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 172 languages from 65 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 95 languages, and found the predicted trend, although statistical significance was not attained when controlling for areal contact. 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 *wh-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 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] 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 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 of 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.

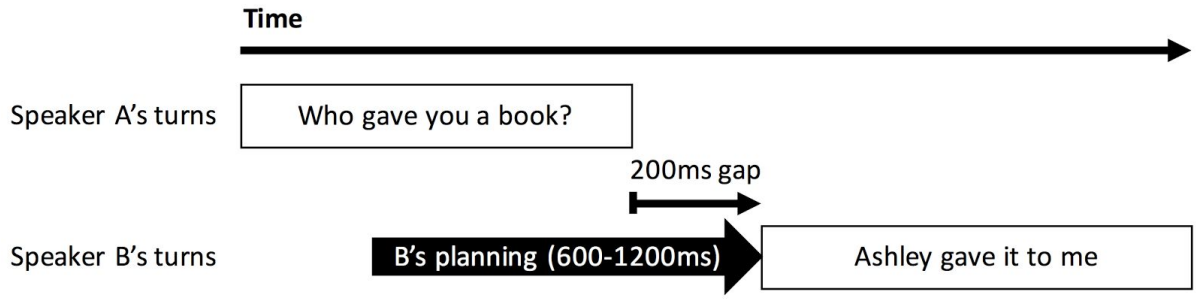

Figure 1. Overlap of comprehension and production in conversation (Levinson 2013, p.104). The typical gap between two turns is of the order of 200 milliseconds, but production planning takes at least 600 milliseconds, meaning that B must start planning their turn in the middle of comprehending the previous 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, Gisladottir, 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 Gisladottir, 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 to greet 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 particular 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 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). 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, but we note that there is no relation between degree of resonance and cognitive complexity.

Wh-word (English)	**Aymara** Aymaran	**Telugu** Dravidian	**Bulgarian** Indo-European	**Dehong** Tai-Kadai	**Vietnamese** Austroasiatic
how	***k**umasa*	***e**laa*	***k**ák*	*com.2 sə.2*	*sao*
how many	***k**awkanaksa*	***e**nni*	***k**ólko*	*xo.1*	*mấy*
how much	***k**awksa*	***e**nta*		*la.3 lai.6; jom.4 lai.6*	*bao nhiêu*
what	***k**unasa*	***e**em;**e**emi[Ti]*	***k**akvó; štó*	*en.3*	*gì*
when	***k**unarsa*	***e**ppuDu*	***k**ogá*	*hak.8; kek.8*	*khi nào*
where	***k**awkinsa*	***e**Ta; **e**edi; **e**kkaDa*	***k**ədé*	*cup.7*	*đâu*
which	***k**awkisa*	***e**evi*	***k**ój*	*h p.9 um.3*	*nào*
who	***k**itisa*	***e**waru*		*x n.3 cep.9*	*ai*
why	***k**unatsa*	***e**n[du]ceeta; **e**nduku*	*zaštó*	*mai.3 ca .6*	*tại sao*
E_f initial segment	0.0	0.0	0.27	0.84	1.0

Table 1. Examples of question words in different languages. Question words in Aymara, Bulgarian and Telugu 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 (Haspelmath & Tadmor, 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, homophones 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 a counteraction to the 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. This pressure goes against the pressure for distinctiveness.

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 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).

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. Put another way, many other groups of words, for example basic colour words, could have undergone the same changes to bring about similarities in form, but this tends not to happen in order to maintain distinctiveness.

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 content question words can be detected, and whether this is independent of historical contact between languages.

Meaning	**Word**	**Gloss**	**Analyzability**
how many?	ikutsu	iku-tsu *some-CLASS*	analyzable derived
how much?	ikura	iku-ra *some-PL*	analyzable derived
how?	dō		unanalyzable
what?	nani		unanalyzable
when?	itsu		unanalyzable
when?	nanji	nan-ji *what-hour*	analyzable derived
where?	doko	do-ko *Q-place*	semi-analyzable
which?	dono	do-no *Q-ATTR*	semi-analyzable
which?	dore		unanalyzable
who?	dare		unanalyzable
why?	dōshite	dō-s-ite *how-do-CONV*	analyzable phrasal
why?	naze	nani-semu-ni *what-do-ADV*	semi-analyzable

Table 2: A list of question words in Japanese with analyzability. Data from Data from the World Loanword Database (Haspelmath & Tadmor, 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 Spraakbanken 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). The final dataset included only languages with at least 5 out of 9 interrogative words (*how, how many, how much, what, when, where, which, who, why,* these are all separate concepts according to the IDS concept list*)*. If a particular language had two or more words referring to one of these question words, they were all taken into account.

The final data set for analyses consisted of 172 languages which come from 65 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 7% of possible entries had no data, 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 words in the corpus (934 concepts). IDS (and by inheritance WOLD and Spraakbanken) 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 132 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 49 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[1].

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 31 languages.

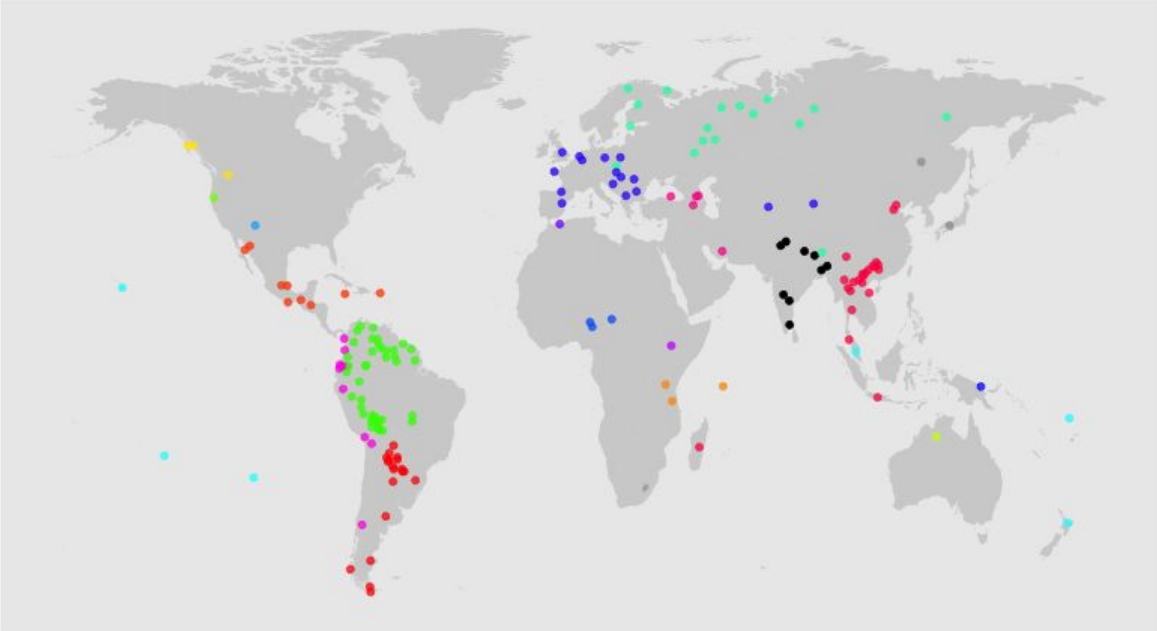

Figure 2. Distribution of languages in the study, coloured according to geographic contact areas.

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

[1] 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.

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, … x_n, the probability of observing a given segment type x_i is $p(x_i)$, and the entropy efficiency is calculated as:

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

(where b represents the log base, e.g. 2)

The following examples are provided to clarify the notion of entropy. In Aymara language all interrogative words start with a phoneme /k/. In such case there is no uncertainty within this set of words. Entropy therefore is equal to 0 because the probability of the word starting with /k/ within this particular set is absolute. On the other hand, in Bulgarian all but one interrogative words starts with /k/ (the /k/ phoneme is coincidental). In this case Ef = 0.16, still low but higher than for Aymara. In Dehong, all 9 interrogative words start with a different phoneme except two words, which start with /t/. Accordingly, Ef = 0.93. A language where all interrogative words start with a different phoneme would have Ef = 1.

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 mean entropy scores for each 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). So the Spanish word for “who” might be swapped with the Dutch word for “who”. The entropy score for each language is recalculated, and the mean of these values is 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.

True data

English	Latvian
haw	ka:
haw mɛni	tsik
haw mətʃ	tsik
wət	kas
wɛn	kad
wɛr	kur
wɪtʃ	kurʃ
hu	kas
waj	'ka:pe:ts
E_f = 0.31	E_f = 0.24

Permuted data

English	Latvian
haw	ka:
tsik	haw mɛni
haw mətʃ	tsik
kas	wət
wɛn	kad
wɛr	kur
kurʃ	wɪtʃ
hu	kas
waj	'ka:pe:ts
E_f = 0.60	E_f = 0.52

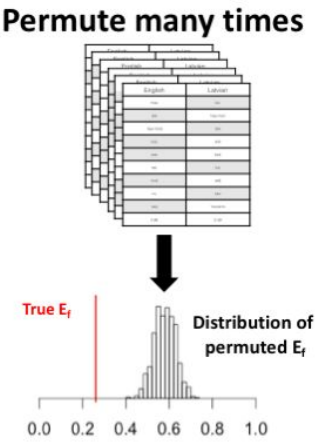

Figure 3: Demonstration of permutation method. The true data (left) is analyzed, calculating the entropy efficiency E_f of the first segment of words within each language. The data is randomly permuted (middle) and the entropy efficiency is calculated again. If words are more similar within languages 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 Indo-European language family (Cysouw, 2004). 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 entropy of first segments (E_f) 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. The entropy 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	0.438	Free permutation	0.78	< 0.0001	-64.49
		Permutation only within language families	0.57	< 0.0001	-31.29
		Permutation only within linguistic areas	0.71	< 0.0001	-52.93
		Permutation only within language families and areas	0.54	< 0.0001	-26.65
Initial segments of unanalyzable interrogative words	0.549	Free permutation	0.75	< 0.0001	-10.65
		Permutation only within language families	0.7	< 0.0001	-4.99
		Permutation only within linguistic areas	0.7	< 0.0001	-4.86
		Permutation only within language families and areas	0.69	0.0004	-3.39

Table 3: Permutation test results comparing the entropy of interrogative words within languages to various baselines.

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 supporting information S3 for full results), these words are also more similar within a language than between languages: same semantic domain (Ef = 0.78, z = -1.99, p = 0.04); body (Ef = 0.72, z = -12.4, p < 0.0001), basic actions (Ef = 0.75, z = -11.3, p < 0.0001) and pronouns (Ef = 0.65, z = -15.61, p < 0.0001). This is probably driven by the differences in phonologies between languages, raising the possibility that the question words are not special. However, the Ef 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 Ef to the permuted distributions. In the next study, we test whether these differences are significant.

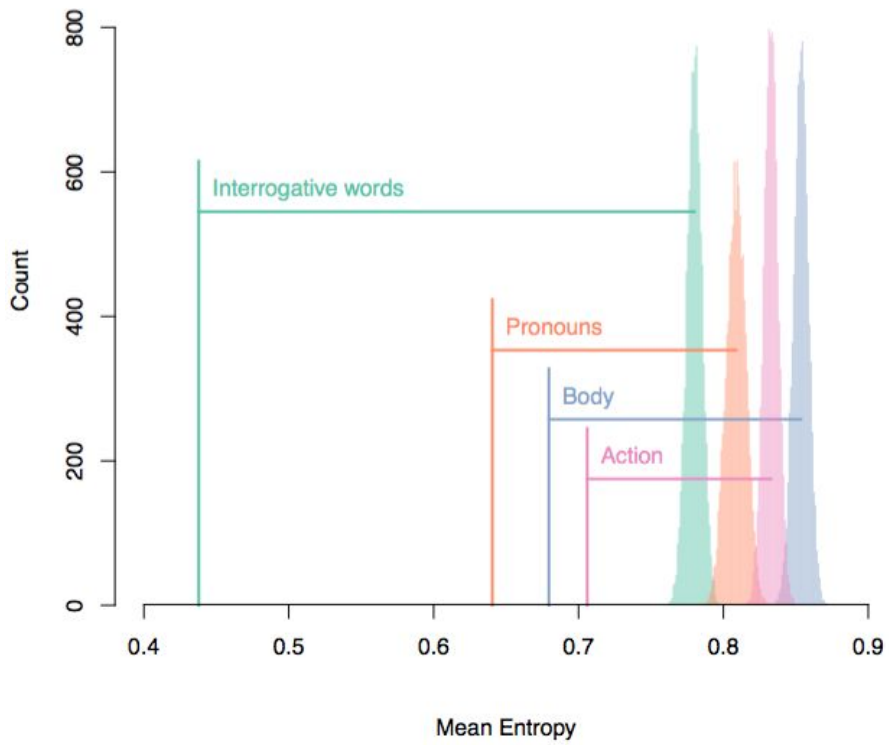

Figure 4: Results of study 1. Vertical lines show the mean entropy efficiency of a group of words (lower values indicate that words are more similar). The horizontal lines connect these values to the distribution of mean entropy efficiency when those words are permuted. For all groups of words, the mean 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 question words themselves should 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 sets of 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.

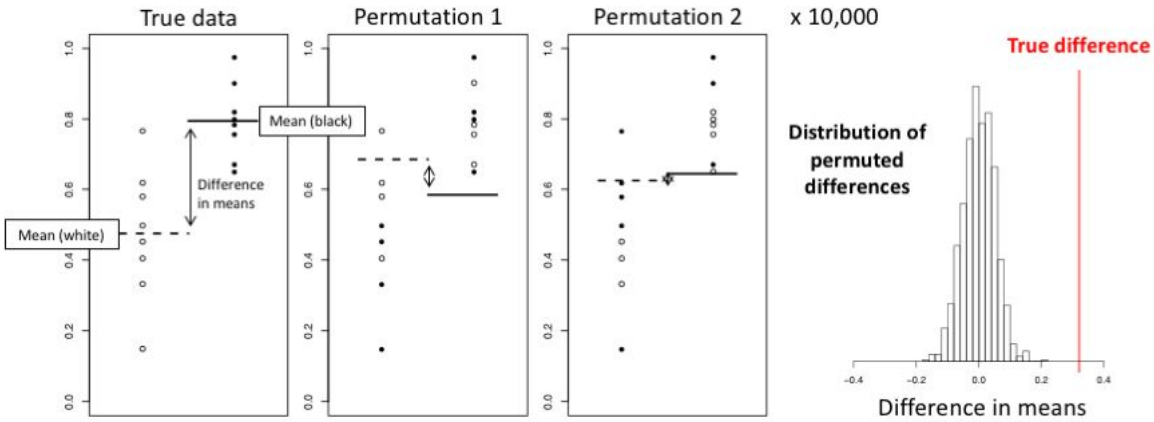

Figure 5: Demonstration of a permutation test of the difference between two groups. The true data has 8 data points in two groups (white and black). The difference in means between the two groups is calculated. When permuting the data (middle), the values are kept the same, but the membership to the 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 - head concepts, basic actions and pronouns.

Sample	Mean E_f	Baseline	Mean E_f	*p*	*z*
Initial segments of interrogative words	0.439	Random words	0.78	< 0.0001	28.1
		Conceptually related words (20 sets)	0.75	< 0.000005	15.09
		Body concepts	0.68	< 0.0001	8.85
		Basic actions	0.70	< 0.0001	9.88
		Pronouns	0.63	< 0.0001	7.96
Initial segments of unanalyzable interrogative words	0.549	Random words	0.84	< 0.0001	10.43
		Conceptually related words	0.72	0.04722	1.53
		Body concepts	0.79	< 0.0001	3.83
		Basic actions	0.81	0.0002	3.54
		Pronouns	0.68	0.029	1.88

Table 4: Permutation test results comparing the entropy of interrogative words within languages to random and conceptually related words. N = Number of languages in the sample.

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 some examples.

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 …), 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:

$$\text{Detectability} = \frac{\prod_{i=1}^{c} \binom{K_i}{k_i}}{\binom{N}{n}}$$

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

161 out of 172 languages had interrogative words with initial segments that were more detectable than randomly selected words, and that this was significant for 117 languages ($p < 0.05$, compared to 10,000 randomly chosen sets of words for each language), though the effect size is small (mean $z = -0.54$).

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 Study 1 and Study 2. In addition we gathered typological data on the languages' positioning of the interrogative phrase. The World Atlas of Language Structures lists data for 71 languages in our sample (Dryer, 2013). Thirty-five languages obligatorily place question words at the beginning of clauses, 33 languages do not and 3 languages have different strategies for different question words (the latter were excluded). We coded a further 38 languages following the coding scheme of Dryer (2013). 29 were initial, 7 were non-initial and 2 had mixed strategies (languages with mixed strategies were excluded). See the Supporting Information S1 for details including sources and reliability coding procedure. The final analysis included 64 initial languages and 40 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.

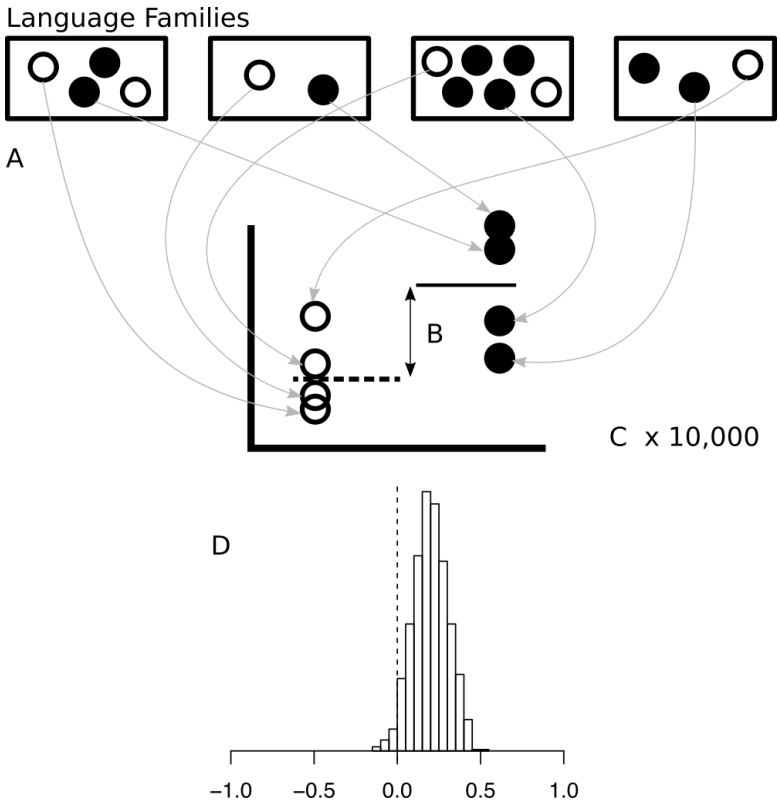

Figure 6: Demonstration of an independent samples test. Languages are represented as circles , with white circles being initial interrogative languages and black circles being non-initial interrogative languages. One language is randomly chosen from each language family for each group (A). The mean for 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 significantly lower than non-initial interrogative languages (initial interrogative words have more similar initial segments) when controlling for language family, but not significant when controlling for geographic area. 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 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	64	40	20000	-0.1	0.047 *	1.69
	Restricted by area	64	40	20000	-0.06	0.201	0.84
Unanalyzable interrogative words	Restricted by family	12	8	10000	-0.28	0.003 *	2.93
	Restricted by area	12	8	10000	-0.28	0.004 *	2.83
Random words (50 sets)	Restricted by family	64	40	47507	-0.01	0.393	0.27
Conceptually related (20 sets)	Restricted by family	64	40	933282	-0.02	0.353	0.39
Conceptually related: Head	Restricted by family	64	40	20000	-0.12	0.010 *	2.22
	Restricted by area	64	40	20000	-0.09	0.096	1.28
Conceptually related: Basic actions	Restricted by family	64	40	20000	-0.03	0.297	0.55
	Restricted by area	64	40	20000	-0.02	0.360	0.39
Conceptually related: Pronouns	Restricted by family	64	39	20000	0.02	0.690	-0.48
	Restricted by area	64	39	20000	-0.02	0.374	0.33

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 Ef is negative, then the initial interrogative languages had a lower Ef (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.

A straightforward permutation test between the detectability z-scores found that initial interrogative languages had more detectable interrogative words than non-initial languages (mean detectability z-score for initial = -0.64, mean for non-initial = -0.44, 10,000 permutations, $p = 0.0342$). When selecting random samples from independent language families, the initial interrogative languages had more extreme detectability z-scores than non-initial interrogative languages ($p = 0.0047$). However, when selecting random samples from independent geographic areas, the result was only marginally significant ($p = 0.0855$).

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 a 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 = 0.66, chisq = 1.31, df = 1, $p = 0.25$). This suggests that, regarding detectability, the difference in the permutation test is driven by historical or areal confounds.

7. Additional analyses

Some additional analyses were done, considering all segments of words instead of just the first 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 at 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 at first consonants (mean difference = 0.15, $z = 2.37$, $p = 0.01$) but this was marginal when considering first vowels (mean difference = 0.07, $z = 1.67$, $p = 0.05$). 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 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 found that question words are more similar on average in languages that place them at the beginning of clause, though the effect size was small and this did not hold when controlling for geographic area. This was not true for other concept sets, except for body concepts which showed the same pattern as question words. Also, initial interrogative languages were more likely to have more salient question words than non-initial languages, but the difference was not significant in all tests.

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.). However, it's plausible that the pressure for action ascription also had 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. The results suggest that question words sound more similar than would be expected by chance, and there is a trend for this to be exaggerated in initial interrogative languages. This contrasts with previous assumptions that formal regularities within question words are not present (Cysouw, 2004) and support 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 the 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 also 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.

Perhaps the most surprising finding is that question words sound more similar in languages which put them at the beginning of turns than in languages that do not. This is a direct prediction of the interaction theory. However, the finding was not robust to all tests - the difference was not significant when controlling for areal effects, and body concepts also showed the same pattern as question words. Alternatively, these might be explained due to correlations with wider properties of syntax and morphology. 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), 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; the Automated Judgement Similarity Program database, Whichmann & Holman, 2016, has an order of magnitude more languages but an order of magnitude fewer concepts and no 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). 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 entirely 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 (see table 6). For example Quechua languages make use of particle *taq* to identify question words (Cerron-Palomino, 2008). Initial segments of question words, however, differ among each other. On the other hand, Aymaran languages makes use of the particle *sa* at the end of the question words, but the similarities of also initial segments can be observed (see also 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), so it is plausible that additional variance was introduced in our data due to the fact that we account only for systematicity within initial segments of the question words. 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.

Language	Example	Translation
Quechua	pi-**taq** hamu-rqa-n?	who came?
Aymara	khiti-**sa** juta-ya-na?	
Quechua	ima-ta-**taq** muna-nki?	what do you want?
Aymara	kuna-**sa** mun-i?	
Quechua	may-pi-**taq** tiya-saq?	where will I live?
Aymara	kawki-na-**sa** utja-nja?	
Quechua	pi runa-**taq** wañu-rqa-n?	which person died?
Aymara	khiti jaqi-**sa** jiwa-ya-na?	

Table 6. Questions in South-American languages Quechua and Aymara.

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 grammaticlisation 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, hwȳ) than modern English. Further research is required on this front.

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] 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: https://github.com/seannyD/UniversalsInWHWords

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. Columns indicate the meaning id as specified in the original databases, a gloss of the meaning, the word in the original database, the source database, the language the word belongs to, the iso and glotto codes of the language, the borrowability score from WOLD, the analyzability from WOLD, the general semantic domain, the word cleaned of typographic irregularities, the word converted to the simplified phonology and the meaning id with some fixes by the authors. Data include 150,770 words, 943 meanings in 172 languages.

S3 - Results for alternative analyses.

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

References

Berlin, B. 1994. Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nichols & J. J. Ohala (eds.), Sound symbolism, 76 –93. Cambridge: Cambridge University Press.

Bickel, B., Nichols, J., 2007. Inflectional morphology. In: Shopen, T. (Ed.), Language Typology and Syntactic Description, 2nd edition, vol. III. 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 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.

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-areal-linguistics/word-lists

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, 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

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 (EVOLANG11). 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

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

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.

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 (EVOLANG11). 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, William A. (1991). The Yimas Language of New Guinea. Stanford: Stanford University Press.

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

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 University Press.

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 on 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 (EVOLANG11). 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.

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, pages 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). Turn-taking in human communication, origins, and implications for language processing. Trends in Cognitive Sciences, 20(1), 6-14.

Levinson, S. C. (2016). Speech acts. In Y. Huang (Ed.), Oxford handbook of pragmatics. Advanced online publication. Oxford: Oxford University Press.

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. (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

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., & 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.

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.

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.

Parker, S. (Ed.). (2012). *The sonority controversy*. Walter de Gruyter.

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, 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. & 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

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., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. PLoS one, 8(8), e70902.

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.

Segerer G., Flavier S., 2016 RefLex: Reference Lexicon of Africa, Version 1.1. Paris, Lyon. http://reflex.cnrs.fr/

Schegloff, E. A. (2007). Sequence organization in interaction: Volume 1: A primer in conversation analysis (Vol. 1). Cambridge University Press.

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.

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.

Ultan, R. (1969). Some General Characteristics of Interrogative Systems. Working Papers on Language Universals, No. 1.

Ultan, Russell (1978). Some general characteristics of interrogative systems. In: Joseph H. Greenberg (ed.) Universals of Human Language. Vol. 4: Syntax, pp. 211-48. Stanford: Stanford University Press.

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, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2016. The ASJP Database (version 17).

Zipf G. (1949) Human Behavior and the Principle of Least Effort. New York: Addison-Wesley