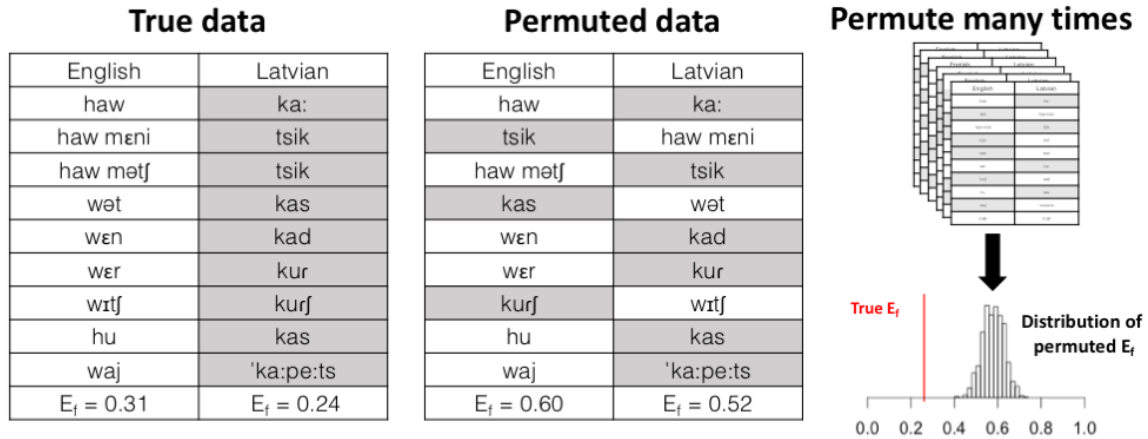


Figure 3: Demonstration of permutation method. The true data (left) is analyzed, calculating the entropy efficiency  $E_t$  of the first segment of words within each language. The data is randomly permuted (middle) and the entropy efficiency is calculated again. If words are more similar within languages than between languages, this should lead to an increase in dissimilarity within each language. Many different permutations are carried out (right, top) leading to a distribution of mean permuted entropy efficiency (left, bottom), which can be compared to the true mean entropy efficiency.

It has been suggested that apparent regularities within question words mostly occur within the Indo-European language family (Cysouw, 2004, p.3). It is likely that phonologies and lexicons of related languages might be more similar to each other. Therefore, permuting languages from different language families, which have less similar phonologies, could increase entropy. To address this, stratified permutation was applied, meaning that random permutation was allowed within language families, but not between them. Note that if there is only one language in a given language family, then the permuted data will always be identical to the true data for that language. This is a conservative process, since it will make the mean permuted entropy efficiency more similar to the true mean entropy efficiency. Since there are also areal patterns in phonological inventories, the same analysis was carried out allowing permutation only within geographical areas, and allowing permutation only within language families within the same geographic areas.

### 3.4 Results

Table 3 shows the numeric results of study 1. To recap the measures:  $E_f$  measures the amount of similarity in a set of words (low = more similar); the  $z$  value shows how different the real value is from the permuted value (the number of standard deviations away from the mean); and the  $p$  value indicates the proportion of permutations where the permuted value was lower than the true value, giving an idea of the probability that the null hypothesis (no difference) is true.

The  $E_f$  of first segments of interrogative words was significantly lower within languages (first segments of interrogative words are similar within languages) compared to a baseline  $E_f$  of freely permuted interrogative words. This result also held when comparing to baselines controlling for historical contact: permutation only within language families, permutation only within linguistic areas and permutation only within language families and areas.  $E_f$  increases when these controls were taken into account, meaning that interrogative words are slightly more similar within their language families and geographic area than between them. This is expected due to related languages sharing similar phonological inventories.

The  $E_f$  of interrogative words was higher when only unanalyzable interrogative words were considered, suggesting that part of the similarity between words is driven by compounds. However, the differences between unanalyzable words and all baselines was still significant. In summary, question words are more similar within a language than between languages.

Sample	True $E_f$	Baseline	Mean Permuted $E_f$	$p$	$z$
Initial segments of interrogative words	0.454	Free permutation	0.78	< 0.0001	72.36
		Permutation only within language families	0.60	< 0.0001	40.09
		Permutation only within linguistic areas	0.71	< 0.0001	58.28
		Permutation only within language families and areas	0.58	< 0.0001	35.93
Initial segments of unanalyzable interrogative words	0.513	Free permutation	0.73	< 0.0001	17.1
		Permutation only within language families	0.67	< 0.0001	7.19
		Permutation only within linguistic areas	0.67	< 0.0001	6.96
		Permutation only within language families and areas	0.65	< 0.0001	3.72

The same tests were done for the alternative sets of words (body, basic actions and pronouns).

When considering first segments and allowing permutation only within language families and areas (see figure 4 and supporting information S3 for full results), these words are also more similar within a language than between languages: same semantic domain ( $E_f = 0.76$ ,  $z = 1.94$ ,  $p = 0.04$ ); body ( $E_f = 0.69$ ,  $z = 13.64$ ,  $p < 0.0001$ ), basic actions ( $E_f = 0.70$ ,  $z = 15.3$ ,  $p < 0.0001$ ) and pronouns ( $E_f = 0.64$ ,  $z = -15.71$ ,  $p < 0.0001$ ). This is probably driven by the differences in phonologies between languages, (or possibly due to iconic sound-meaning associations across languages, Blasi et al., 2016; Dautriche et al., 2016), raising the possibility that the question words are not special. However, the  $E_f$  for question words is lower than for the other sets and the  $z$ -values are twice as extreme, as can be seen in figure 4 which shows the comparisons of  $E_f$  to the permuted distributions. In the next study, we test whether these differences are significant.

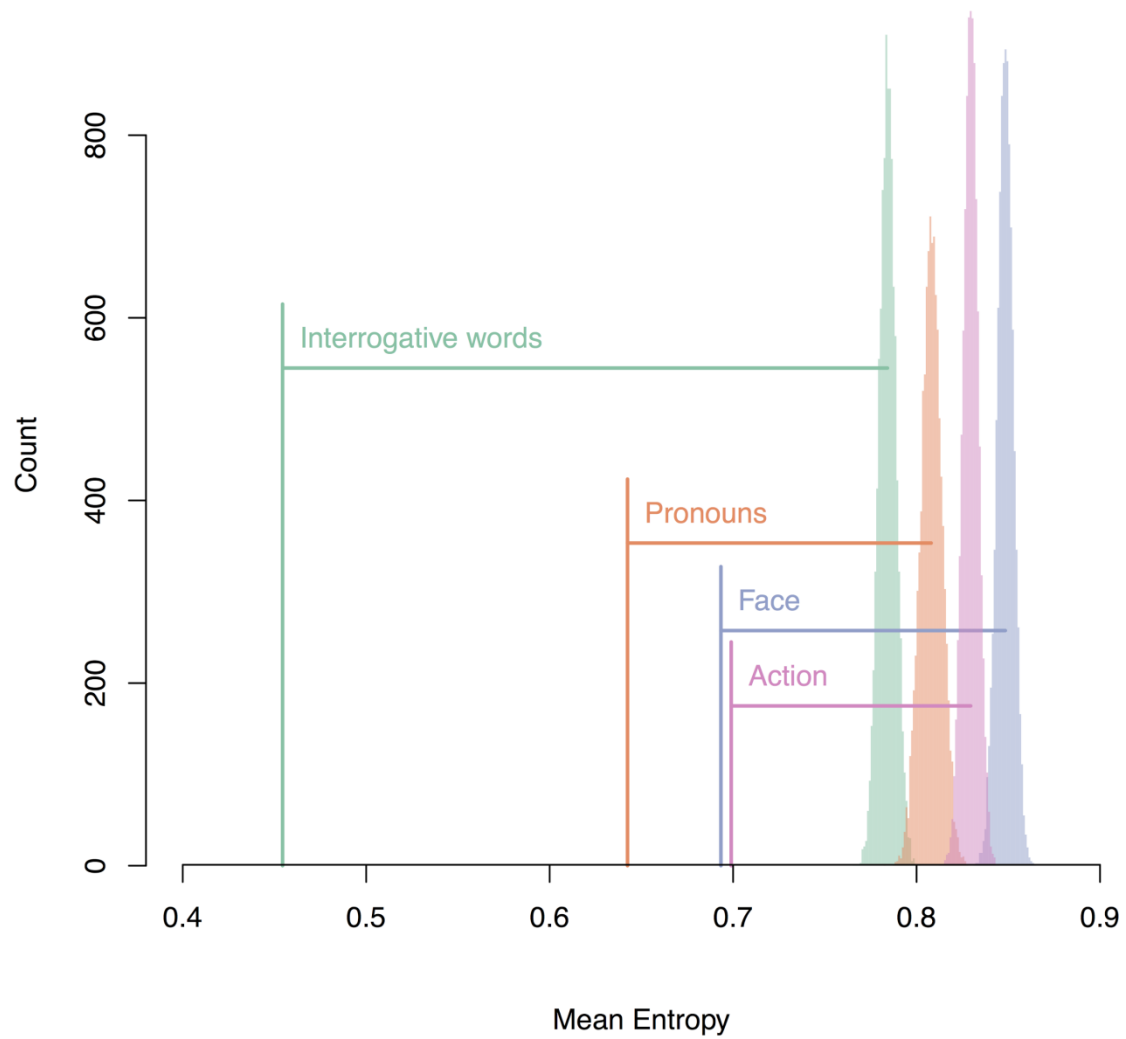


Figure 4: Results of study 1. Vertical lines show the mean entropy efficiency of a group of words (lower values indicate that words are more similar). The horizontal lines connect these values to the distribution of mean entropy efficiency when those words are permuted. For all groups of words, the mean entropy efficiency is clearly lower than the distribution of permuted values, indicating that they are significantly different from chance.

#### **4. Study 2 - Interrogative words vs. random words**

If the beginning of the question word has a pragmatic function, namely action ascription, then it should result in question words having more similarity within a language than a set of random words or a set of words that are conceptually related but do not have a particular pragmatic function. Accordingly, we hypothesize that in order for question words to be a plausible candidate in action ascription, not only should question words themselves be similar (as demonstrated in Study 1) but they should also be more similar than random or conceptually related words.

##### **4.1 Materials**

We used the same data as in Study 1 with the addition of the alternative concept sets.

##### **4.2 Method: Random samples**

To compare the difference in entropy between different concept sets we also used permutation (see figure 5). The entropy for each language is calculated for concept set A (e.g. question words) and concept set B (e.g. basic action words), giving two values for each language. The difference in the mean for group A and the mean from group B represents the true difference between the groups. Then the membership of these numbers to concept A or concept B is randomly permuted and the mean difference is re-calculated. If the two groups do not differ in mean entropy, the random permutation should result in roughly the same difference. If the difference is smaller than the true difference in more than 95% of permutations, then we can reject the null hypothesis and claim there is a significant difference between the groups. This procedure shares some principles with a standard t-test, except the t-test assumes that the values have a t-distribution (similar to a normal distribution), while permutation tests are valid with any kind of distribution.

When comparing with a baseline set of concepts with more than 9 items, a random selection of 9 concepts from the baseline set were selected for each comparison.

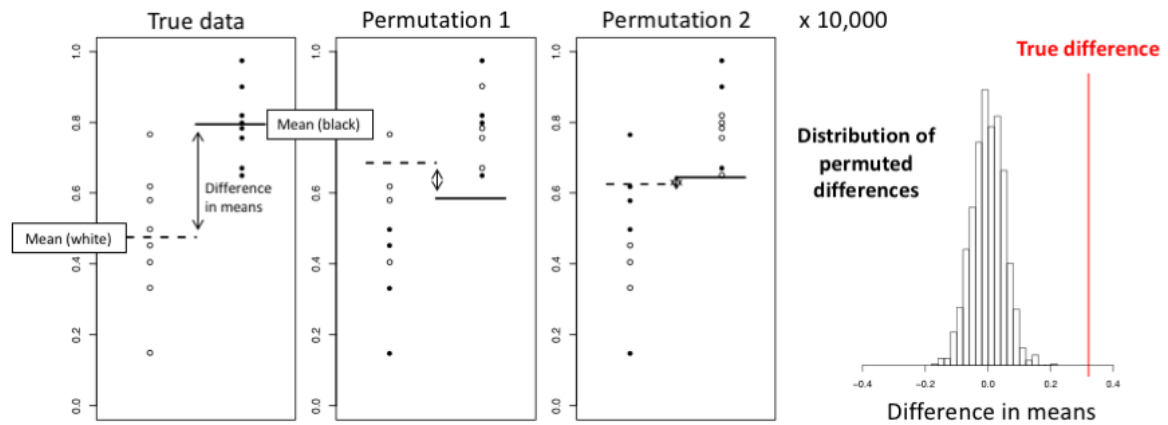


Figure 5: Demonstration of a permutation test of the difference between two groups. The true data has 8 data points in two groups (white and black). The difference in means between the two groups is calculated. When permuting the data (middle), the values are kept the same, but the membership to the group is randomly changed. A distribution of permuted differences is produced (right), which is usually centered around 0, to which the true difference can be compared.

### 4.3 Results

Table 4 shows the results. The mean  $E_f$  of interrogative words was significantly lower compared to random words and conceptually related words. The result also held for more strictly selected words from conceptually related domains - body concepts, basic actions and pronouns.

Sample	Mean $E_f$	Baseline	Mean $E_f$	$p$	$z$
Initial segments of interrogative words	0.454	Random words	0.79	< 0.0001	27.47
		Conceptually related words (20 sets)	0.76	< 0.0001	14.4
		Body concepts	0.69	< 0.0001	10.31
		Basic actions	0.70	< 0.0001	11.04
		Pronouns	0.64	< 0.0001	8.91
Initial segments of unanalyzable interrogative words	0.513	Random words	0.83	< 0.0001	12.97
		Conceptually related words	0.70	0.01	3.13
		Body concepts	0.75	< 0.0001	3.88
		Basic actions	0.76	0.0001	3.65
		Pronouns	0.67	0.005	2.53

Table 4: Permutation test results comparing the entropy of interrogative words to random and conceptually related words, within languages.



## 5. Study 4 - Detectability of interrogative words

The studies above show that interrogative words are more similar to each other than expected by chance. However, we would also predict that the words are easily detectable compared to other words. For example, they use distinctive initial phonemes that are less likely to be found in other words. The two measures are, in principle, independent, as we demonstrate with an example below.

Consider a language where 99% of words start with [s] and 1% start with [w]. If all wh-words start with [w], then they are both very similar to each other and very detectable (few other words start with [w]). However, if all wh-words started with [s], the similarity would still be high but they would not stand out from other words, so the detectability would be low.

### 5.1 Materials

We used the same materials as study 1, 2 and 3.

### 5.2 Method: Measuring detectability

The detectability of the initial segments of interrogative words can be measured in the following way. For each language, two lists of segments are extracted: the initial segments of interrogative words and the initial segments of all words. The probability of picking the interrogative segments from the list of all segments can then be calculated. If the interrogative segments are very common in the set of all segments (low detectability), then the likelihood of picking them at random is higher.

This probability can be calculated directly using the multivariate hypergeometric probability mass function. The list of all segments has  $K_i$  segments of type  $i$ , and the interrogative segments can be summarised as  $(k_s, k_z, k_w \dots)$ , where  $k_s$  is the number of [s] segments in interrogative words. The probability of selecting the interrogative segments from the list of all segments is then:

$$\text{Detectability} = \frac{\prod_i^c \binom{K_i}{k_i}}{\binom{N}{n}}$$

Where  $N$  is the number of segments in all words,  $n$  is the number of segments in the interrogative word list and  $c$  is the number of distinct segments in all words.

The value for the true interrogative words can be calculated, then compared to the same measure for many randomly selected words to produce a z-score and p-value. That is, the z-value represents how detectable the interrogative words are compared to a set of randomly selected words for each language.

### 5.3 Results

214 out of 226 languages (95%) had interrogative words with initial segments that were more detectable than randomly selected words, and that this was significant for 154 languages ( $p < 0.05$ , compared to 10,000 randomly chosen sets of words for each language), though the effect size is small (mean  $z = -0.53$ ).

## **6. Study 4 - Initial vs. non-initial interrogative phrase languages**

If front-loading of question words functions as a cue to determining a speech act, interrogative words that appear in initial position in the sentence should be under a greater pressure to change in order to exhibit a general cue to questionhood. We therefore hypothesise that interrogative words of languages that use initial interrogative phrases should be more similar than for languages with non-initial interrogative phrases.

### **6.1 Materials**

We used the same data as in studies 1 and 2. In addition we gathered typological data on the languages' positioning of the interrogative phrase. The World Atlas of Language Structures lists data for 81 languages in our sample (Dryer, 2013). 35 languages obligatorily place question words at the beginning of clauses, 44 languages do not and 2 languages have different strategies for different question words (the latter were excluded). We coded a further 32 languages following the coding scheme of Dryer (2013). 25 were initial, 6 were non-initial and 1 had a mixed strategy (languages with mixed strategies were excluded). See the Supporting Information S1 for details including sources and reliability coding procedure. The final analysis included 60 initial languages and 50 non-initial languages.

### **6.2 Method: Random independent sample test**

To compare the similarity of interrogative words in languages that use initial interrogative phrases and languages that do not, a random independent sample test was used. In each group (initial and non-initial), one language is chosen randomly from each language family, so that the data points within each group are (relatively) independent of historical influence. The same number of points are selected in each group (larger samples are more likely to include extreme values which could bias the result). The mean entropy efficiency of languages within each of these sub-groups is calculated. We then test whether the mean entropy for the initial languages is lower than for the non-initial languages. This process is repeated many times. The result of the test is the proportion of random independent samples in which initial languages have a lower mean entropy than non-initial languages. If this is more than 95% of the samples, then the null hypothesis (no difference) can be rejected. This test has the advantage of controlling for historical influence, but also does not require the values to be normally distributed. The same test can be done for linguistic areas instead of families.

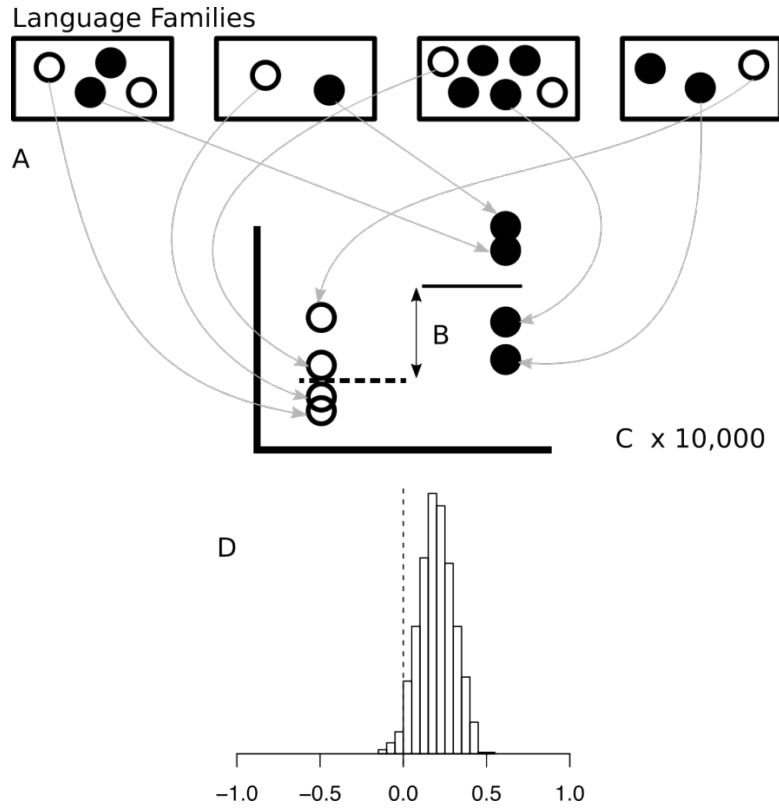


Figure 6: Demonstration of an independent samples test. Languages are represented as circles, with white circles being initial interrogative languages and black circles being non-initial interrogative languages. One language is randomly chosen from each language family for each group (A). The mean for each group in the sub sample is calculated and compared (B). This process is repeated many times (C) to form a distribution of differences in means (D). If 95% of random samples result in a value greater than zero, then the null hypothesis can be rejected.

### 6.3 Results

The results are shown in table 5. For interrogative words, the entropy efficiency of initial interrogative languages was not significantly different from the entropy efficiency of non-initial interrogative languages when controlling for language family or geographic area<sup>2</sup>. Both results for unanalyzable words are significant, though this is based on far fewer languages.

There was no significant difference between initial and non-initial interrogative languages for random sets of words, conceptual related sets, basic actions nor pronouns. However, body words in initial

<sup>2</sup> In the initial submission of this paper, and in some aural presentations, we reported that initial languages had more similar wh-words than non-initial languages. These results were based on earlier versions of WOLD and the IDS, which were updated during the review of this paper. In the first version there were 172 languages, and the result was significant when controlling for language families ( $p < 0.05$ ), though all other results were qualitatively the same. The current results are based on the most up-to-date data from WOLD and the IDS. The difference in results is due to the historically contingent nature of the tests: a few extra languages in a small language family can have a big impact on the results.

interrogative languages were more similar than body words in non-initial interrogative languages when controlling for language family (and marginal when controlling for geographic area).

Sample		IIP languages (N)	Not-IIP languages (N)	Number of samples	Mean difference $E_f$	$p$	$z$
Interrogative words	Restricted by family	60	50	20000	-0.08	0.098	1.3
	Restricted by area	60	50	20000	-0.05	0.256	0.66
Unanalyzable interrogative words	Restricted by family	11	10	10000	-0.24	0.004*	2.84
	Restricted by area	11	10	10000	-0.23	0.005*	2.6
Random words (50 sets)	Restricted by family	60	50	48871	-0.01	0.396	0.27
Conceptually related (20 sets)	Restricted by family	60	50	94112 1	-0.03	0.270	0.61
Conceptually related: Body	Restricted by family	60	50	20000	-0.14	0.005*	2.48
	Restricted by area	60	50	20000	-0.1	0.067	1.45
Conceptually related: Basic actions	Restricted by family	60	50	20000	-0.03	0.273	0.62
	Restricted by area	60	50	20000	-0.01	0.426	0.22
Conceptually related: Pronouns	Restricted by family	60	50	20000	-0.01	0.410	0.22
	Restricted by area	60	50	20000	-0.03	0.326	0.46

*Table 5: Results of study 3: random independent samples comparing initial interrogative phrase languages and non-initial interrogative phrase languages. If the mean difference in  $E_f$  is negative, then the initial interrogative languages had a lower  $E_f$  (more similar) than non-initial interrogative languages. Note that duplicate sample sets were excluded, so the number of samples does not always reach the target.*

### 6.3.1 Detectability in initial and non-initial languages

We also compared the detectability z-scores of initial and non-initial interrogative languages by a permutation test and by random independent sample test. In the latter, the same number of independent languages were selected from each group and the difference in means was calculated. This was repeated 10,000 times to produce a distribution of differences. If initial interrogative languages are more detectable than non-initial interrogative languages, then we would expect the mean z-value for initial languages to be more extreme in more than 95% of samples.

There was no evidence that initial interrogative languages had more detectable interrogative words than non-initial interrogative languages (mean detectability z-score for initial = -0.63, mean z-score for non-initial = -0.59, 10,000 permutations,  $p = 0.72$ ). The z-scores were highly non-normal, but additional analyses were done by excluding data more than two standard deviations above or below the mean (excluding 6 languages). In order to control simultaneously for language family and geographic area, we ran a mixed effects model (in R using the package lme4, Bates et al., 2015) predicting z-score by interrogative position, with random intercepts for language family and geographic area. The difference was in the predicted direction, but model comparison showed that interrogative position did not significantly improve the fit of the model (log likelihood difference = 0.27,  $\text{chisq} = 0.26$ ,  $\text{df} = 1$ ,  $p = 0.60$ ).

## 7. Additional analyses

Some additional analyses were carried out, considering all segments of words instead of just the first segment and considering vowels and consonants separately. The summaries of these tests are available in supporting information S3. Broadly speaking, the same patterns held. When analysing all segments instead of first segments, the entropy efficiency increases, but question words are still more similar within languages than between languages (study 1) and were significantly more similar than randomly selected words (study 2). When considering all segments of unanalyzable question words, initial interrogative languages were not significantly more similar than non-initial languages (study 3). Initial interrogative languages were more similar than non-initial languages when considering first consonants (mean difference = 0.12,  $z = 1.89$ ,  $p = 0.03$ ) and slightly weaker when considering first vowels (mean difference = 0.08,  $z = 1.80$ ,  $p = 0.04$ ). The significance does not hold when controlling for geographic area. In general, then, the predicted effects were more evident in initial segments and for consonants.



## **8. General summary**

Study 1 showed that question words are more similar on average within languages than between languages. Study 2 showed that question words were more similar on average than other sets of words (within languages), including randomly chosen words, conceptually related words and tightly related words (body concepts, basic actions and pronouns). Findings for study 1 and 2 also held when using only unanalyzable words. Study 3 showed that for 68% of languages in the sample, question words started with significantly salient phonemes (phonemes used at the start of few words). Study 4 tested whether question words are more similar on average in languages that place them at the beginning of clause. The trend was in the predicted direction, but significance was not obtained for many tests. In addition, there was no difference in salience between the two language types. Furthermore, body concepts showed the same pattern as question words.

## **9. Discussion and conclusions**

Real-time conversation is a harsh ecology to which the forms and structures of language must adapt in order to replicate and survive. One understudied evolutionary pressure is the speed at which interlocutors must recognise the pragmatic action of their partner's turn in order to plan their own turn. We identified answers to content questions as a particularly challenging environment, and reviewed previous work on paralinguistic cues that help interlocutors identify upcoming questions (eye gaze, intonation etc.). It is also plausible that the pressure for action ascription has an impact on the structures of words and phrases. For example, question words are often front-loaded in the turn. We hypothesised that question words might undergo a cultural evolutionary pressure to sound similar in order to provide an additional, low-level cue. This lead to an additional prediction that the pressure would be greater for languages that place question words at the beginning of a turn.

The aim of the present study was to explore whether there was a statistical trend for languages to have similar sounding question words. Although previous analyses suggested that there are no strong universal patterns in wh-words (Cysouw, 2004), the present results suggest that question words sound more similar than would be expected by chance. This supports the hypothesis that language adapts to the pressures of interaction in conversation.

However, the present study should be considered with caution. First, there is undeniable impact of contact between languages - the results weaken when language family or geographic area are controlled for. The controls here are reasonably coarse, and more detailed controls for relatedness could be applied (e.g. for language families where reliable phylogenetics are available, phylogenetic generalised least squares). Nonetheless, the tests still show significant effects which signals that results are not likely to be confounded with these factors.

Secondly, and probably most importantly, the biggest chance of bias is present in regard to the diverse phonologies of the languages of the world. In the present study the issue was addressed by means of simplified phonology. This made the languages more comparable and the permutation test in study 1 more conservative. However, it could also have obscured differences between question words that are important in some languages. Furthermore, we found that other sets of words were more similar within languages than between languages. Therefore, this was not a surprising result.

A more important result for the theory is that question words were more similar than other sets of words. This included randomly selected words, words within the same semantic domain, words that referred to parts of the face, basic action words and pronouns. In general, conceptually related words are under a pressure to be easily identifiable. We argue that words that indicate the pragmatic role of the turn in conversation have an additional pressure to sound similar in order to aid action ascription. 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.

We also found some preliminary evidence that question words tend to start with distinctive segments, though not all languages exhibited this property. The operationalisation of distinctiveness was based on frequency of segments in a small lexicon. This could be improved to take into account acoustic saliency (e.g. sonority, see Parker, 2012), though this might be difficult for a wide array of languages.

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 tests on the main data were not significant. Tests were significant when using only unanalyzable words, but this was only run on the 21 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 supported by this study. Even the general trend may be explained by other mechanisms. For example, initial interrogative languages are more likely to place the verb before the subject in canonical sentences (see supporting information S4). Basic word order has well known areal patterns, and may also have a knock-on effect on the distribution of information in words (Maurits, Perfors & Navarro, 2010). For example, if verbs come before subjects, there might be greater information conveyed through verbs, leading to a lower pressure to make subject elements (like question words) more distinct. Controlling for these issues requires the theory to be fleshed out, including how a bias for uniform information density (see Jaeger, 2010; Mahowald, 2013) and ease of processing interacts with a bias for front-loading (Levinson, 2013; Hofmeister, et al., 2007; Roberts & Levinson, in press), and for more complex statistical models which can untangle networks of causal effects (e.g. causal graph inference, see Roberts & Winters, 2013). The current study also used very a coarse typology, and richer information could be used.

As far as we know, the database collected in this project represents one of the largest and widest collection of phonemically transcribed lexical items currently available (over 900 concepts in 172 languages, around 250,000 entries; the Automated Similarity Judgement Program database, Wichmann & Holman, 2016, has around 260,000 entries for over 7,000 varieties, but an order of magnitude fewer concepts and very few question words; Mackenzie, 2009 has a more detailed analysis of content question words but only 50 languages; RefLex, Segerer & Flavier, 2016, has greater transcription consistency and more languages and concepts but is mainly restricted to Niger Congo languages for question words; DiACL, Carling, 2017, has 200-word Swadesh lists for 500 doculects). However, the current study uses a simple coding scheme, assuming that 9 distinctions in question words are broadly applicable in all languages. Also, the sample is not well balanced (there are proportionately few languages from Africa, Australia and Papua), and the linguistic treatment of languages could be improved. For example, a considerable proportion of languages come from South America, which often bear a common interrogative particle not at the beginning of the question words, but at the end of it. For example Ayacucho Quechua languages make use of particle *taq* to identify question words (Cerron-Palomino, 2008), though initial segments differ. Aymaran languages makes use of the particle *sa* at the end of the question words, but the similarities in initial segments can also be observed (see also e.g. /-ro/ in Ache, Heckart & Hill, 2015; /-aj.1/ in Thai, Peiros, 2015; and examples in Mackenzie, 2009). We simplified the analysis by only looking at the initial segment (because that would be most useful for rapid action ascription, though see the supporting information S3 for a summary of analyses using all segments). Future research should continue the search for the patterns in order to establish a complete picture of how question words are identified and accordingly how they could prompt question recognition.

The current study demonstrated a synchronic pattern, assuming that processes of cultural evolution brought them about. However, more a detailed theory should be worked out. For example, since the pressures from conversation have been around for a very long time (Levinson, 2006), one might expect a stronger statistical signal if the pressure for action ascription in content questions was very strong. Instead, it is more likely that a number of evolutionary processes are at work, including grammaticisation and sound changes which affect the adaptive environment for question words. As such, question word similarity may be more of an exaptation - a product of a weak bias to tweak existing forms to better serve rapid turn taking.

An important source of evidence would come from a diachronic analysis. For example, do languages evolve to become more supportive of action ascription, or perhaps do question words become more similar after the languages begin placing the question word at the start of turns? Diachronic change in interrogative structures is a complex area which is outside the scope of this paper (see Mackenzie 2009; Mao, 2012; Huang, 2012), but we note that at least in some cases languages drift in the opposite direction to the prediction. For example, there is more similarity in the initial segments of question words in Old English question (*hū, hwā, hwæt, hw*⌘) than modern

English. Further research is required on this front, and we note that new digital resources are being developed that could help here (e.g. DiACL, Carling, 2017).

However, the most important remaining issue for the general theory, in our view, is to link the cross-cultural finding here with actual use in conversation. There are two obvious questions. Firstly, are phonetic segments like /w/ and /h/ in English actually reliable cues to content questions in real conversations? Secondly, do interlocutors actually use these cues to predict upcoming pragmatic actions? [Another study by the current authors, XXXX] attempts to address these questions, and find that, in English at least, the answer is affirmative.

## **10. Conclusion**

We argued that the social norms of conversation put a pressure on interlocutors to take precisely timed turns, which leads to a substantial cognitive load. Answering content questions in particular involves simultaneous comprehension of the question and planning of the answer. Answering on time would be facilitated by the ability to rapidly recognise the pragmatic action of turns, for example by the presence of cues for questions. We hypothesised that similarity in question words could form such a cue, and provided evidence that such a systematicity exists in the languages of the world. We take this as initial evidence in favour of the hypothesis. At the very least, an assumption that there are no regularities in question words should be reconsidered.

While there are many shortcomings of this study, and much more ground to cover before the theory is fully supported, we hope to have demonstrated that falsifiable hypotheses can be formulated relating the socio-cognitive pressures from conversation to statistical patterns in the form and structure of the world's languages, through cultural evolution. Importantly, these hypotheses can be tested using rigorous quantitative methods. We look forward to future studies which address how language adapts to interaction.

## **Acknowledgements**

To be added after review.

## **Supporting information**

All code and data is available through github: [link to github removed to protect anonymity]

S1 - List of languages used in the analyses, together with notes on the typological coding of interrogative position.

S2 - Lexical data used in the analyses. Available at [link to github removed to protect anonymity]

S3 - Results for alternative analyses.

S4 - Statistical test of correlation between interrogative position and basic word order.

## References

- Alves, M. (2009) Vietnamese. In Haspelmath, M. & Tadmor, U. (eds.) *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online <http://wold.cld.org/language/24>
- Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final 'go-signals'. *Frontiers in Psychology*, 8, 393.
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.
- Berlin, B. (1994). Evidence for pervasive synthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nicholls, & J.J. Ohala (Eds.), *Sound symbolism* (pp. 76–93). New York: Cambridge University Press.
- Berlin, B. (1994). Evidence for pervasive synthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nicholls, & J.J. Ohala (Eds.), *Sound symbolism*. New York: Cambridge University Press, pp. 76–93).
- Bickel, B., Nichols, J., 2007. Inflectional morphology. In: Shopen, T. (Ed.), *Language Typology and Syntactic Description*, 2nd edition, vol. III. Cambridge: Cambridge University Press, Cambridge, pp. 169–240.
- Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39), 10818–10823.
- Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.
- de Boer, B. (2000). Self-organization in vowel systems. *Journal of phonetics*, 28(4), 441–465.
- Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports*, 5, 12881.
- Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports*, 5.
- Borin, L., Comrie, B. & Saxena, A. (2013). The Intercontinental Dictionary Series – a rich and principled database for language comparison. Online <https://spraakbanken.gu.se/eng/research/digital-area-linguistics/word-lists>
- Carling, G. (ed.) 2017. *Diachronic Atlas of Comparative Linguistics Online*. Lund: Lund University. (DOI/URL: <https://diachl.lu.se/>. Accessed on: 29/3/2017)
- Casillas, M. (2014). Turn-taking. In: D. Matthews (ed.), *Pragmatic Development in First Language Acquisition* (pp. 53–70). Amsterdam/ Philadelphia: John Benjamins Publishing Company
- Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30(3), 348–369.
- Cassidy, K.W., Kelly, M.H. (1991) Phonological

information for grammatical category assignments. *Journal of Memory and Language*. 30, pp. 348–369

- Cerron-Palomino, R.M., 2008. Quechumara. Estructuras paralelas de las lenguas quechua y aimara. Universidad Mayor de San Simón, PROEIB Andes, Plural Editores, La Paz, [First edition 1994, Centro de Investigacion y Promocion del Campesinado (CIPCA), La Paz].
- Clark, H. H., and Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition* 84, 73–111. doi: 10.1016/S0010-0277(02)00017-3
- Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.
- Cysouw, M. (2004). Interrogative words: an exercise in lexical typology. Unpublished Manuscript. Online: [http://www.cysouw.de/home/manuscripts\\_files/cysouwQUESTION\\_handout.pdf](http://www.cysouw.de/home/manuscripts_files/cysouwQUESTION_handout.pdf)
- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2016). Wordform Similarity Increases With Semantic Similarity: An Analysis of 100 Languages. *Cognitive Science*.
- Dediu, D., & Moisik, S. R. (2016). Anatomical biasing of click learning and production: An MRI and 3d palate imaging study. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVLANG11)*. Retrieved from <http://evolang.org/neworleans/papers/57.html>.
- Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS One*, 8(11): e78273. doi:10.1371/journal.pone.0078273
- Enfield, N. J., Stivers, T., & Levinson, S. C. (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42(10), 2615-2619.
- Estrada Fernández, Z. (2009) Yaqui vocabulary. In: Haspelmath, Martin & Tadmor, Uri (eds.) *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wold.clld.org/vocabulary/32>, Accessed on 2017-04-12.)
- Everett, Daniel L. (1986). Pirahã. In: Desmond C. Derbyshire & Geoffrey K. Pullum (eds.) *Handbook of Amazonian Languages*. Vol. 1, pp. 200-325. Berlin: Mouton de Gruyter.
- Feher O., Smith K., Wonnacott E. and Ritt N. (2016). Communicative Interaction Leads To The Elimination Of Unpredictable Variation. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVLANG11)*. Available online: <http://evolang.org/neworleans/papers/137.html>
- Ferrer-i-Cancho, R. (2008). Some word order biases from limited brain resources: A mathematical approach. *Advances in Complex Systems*, 11(03), 393-414.
- Foley, W. A. (1991). *The Yimas Language of New Guinea*. Stanford: Stanford University Press.
- Gasser, M. (2004, January). The origins of arbitrariness in language. In *Proceedings of the cognitive science society* (Vol. 26, No. 26). Gasser, M. (2004). The origins of arbitrariness in language. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (Vol. 26, pp. 4-7).

- Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PloS one*, 10(3), e0120068.
- Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. (2016) Glottolog 2.7. Jena: Max Planck Institute for the Science of Human History. (Available online at <http://glottolog.org>, Accessed on 2016-09-01.)
- Harper, D. (2016) Etymologies of \*hwa. Online etymology dictionary. Online: [http://etymonline.com/index.php?allowed\\_in\\_frame=0&search=hwa](http://etymonline.com/index.php?allowed_in_frame=0&search=hwa)
- Harper, D. (2016b) Etymology of “what”. Online etymology dictionary. Online: <http://etymonline.com/index.php?term=what>
- Haspelmath, Martin & Tadmor, Uri (eds.) 2009. World Loanword Database. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wold.clld.org>, Accessed on 2016-09-01.)
- Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge Studies in Linguistics, volume 73. Cambridge: Cambridge University Press.
- Heckart, C.J., Hill, K. (2015) Aché dictionary. In: Key, Mary Ritchie & Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <http://ids.clld.org/contributions/288>, Accessed on 2017-02-18.
- Heritage, J. (2012). The epistemic engine: Sequence organization and territories of knowledge. *Research on Language & Social Interaction*, 45(1), 30-52.
- Hinds, J., 1986. Japanese (Croom Helm Descriptive Grammars). Croom Helm, London.
- Hofmeister, P., Jaeger, T. F., Sag, I. A., Arnon, I., & Snider, N. (2007). Locality and accessibility in wh-questions. In S. Featherston & W. Sternefeld (eds.) *Roots: Linguistics in search of its evidential base*. de Gruyter, 185-206.
- Huang, L. (2012). Grammaticalization in Pragmatics: A New Approach to Contrastive Study on English and Chinese. *Foreign Language Research*, 3, 020.
- Hutchins E. (2006). The Distributed Cognition Perspective on Human Interaction. In Enfield N. J. & Levinson S. C. (eds.) *Roots of human sociality: Culture, cognition, and human interaction*. 375-398. Oxford: Berg.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1), 23-62.
- Janssen, R., Winter, B., Dediu, D., Moisik, S. R., & Roberts, S. G. (2016). Nonlinear biases in articulation constrain the design space of language. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVLANG11)*. Retrieved from <http://evolang.org/neworleans/papers/86.html>
- Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: a quantitative study. *Discourse Processes*, 52(4), 255-289.



- Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., & Levinson, S. C. (2014). Sequence organization: A universal infrastructure for action. *4th International Conference on Conversation Analysis*. UCLA, CA.
- Key, M. R. & Comrie, B. (eds.) 2015. The Intercontinental Dictionary Series. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://ids.clld.org>)
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681-10686.
- Legendre, P. L., & Legendre, L. (1998). L. 1998. *Numerical ecology*. Second English Edition. Amsterdam Elsevier Science.
- Lehmann, C., 1993. On the system of semasiological grammar. Unpublished paper, available at [http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/d\\_lehmann.html](http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/d_lehmann.html).
- Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.
- Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(01), 1-38.
- Levinson, S. (2006). On the human interaction engine. In Enfield, N. and Levinson, S., (editors), *Roots of Human Sociality: Culture, Cognition and Human Interaction*, pp.ages 39–69. Oxford: Berg.
- Levinson, S. C. (2013). Action formation and ascription. In J. Sidnell & T. Stivers (eds.) *The handbook of conversation analysis*, 101-130.
- Levinson, S. C. (2016). Speech acts. In Y. Huang (Ed.), *Oxford handbook of pragmatics. Advanced online publication*. Oxford: Oxford University Press.
- Levinson, S. C. (2016). Turn-taking in human communication, origins, and implications for language processing. *Trends in Cognitive Sciences*, 20(1), 6-14.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731.
- Lindström, Eva (1995). Animacy in interrogative pronouns. In: Inger Moen, Hanne Gram Simonsen & Helge Lødrup (eds.) *Papers from the 15th Scandinavian Conference of Linguistics*, pp. 307-15. Oslo: University of Oslo.
- Lupyan, G. & Dale, R. (2016) Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *Trends in Cognitive Sciences* 20(9), 649–660
- Mackenzie, J. L. (2009). Content interrogatives in a sample of 50 languages. *Lingua*, 119(8), 1131-1163.
- Macuch Silva V. and Roberts S. G. (2016). Language Adapts To Signal Disruption In Interaction. In S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér & T. Verhoef (eds.) *The Evolution of Language: Proceedings of the 11th International Conference (EVOlang11)*. Available online: <http://evolang.org/neworleans/papers/20.html>

- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313-318.
- Mao, A. (2012). Grammaticalization Mechanism of English Interrogative Pro-forms. *Overseas English*, 16, 113.
- Matthew S. Dryer. 2013. Position of Interrogative Phrases in Content Questions. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/93>)
- Maurits, L., Perfors, A., Navarro, D. (2010) Why are some word orders more common than others? A uniform information density account. *Advances in Neural Information Processing Systems*, 23 (pp. 1585-1593)
- Mayr, E. (1961). Cause and effect in biology. *Science*, 134(3489), 1501-1506.
- Micklos, A. (2014) The Nature of Language in Interaction. In E. Cartmill, S. Roberts, H. Lyn & H. Cornish (eds.) *Proceedings of the 10th Evolution of Language conference*. World Scientific: Vienna, Austria.
- Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive psychology*, 55(4), 259-305.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3), 325.
- Nichols, J., Witzlack-Makarevich, A., and Bickel, B. (2013). The AUTOTYP genealogy and geography database: 2013 release. Zurich: University of Zurich.
- Parker, S. (Ed.). (2012). *The sonority controversy*. Berlin, Germany: de Gruyter Mouton. Walter de Gruyter.
- Peiros, I. (2015) Dehong dictionary. In: Key, Mary Ritchie & Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://ids.cld.org/contributions/405>, Accessed on 2017-02-18.)
- Peiros, I. (2015). Central Thai dictionary. In: Key, Mary Ritchie & Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <http://ids.cld.org/contributions/401>, Accessed on 2017-02-18.
- Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1).
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Roberts S. G. & Mills G. J. (2016) Language Adapts to Interaction. In S. Roberts & G. Mills (Eds.) *Proceedings of EvoLang XI*, Language Adapts to Interaction Workshop, 21 March, 2016. Available online: [http://evolang.org/neworleans/workshops/papers/LATI\\_1.html](http://evolang.org/neworleans/workshops/papers/LATI_1.html)

- Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6), EL471-EL477.
- Roberts, S. G., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS one*, 8(8), e70902.
- Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in psychology*, 6, 509.
- Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6.
- Roberts, S. G. & Levinson S. C. (in press). Conversation, cognition and cultural evolution: a model of the cultural evolution of word order through pressures imposed from turn taking in conversation. *Frontiers in Psychology*.
- Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis*. Malden, MA: Wiley-Blackwell, 308–329.
- Rossano, F. (2013). Gaze in conversation. *The handbook of conversation analysis*, 308-329.
- Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. *Conversation analysis: Comparative perspectives*, 187-249.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 696-735.
- Saxena, A. (2015) Telugu dictionary. In: Key, Mary Ritchie & Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://ids.cld.org/contributions/709>, Accessed on 2017-02-18.)
- Schegloff, E. A. (2007). Sequence organization in interaction: Volume 1: A primer in conversation analysis (Vol. 1). Cambridge: Cambridge University Press.
- Schmidt, K. (2009) Japanese. In Haspelmath, M. & Tadmor, U. (eds.) *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online <http://wold.cld.org/language/21>
- Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, 29(1), 86-102.
- Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6(1), 38-47.
- Segeer G., Flavie S., 2016 RefLex: Reference Lexicon of Africa, Version 1.1. Paris, Lyon. <http://reflex.cnrs.fr/>
- Sicoli, M. A., Stivers, T., Enfield, N. J., & Levinson, S. C. (2015). Marked initial pitch in questions signals marked communicative function. *Language and Speech*, 58(2), 204-223.
- Sicoli, M. A., Stivers, T., Enfield, N. J., & Levinson, S. C. (2014). Marked initial pitch in questions signals marked communicative function. *Language and Speech*, 0023830914529247.

- Stephens, G. & Matthews, D. (2014). The communicative infant from 0-18 months. In: D. Matthews (ed.), *Pragmatic Development in First Language Acquisition* (pp. 13-35). Amsterdam/Philadelphia: John Benjamins Publishing Company
- Stivers, T., & Robinson, J. D. (2006). A preference for progressivity in interaction. *Language in society*, 35(03), 367-392.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587-10592.
- Strömbergsson, S., Hjalmarsson, A., Edlund, J., and House, D. (2013). "Timing responses to questions in dialogue," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2013* (Lyon: International Speech and Communication Association), 2584–2588.
- Tamariz, M., Cornish, H., Roberts, S. & Kirby, S. (2012) The effect of generation turnover and interlocutor negotiation on linguistic structure. In T. C. Scott-Phillips, M. Tamariz, E.A. Cartmill & J.R. Hurford, *The Evolution of Language: Proceedings of the 9th International Conference (EVOlang9)*. World Scientific. p. 555
- Thompson, S. A. (1998). A discourse explanation for the cross-linguistic differences in the grammar of interrogation and negation. *Case, typology and grammar: In honor of Barry J. Blake*, 309-341.
- Ullan, R. (1978). Some general characteristics of interrogative systems. In: J. H. Greenberg (ed.) *Universals of Human Language*. Vol. 4: Syntax, pp. 211-48. Stanford: Stanford University Press.
- Vakareliyska, C. & Horissian, K. (2015) Bulgarian dictionary. In: Key, Mary Ritchie & Comrie, Bernard (eds.) *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://ids.clld.org/contributions/200>, Accessed on 2017-02-18.)
- Verhoef, T., Roberts, S. G., & Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In D. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)* (pp. 2481-2486). Austin, Tx: Cognitive Science Society.
- Wichmann, S., Holman, E.W. & Brown, C.H. (eds.). 2016. The ASJP Database (version 17). <http://asjp.clld.org/>.
- Zipf G. (1949) *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley

## Vitae

Anita Slonimska is a research fellow at ISTC, CNR in Rome and a PhD candidate at CLS, Radboud University in the Multimodal Language and Cognition lab. She holds a master's degree in general linguistics and a research master's degree in language and communication, both from Radboud University. She studies how pragmatics affects interaction in both vocal and visual modalities in spoken and signed languages.

Seán G. Roberts is a Leverhulme fellow at the Department of Archaeology and Anthropology at the University of Bristol. He holds a doctorate in evolutionary linguistics from the University of Edinburgh. He studies how evolutionary theory can be applied to studying the similarities and differences between languages and cultures.

## Highlights

- Interlocutors need to rapidly identify the pragmatic content of turns in order to respond appropriately in real-time conversation.
- In particular, responding to content questions constitutes a harsh cognitive environment.
- Languages may adapt to this ecology in order to facilitate rapid turn taking such as providing a low-level phonetic cue to upcoming questions.
- Wh-words sound systematically similar within a language, controlling for some basic confounds.

## Title page

A case for systematic sound symbolism in pragmatics: Universals in *wh*-words

Anita Slonimska

Centre for Language Studies, Radboud University, PO Box 9103, 6500 HD, Nijmegen, Netherlands.

Institute for Cognitive Science and Technology, Italian National Research Council (CNR), Via S.Martino della Battaglia 44 - 00195, Rome, Italy.

Seán G. Roberts\*

Department of Archaeology and Anthropology, University of Bristol, 43 Woodland Road, BS8 1UU, Bristol, UK.

Language and Cognition Department, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, Netherlands.

\* Corresponding Author: [sean.roberts@mpi.nl](mailto:sean.roberts@mpi.nl)