# How to distinguish between self- and other-directed *wh*-questions?

*Katalin Mády, Uwe D. Reichel*

Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
{mady|uwe.reichel}@nytud.mta.hu

## Abstract

The most general aim of *wh*-questions is to seek for information, but they can have a wide range of other pragmatic functions. In this paper we investigate self-directed questions in dialogues that are lexicalised forms of vacillation (“how should I explain?”) and do not directly address the interlocutor. Their prosodic properties are compared with real *wh*-questions that seek for information.

**Index Terms**: *wh*-questions, self-directed speech, prosody, stylisation, Adaboost

## 1. Introduction

According to [1] most languages have three basic sentence types: declarative, interrogative, and imperative. For the interrogative type [1, p 160] point out as a first approximation that it “elicits a verbal response from the addressee. It is used principally to gain information”. In accordance to Searle’s question analysis [2] a question is an attempt to elicit information from the addressee the speaker wants to gain. [1] and many other researchers (e.g. [3, 4]) give numerous counter-examples suggesting a more fine-grained subdivision of interrogatives. Among these counter-examples are self-directed (“self-addressed” [3]) questions by which the speaker does not expect information from an addressee but is rather thinking aloud. Self-directed questions can be marked syntactically, e.g. in German by verb-last word order [4] (*“ob das wohl stimmt”?* – ‘whether it is true?’).

According to [3], self-directed questions do not request an answer, instead, they express the status of the speaker. In the example *“Now why did I say that?”* the speaker verbalises her surprise about her own utterance. A different view is provided by [5] who investigate self-directed queries in connection with disfluency signals. They claim that in this case, the speaker is the “addressee” of the query, since he/she can straightforwardly answer their own question.

In this paper *wh*-questions that act as verbalised vacillation are investigated. In these cases, speakers use self-directed questions to gain time to collect their thoughts and find a better way to explain something to their partner. As opposed to real questions, these questions do not aim at encouraging the interlocutor to be cooperative, instead, they can be characterised as offtalk. The goal of the paper is to compare prosodic features of real *wh*-questions that seek for information and require cooperativeness from the partner to self-directed questions that primarily signalise vacillation on the speaker’s side. Based on Ohala’s frequency code concept [6] we expect higher energy, higher f0 level and range values as well as more pronounced local f0 shapes for the interlocutor-directed than for the self-directed questions.

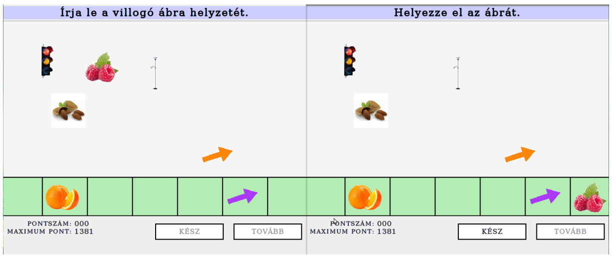
Írja le a villogó ábra helyzetét.
Helyezze el az ábrát.

*Figure 1: Image of the objects that appear on the screens of the two players. Left: screen of the describer with the raspberry blinking, right: screen of the follower who is supposed to place the raspberry into the position explained by the first player. Instruction left:* Describe the position of the blinking object, *right:* Drag the object into the correct position.

## 2. Data

### 2.1. Corpus

Data are taken from the Hungarian version of the object game of the Columbia Game Corpus [7]. It is a computer-aided game with two participants. Participants use separate laptops, and they do not have visual contact with each other. The players see objects on their screen that are identical except for one object that is blinking on the screen of one player, while it is located in the lower part of the screen of the other player. The first player describes the position of the blinking object in relation to the other objects that are placed on the screen of the second player in the same position. The second player is supposed to place the object in exactly the same position. Participants get a score after each turn on a 0 to 100 scale. Their roles alternate in the course of the game, so that both speakers are describers in half of the altogether 14 turns. Figure 1 shows the objects from a turn as were shown on the the two screens.

In the Hungarian version of the game, players formed 4 triplet groups, and they played two games with partly different, partly identical objects with both other members of the group (A with B, B with C, C with A). They were payed for their participation. Additionally, the group that scored highest was promised additional payment, in order to enhance the accuracy of the descriptions. Participants within a group were familiar with each other (relatives or close friends), which lead to a high degree of naturalness during the task.

The corpus is currently being annotated among others for dialog acts. The current version of the paper presents first results on self- and other-directed *wh*-questions that were manually segmented and labelled. All interrogatives began with a *wh*-word that carries an accent in Hungarian as a default. Self-directed questions did not differ from other-directed questions in their syntax. One self-directed question contained a lexical unit that would be improbable in a real question: “And this is located between the traffic light and the standard lamp. In addition, how is it located?”. Another self-directed question expressed that the describer has difficulties to express himself: “Ow, how should I tell you?” The lexical form of the remaining self-directed questions was identical with potential string-identical other-directed questions.

### 2.2. F0 extraction and preprocessing

Fundamental frequency (f0) was extracted by autocorrelation (Praat 5.3, sample rate 100 Hz, [8]). Voiceless utterance parts and f0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering [9] using third order polynomials in 5 sample windows and transformed to semitones relative to a base value. This base value was set to the f0 median below the 5th percentile of the speaker’s f0 within the entire dialog and served to normalize f0 with respect to its overall level.

## 3. Prosody stylisation

### 3.1. Parameterisation

Within the utterance chunks three types of features were extracted: (1) f0 register features, (2) local f0 movements on the *wh*-word, and (3) energy. The features are listed in table 1. As register features we measure the f0 level and range starting points, trends and mean values. For this purpose a base- mid- and topline were fitted to the chunk as illustrated in the left half of Figure 2. As described in greater detail in [10] this fitting method does not depend on fuzzy f0 peak and valley detection but consists of three linear regressions through local f0 median values in the lower, mid and upper f0 range. As shown in [10] this method therefore is less error prone and more robust against the influence of local pitch events. The level trend within the chunk is defined as the midline slope. The range trend is defined as the slope of the regression through the pointwise distances between top- and baseline. These linear level and range stylisations are shown in the right half of Figure 2.

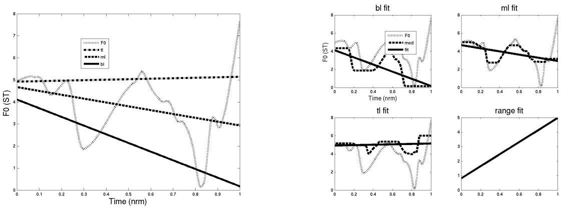

Figure 2: **A (left):** *Stylisation of base-, mid- and topline based on F0 median sequences below the 10th percentile for the baseline, above the 90th percentile for the topline and for all values for the midline. The F0 range is represented by a regression line fitted through the pointwise distances between the base- and topline.* **B (right):** *Base-, mid-, topline and linear range stylisation results.*

Next to the global f0 register variables we parameterized the local f0 movement of the stressed first syllable on the always chunk-initial wh-word by a third-order polynomial. The 30 ms window was placed on the vowel midpoint, the left half limited by the chunk onset. Within that window time was normalized from $-1$ to 1 with 0 placed on the vowel midpoint. Figure 3 shows the decomposition of a local f0 movement by a third order polynomial.

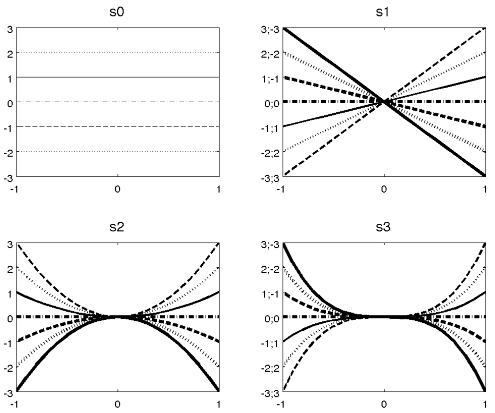

Figure 3: *Influence of each coefficient of the third order polynomial $t = \sum_i s_i \cdot t^i$ on the contour shape. All other coefficients set to 0. For the purpose of compactness both function and coefficient values are shown on the y-axis if they differ.*

Finally, energy was measured by RMS over the entire speech chunk.

### 3.2. Feature weights

Table 1 summarizes the examined features and their discriminatory power to hold apart self- and other-directed questions. These weights $w$ for features $i$ are derived from the Silhouette measure usually used for cluster validation as follows:

$$w(i) = \frac{\frac{\sum_{j=1}^{n} S(j)}{n} + 1}{2},$$

where the silhouette $S(j)$ measures for each of the $n$ data points – i.e. for a feature vector – $j$ how well it can be assigned to one of the classes *self-* and *other-directed.* More precisely

$$S(j) = \frac{d_B(j) - d_A(j)}{\max(d_A(j), d_B(j))}.$$

$d_A(j)$ stands for the mean squared Euclidean distance between vector $j$ and other vectors of the same class. $d_B(j)$ stands for the mean distance between vector $j$ and vectors of the other class. Adding 1 and dividing by 2 transposes the weight range to the interval $[-1\ 1]$.

## 4. Results

Figure 4 shows the values of the examined parameters for other- and self-directed questions. A visual inspection reveals that the difference between these question types is primarily quantitatively but not qualitatively expressed, i.e., there is a difference in the absolute values but not in the algebraic sign. As an example, for both other- and self-directed questions there is a falling local f0 movement on the accented syllable (negative $c_1$) which is more pronounced in other-directed speech. Generally absolute values are higher in other-directed speech indicating a more pronounced usage of intonation. 2-sided Wilcoxon tests on the absolute values reveal significant differences for *ml_slope, ml_icpt,*

Table 1: *Prosodic features and their weights in terms of mean Silhouette normalized to sum 1. Weights were calculated for absolute feature values.*

| Feature | Description | Weight |
|---|---|---|
| | Register | |
| ml_slope | f0 midline slope | 0.5905 |
| ml_icpt | f0 midline intercept | 0.5781 |
| ml_mean | f0 midline mean | 0.5650 |
| rng_slope | f0 range slope | 0.3981 |
| rng_icpt | f0 range intercept | 0.5895 |
| rng_rms | f0 range RMS | 0.5718 |
| | Pitch accent | |
| $c_3$ | cubic polynomial coefficient | 0.4182 |
| $c_2$ | quadratic polynomial coef | 0.6074 |
| $c_1$ | linear polynomial coef | 0.4783 |
| $c_0$ | offset from midline | 0.4511 |
| | Energy | |
| $en$ | signal RMS over chunk | 0.6437 |

*rng_rms*, $c_3$, $c_2$ ($p < 0.05$), and *en*, and tendencies for *rng_icpt*, $c_1$, and $c_0$ ($p < 0.1$). For *ml_mean* and *rng_slope* no significant differences were found. However, the boxplots suggest that additional data will move the differences towards significance.

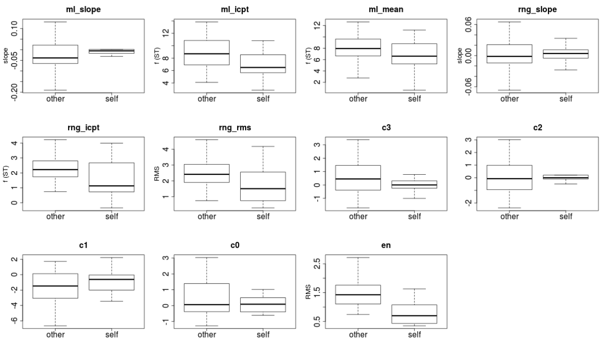

Figure 4: *Prosodic parameter values in other- and self-directed* wh-*questions. ml: f0 level, rng: f0 range, c: polynomial coefficients, en: energy*

## 5. Detection

We used the tree ensemble classifier AdaBoost M1 [11] designed for two-class problems. By brute force optimisation on a small development set the ensemble learner parameters were set as follows: number of learners: 100, maximum number of decision splits: 5, minimum number of observations at a leaf: 5, minimum number of observations at a non-terminal node: 10. The preliminary results of a tenfold cross-validation are presented in table 2. At the current state accuracy amounts to 87% and is expected to rise with additional training data.

## 6. Discussion

We found clear evidence for prosodic differences in self- and other-directed questions. Overall prosody is more expressive in other-directed than in self-directed speech: f0 level and range as well as energy are higher, and local f0 movements are more pronounced. This is reflected in overall higher absolute values of all examined features in other-directed questions.

Table 2: *10-fold cross-validation. Mean accuracy, weighted recall, precision, and F1 score (and yes: the F1 score can indeed be below precision and recall).*

| | |
|---|---|
| Accuracy | 0.87 |
| Weighted Recall | 0.87 |
| Weighted Precision | 0.94 |
| Weighted F1 score | 0.86 |
| Kappa | 0.65 |

The weights in Table 1 show that energy (*en*) and the sharpness of the local f0 movement ($c_2$) deviating from the midline on the question word are most influential in marking self- and other-directedness.

All differences are gradual and quantitative, not qualitative. To give examples, for both conditions there is f0 declination (negative *ml_slope*), which is flatter in self-directed speech. In both conditions there are concave as well as convex local f0 shapes on the question word (negative and positive $c_2$), but again the shape is less pronounced in self-directed questions (much less variation around 0).

These differences in expressiveness are in line with the findings of [12] who compared linguistic and prosodic features in on- and offtalk. Since [12] examined human-machine communication, they partly attributed this difference in expressiveness to the artefact that humans tend to hyperarticulate when talking to machines which therefore enlarges the differences between other- (here: computer) and self-directed speech. However, our data suggests that these differences also hold for human–human communication. They might be actively used by the speaker to signal whether or not a question is information-seeking and requires a reaction by the interlocutor.

Finally, automatic question type prediction based on the extracted features yields high accuracies which will be beneficial for more general offtalk detection for dialog systems.

## 7. Acknowledgments

The work of the first author was funded by OTKA K 115922 and by an institute partnership programme of the Alexander von Humboldt Foundation. The second author is financed by a fellowship of the Alexander von Humboldt Foundation.

# 8. References

[1] J. Sadock and A. Zwicky, “Speech act distinctions in syntax,” in *Language Typology and Syntactic Description I: Clause Structure*. Cambridge: CUP, 1985, pp. 155–96.

[2] J. Searle, *Speech Acts*. CUP, 1969.

[3] D. Wilson and D. Sperber, “Mood and the analysis of non-declarative sentences,” in *Human Agency*, J. Dancy, J. Moravcsik, and T. C., Eds. Stanford, CA: Stanford University Press, 1988, pp. 77–101.

[4] H. Truckenbrodt, “Zur Strukturbedeutung von Interrogativsätzen,” in *Linguistische Berichte*. Hamburg: Helmut Buske Verlag, 2004, vol. 199, pp. 313–350.

[5] J. Ginzburg, R. Fernández, and D. Schlangen, “Self-addressed questions in disfluencies,” in *Proc. 6th Workshop on Disfluency in Spontaneous Speech*, Stockholm, 2013, pp. 33–36.

[6] J. Ohala, “The frequency code underlies the sound symbolic use of voice pitch,” in *Sound Symbolism*. Cambridge: Cambridge University Press, 1994.

[7] A. Gravano, v. Beňuš, H. Chávez, J. Hirschberg, and L. Wilcox, “On the role of context and prosody in the interpretation of ‘okay’,” in *Proc. 45th Annual Meeting of Association of Computacional Linguistics*, Prague, 2007, pp. 800–807.

[8] P. Boersma and D. Weenink, “PRAAT, a system for doing phonetics by computer,” Institute of Phonetic Sciences of the University of Amsterdam, Tech. Rep., 1999, 132–182.

[9] A. Savitzky and M. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[10] U. Reichel and K. Mády, “Comparing parameterizations of pitch register and its disconti nuities at prosodic boundaries for Hungarian,” in *Proc. Interspeech 2014*, Singapore, 2014, pp. 111–115.

[11] Y. Freund and R. Schapire, “A short introduction to boosting,” *J. Japanese Society for Artificial Intelligence*, no. 5, pp. 771–780, 1999.

[12] A. Batliner, C. Hacker, and E. Nöth, “To talk or not to talk with a computer – Taking into account the users focus of attention,” *J. Multimodal User Interfaces*, vol. 2, pp. 171–186, 2008.