ITRW on Disfluency in
Spontaneous Speech (DiSS'01)
Edinburgh, Scotland, UK
August 29-31, 2001

Sound and Function Regularities in Interjections

Nikolinka Nenova, Gina Joue, Ronan Reilly, and Julie Carson-Berndsen

Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
{ nikolinka.nenova, gina.joue, ronan.reilly, julie.berndsen }@ucd.ie

Abstract

This paper investigates the relation between the sound patterns of interjections and their functional realisation in the discourse process. It considers whether certain interjection functions tend to have particular sound distributions. In order to address these questions a classification scheme for American English nonlexical interjections in terms of discourse markers is also presented.

1. Introduction

In the attempt to create a robust and relevant computational model for spontaneous speech interaction, speech system projects have only recently begun to consider dysfluencies as functional devices in the process of communication [1]. Save for the few instances in which interjections are analysed as part of the reparandum [2] or mentioned as back channelling moves [3,4], the contextual richness of interjection function has been hardly discussed [5,6]. Researchers also have casually but consistently noted that nonlexical interjections in different languages share phonetic similarities. For example, nonlexical interjections in English, Swedish [7] and Spanish [8] commonly involve infrequent or illegal phonotactic combinations. In a study involving Icelandic, English, Polish, Hungarian, Finnish, Ososo, Malagasi and Slovenian interjections, Abelin [7] noted that interjections in all these languages involve mostly labial or alveolar sounds. However, again, the phonological tendencies of nonlexical interjections have not been properly investigated.

This work contributes to filling in some of these functional and phonological gaps and demonstrates the sound regularities and the functional importance of nonlexical interjections in discourse. In this paper we contend that the sound patterns of interjections are dependent on their function, propositional meaning and position (both physical and contextual) in which they are realised in the discourse process.

Section 2.2 presents the phonological paradigm we used for the functional analysis of the constraints influencing the phonology of interjections. Then follows the analyses themselves and the discourse notions on which these analyses were based. The last section evaluates the analyses in the context of the suggested hypothesis.

2. Phonology of Nonlexical Interjections

We approached our investigation for a functional explanation of the constraints influencing nonlexical interjection based on Phonology as Human Behavior (PHB) [9,10]. PHB is a cognitive approach to phonology. Its aim is not simply to describe the systematic distributions in the sound structure of a language but also to explain these patterns. Appealing to functional and semiotic explanations, PHB purports that these patterns are directly shaped by the synergetic interactions of communicative and human physiological/behavioral constraints. That is, sounds in languages are not random because the (sometimes) conflicting goals of minimising articulatory effort and of maximising communication will tend to favour certain sounds over others. For example, most pause fillers are made up entirely by vowels (e.g. *uh, ah, oh*) as vowels require less effort to articulate than consonants. Distinctions among voiced vowels, however, become much more difficult (much subtler) with the increased number and variety of vowels which need to be distinguished in a phonological system. Therefore the speaker may have to increase efforts to enhance communication as vowels alone are limiting. Thus, although consonants are more difficult to articulate, they provide greater distinctions needed between vowels. Certain consonants and certain vowels will be more common than others. For example, consonants involving the lips and the tip of the tongue are easier to produce (and the lips being more visual so easier to perceive); therefore, these consonants occur most frequently in interjections across languages.

2.1. Interjection sound pattern hypotheses

Such a paradigm leads to a few hypotheses and explanations about the sound structure of interjections. For example, it supports Abelin's [7] observation that pause fillers tend to involve sounds produced by either the lips or the apex of the tongue, depending on their discourse function.

We hypothesise that interjections which signify **static** functions, that is those that do not change the current belief or knowledge of the participants or the intentional direction of the discourse moves (but merely indicate the speaker's attendance in the conversation, for example), will overall be much simpler and vary less phonetically than interjections indicating more **dynamic** participation. In other words, static-function interjections will most likely involve the most easily articulated sounds, which entails a more limited phonetic inventory, very simple syllable structures and most likely monosyllables. This hypothesis is motivated by the assumption that dynamic-function interjections indicate a speaker's willingness to increase articulatory effort for greater communicative holds and to produce particles with greater perceptual distinctions (or marked sounds). Likewise, static-function interjections imply more reluctance for too much articulatory effort or the avoidance of too salient sounds (or unmarked sounds).

3. The Analysis

The hypotheses outlined in the previous section, were formed to answer the following questions:

- Is there any significant difference in the sound distribution of the interjections in relation to their position in turns?
- Do certain interjection functions tend to particular sound distributions?

In order to test these hypotheses, we created a functional taxonomy for interjections that was simple enough for computational purposes but which also sufficiently captured the functions of interjections as discourse markers. We analysed the set of all interjections that were encountered in the TRAINS 91 corpus [11] based on this taxonomy. Although we did not have any phonetic transcriptions of the interjections, we assumed that the orthographic transcription of interjections are faithful to general English sound spelling rules and broadly examined them with the principles of PHB.

3.1. The choice of corpus

As was mentioned above, early research in spoken language systems filtered interjections as irrelevant to the process of communication. That is why most speech corpora transcribed for computational analyses have ignored interjections in transcription or were inconsistent in their transcription. This problem restricted our corpus choice to the TRAINS 91 dialogues. The TRAINS corpus provides orthographic transcriptions of the variations (e.g. *ohhh*) of interjection baseforms (*oh*) to approximate the actual token articulation of the given interjection. The transcription also includes overlapping speech, which, for example, was unfortunately not the case with the phonetically transcribed portion of the Switchboard corpus. The corpus is a collection of 16 task-oriented Wizard of Oz dialogues. The dialogues were approximately 80 minutes in length and included a balanced number of male and female American English speakers.

3.2. Function taxonomy

We view interjections as discourse markers, that is, the functions that they complete are based on the factors that constrain the discourse process. Three factors we identified are the information **direction** (new vs. old information), the **relation** or the hierarchical interdependency between the utterances in the dialogue (main vs. sub topic), and the participants' **intention** and expectations (what the move was intended as vs. what it was implemented as).

- The **direction** shows how the information currently presented is related to the one that has been already exchanged. When the utterance is related to a discussed topic, the direction is backward.
- The **relation** refers to the contextual position of the current utterance in the overall discourse hierarchy. It is a term that shows the focus of what *is being* said to what *has been* said. Relation realizations can be *start*, *finish* or *expansion*.
- The **participants' intentions** towards the dialogue move refer to the speaker's intention for the effect, which the current utterance would have on the other participant. When a speaker produces a move they expect this move to be responded to by a particular move or set of moves from the other participant. In our analyses this is further generalized to represent whether the utterance is intended towards the speaker themselves or the hearer. It specifies whether the utterance is a comment on current self-knowledge of the speaker or the current knowledge of the hearer. Participants' intentions may be *subjective*, where the utterance is an evaluation of self-knowledge; or they can be *objective*, which refers to evaluation of the other participant's knowledge. We also considered an additional factor: the participants' degree of evaluation of the ongoing discourse process. The degree of evaluation can be *positive*, *negative* or *neutral*. This factor is applied only to one of the functions (see Table 1)

Table 1: Function taxonomy

Function	Direction	Relation	Intention	Evaluation
Acknowledgement (Ack)	backward	finish	objective	neutral (AA), positive(AP), negative(AN)
Expansion (Exp)	forward	start, expansion	subjective (ES), objective (ER)	
Correction (Corr)	backward		subjective (CS) objective (CR)	

The interaction among these three factors establishes the three basic discourse functions (see Table1) (as opposed to syntactic or semantic). In this work we considered these functions to constrain the inference and intentional structure of discourse.

This taxonomy was used to annotate the nonlexical interjections in the TRAINS corpus.

3.3. Interjections in TRAINS

We first identified base forms by their relative high frequencies in contrastive **functional realisations**, that is, their functions (as based on the taxonomy in Table 1) and locations in turns. In correlating the variations of the interjections to their baseforms, articulatory or sound similarities are insufficient criteria because interjections with different functional realisations often have close sound structures. We found that the patterns of frequencies in functional realisations, in addition to sound proximity, provided a reliable method of identifying variants of baseforms even when frequencies were very low for some. Table 2 lists the interjections and their variants in decreasing order of frequency. Items in parentheses indicate very low frequency. Items in italics are sound synonyms.

Table 2: Non lexical interjections and their variants in TRAINS (Location: 0=constitutes turn, 1=at the beginning of turn, 2= within turn, 3 = at the end of the turn)

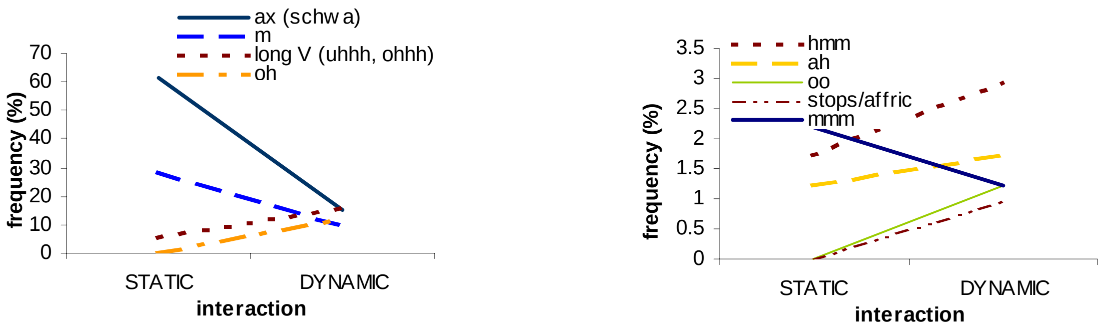
baseforms	variants	functions	locations
ah	(*hmm*)	Ack, Exp, Corr	2, 1
(aha)		Corr	2
(eh)		Corr / Exp	1 / 2
(err)		Corr, Exp	2,1
m-hm		Corr	0, 1
uh	uhh, uhhh, *uhm*, (*uhhm*), uhmm	Exp, Corr, Ack	2, 1, 3, 0
um	*mm*, umm, uumm, *uhm*, (*uhhm*)	Exp, Ack, Corr	2, 1, 3, 0
uh-huh		Corr	0, 2
hm	(*hmm*), m, (*mm*)	Exp, Ack	2, 1
oh	(o), (ohh), (oo), (ooh), (oooh)	Corr, Ack, Exp	1, 2, 3
(oops)	(whoops)	Corr	0, 1 / 2
[ouch]	(uch)	Corr	0
wow		Corr	1

Our analyses provide some support that interjections are context dependant and that their function is a combination of their position, their propositional meaning and the context in which they appear.

3.4. The relation between interjection position and function

Results show that the most frequent position is within the body of the utterance; however, most of these were self-expansion interjections (Exp). They show that the current speaker intends to further expand the utterance by contributing more information. The least frequent position is at the end of the utterance. Therefore, in general, interjections appear to prepare the listener in predicting the following utterances. The 2% that occur at the end are predominantly interjections, which speakers use for self-expansion (indicating an intended beginning of a turn) but were interrupted by the listener.

Figure 2: Relation of marked/unmarked sounds with strength of interaction



The second most frequent function of interjections is as indicators of change (Corr). The change includes self-repair or self-realisation (the most frequent in that type). The change of the direction of the information usually indicates that there is an update of the knowledge, or a change in the current state of the world or of the current topic in the communication. Like the general trend of interjection positions, these types of interjections tend to appear in the body of the turn; however, this case usually occurs at the beginning of a new utterance within the turn. The least frequent function of the three is that of acknowledgement (Ack).

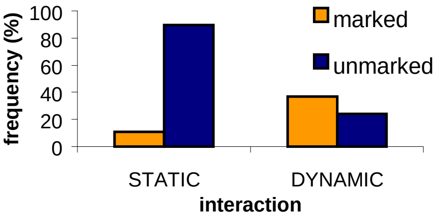
3.5. The Phonetic Analysis

In order to test our hypotheses of the phonological properties of nonlexical interjections (Section 2.1), we classed AA, ES, and ER functions (see the taxonomy in section 3.2) as indicators of more static interaction, and the rest as more dynamic. We identified marked sounds depending on

- the complexity of syllabic structure and
- the acoustic/articulatory salience of the sounds making up the interjection.

The schwa is the most central position of the vocal tract for an American English speaker, and the closed lips the most neutral static position for no utterances; and not involve sounds of more effort such as very lengthened vowels or nonsonorant consonants. We took /m/ and the schwa to be unmarked sounds in American English. where marked of course is relative to the specific language's sound inventory. Marked sounds are rounded (e.g. *oh*), lengthened vocalisations (long vowels, *mmmmm*), noncentral or tense vowels, and nonsonorants (such as stops).

Figure 1: Relation marked/unmarked sound patterns in relation to interaction strength



In support of our hypotheses, results confirm that the syllable structure of interjections indicating static interaction tended away from multisyllabic forms (0.5% multisyllables) more than the dynamic interaction ones (3.4% multisyllabes). Results also show skewings indicating that the degree of markedness in the sound makeup of interjections relates directly to the degree of interaction (see Figure 1), as we also expected.

As seen in Table 2, our method of classifying marked/unmarked uncovers a direct relation between markedness and the degree of interaction in the discourse process. Figure 2 shows that sounds with less acoustic energy (such as *m*) tend to indicate less dynamic interactions (so they are mostly within turns) than those with more acoustic energy such as long vowels.

Almost all the interjections in the TRAINS corpus were monosyllabic (97.6%), as has been commonly observed in interjections across languages. The results show (Figure1) that static-function interjections (ES, ER, AA) tend to have less complex sound structures and less marked sounds than dynamic-function ones (Corr). Specifically, static-function interjections involve mostly the unmarked sounds: schwas and /m/s.

Likewise, in the group of the dynamic-function interjections include bisyllabic forms although, these are reduplication or minor variations of a very simple syllable structure. They also involve less of the perceptually weaker sounds such as schwas and more “marked” sounds. The lengthened forms *mmmmm* and *hmmmm* are neutral (such as giving the other participant a chance to interrupt) but also indicate more dynamic participation (and hence are at the beginning or at the end of turns).

Another example for the interrelation of function and sound choice are *mm-hm* and *uh-huh.* Both usually indicate more dynamic interactions. Thus, it is not surprising that they

are bisyllabic and are almost syllable reduplications. The /h/, however, also acts to increase the perceptual distinction of the second syllable from the first; without the aspiration, the speaker would have to place a pause between the *mm* syllables or a glottal stop between the *uh* syllables to ensure the perception of two syllables. Perhaps the additional syllable complexity is also balanced by the fact that both *mm-hm* and *uh-huh* involve the most neutral (least complex) sounds: /m/ and /ə/. Although /m/ is a labial nasal and thus more visual perceived and more naturally articulated, *uh-huh*, which involves a more open oral position and involving more effort function for more dynamic discourse purposes.

There were a few sounds whose frequencies were too low for us to draw any conclusions. However, as seen in Figure 2, our method of classifying marked/unmarked uncovers a direct relation between markedness and the degree of interaction in the discourse process. The only sound, which appeared not to match our predictions according to Figure 2, is the lengthened *m*, as we assumed that it is marked yet it appears more frequently in static interaction.

However, a difference does exist between lengthened /m/ and its shorter baseform. The lengthened /m/ occurs primarily at the beginning and the end of turns (thus marking the change in turns) whereas the shorter form occurs primarily within turns. This may imply that the sound structure of nonlexical interjections depend on *both* function and location and supports our hypothesis that marked sounds indicate more dynamic interactions.

4. Conclusions

In this paper, we analysed the relation between the phonetic structure and the pragmatic function that the interjections fulfil in the process of task-oriented communication. The consistencies in the sound structure of interjections in relation to their functional realisations lend support to the contention that interjections are discourse markers with functional and phonetic regularities. A stronger support of these regularities would be to conduct a cross-linguistic analysis on nonlexical interjections and investigate their dependencies on the language’s sound inventory.

5. References

[1] Heeman, P.A., Byron D. and Allen, J.F. Identifying Discourse Markers in Spoken Dialog. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, Stanford, March 1998, pages 44-51. 1998

[2] Shriberg, E.E. *Preliminaries to a Theory of Speech Disfluencies*. PhD Thesis, University of California, Berkeley. 1994.

[3] Core, M., and Allen, J., Coding Dialogues with the DAMSL Annotation Scheme. In *Workshop Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*. 1997.

[4] Traum, D. R., *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis. University of Rochester, New York. 1994.

[5] Meteer, M. Dysfluency Annotation Stylebook for the Switchboard Corpus unpublished.·1995.

[6] Fisher, K., *From cognitive semantics to lexical pragmatics: the functional polysemy of discourse particles*. Mouton de Gruyter, 2000

[7] Abelin, A. *Studies in Sound Symbolism*. PhD Thesis. Gotehnburg Monographs in Linguistics, Göteborg University, Sweden. 1999.

[8] Montes, R.G. The development of discourse markers in Spanish Interjections. *Journal of Pragmatics*, vol. 31, pp.1289—1319. 1999.

[9] Diver, W. The Theory. In E. Contini-Morava and B. Sussman Goldberg (eds.), *Meaning as Explanation: Advances in Sign-Oriented Linguistic Theory*, pp. 43-113: Mouton de Gruyer, 1995.

[10] Tobin, Y. Phonology as Human Behavior: Theoretical Implications and Clinical Applications. Duke University Press: Durham and London, 1997.

[11] Allen, J.F. and Schubert, L.K. The TRAINS project. TRAINS Technical Note 91-1, University of Rochester, Department of Computer Science, 1991.