ON THE OBJECTIVITY OF PROSODIC PHRASES

Jan Romportl

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, Pilsen, 306 14 Czech Republic
rompi@kky.zcu.cz, Tel: +420 377 632 533, Fax: +420 377 632 502

Abstract

Objective annotation of prosodic phrases in a corpus for a text-to-speech system is an important issue due to its influence on the naturalness of synthesised speech. The paper discusses drawbacks of common ways of prosodic phrase annotation and proposes a concept of prosodic phrases defined by a maximum likelihood estimation over results of many parallel subjective annotations. Validity of this method is analysed in terms of agreement among the subjects using Cohen's and Fleiss' kappa measures and heuristically modified relative agreement.

Keywords

Prosodic phrase, objective annotation, speech corpus

1.0 Introduction

A text-to-prosody (TTP) system as a subsystem of a text-to-speech (TTS) system can be conceived and developed in terms of a machine learning (ML) paradigm. Such a conception, however, requires the existence of suitable training and testing databases covering desired prosodic phenomena. In this case, what does "suitable" mean? How much of such data do we need? And, most importantly, *who* can prepare such data?

I will try to explain my view on these questions, which can be easily classified as rather pragmatic – that is to say, I will hold the view that those data can be successfully prepared with very modest a priori phonetic knowledge and I will demonstrate this on the case of prosodic phrase and semantic accent annotation in a corpus for the Czech unit-selection TTS system, ARTIC. It is, nevertheless, very important to note that I do not deny the importance of phonetic knowledge per se – I only assert that in this particular situation, such knowledge is not essential.

The concept of *prosodic phrase* basically corresponds to a traditional view or to what is meant by the term "phonemic clause" (or "discourse segment") in Czech literature (Palková, 1974), i.e. a phonetic unit which underlies the perception of a certain level of rhythmical qualities in language. A prosodic phrase is mainly delimited by the acoustical features of its boundaries and it can also contain an "intonation peak". However, as Palková discusses (ibid.), there is no empirical evidence supporting any stronger assumption about the presence/absence of an intonation peak or their number in a Czech utterance.

We further assume that a speaker may emphasise any number of words by acoustic means to express (perhaps even unintentionally) their prominence in comparison with other words. The acoustic prominence of a word can deliver various kinds of information: it can either help structure an utterance and delimit phrase boundaries, or it can modify the semantics and pragmatics of an utterance. The first type of acoustic prominence is automatically realised at the end of a phrase. The second type, called *semantic accent*, can be realised anywhere within an utterance and its usage is semantically functional – it often plays an important role in the articulation of topic-focus.

Prosodic phrase and semantic accent usage can be illustrated by several examples. Prosodic phrase boundaries are designated by "/", words in italics are with the semantic accent:

- "tito živočichové / jsou velice inteligentní / ale také pomalí" – "these animals / are very intelligent / but also slow"
- "i hráčům *jiných* sportů / jdou šipky dobře" – "players of *other* sports / are also good in darts"
- "podle *Iráku* / jich bylo *pouze* osm" – "according to *Iraq* / there were *only* eight of them"

2.0 Aspects of prosodic phrase annotation

The prosodic phrase annotation process I will discuss here has one specific goal: to allow the designed TTP system based on ML techniques to produce prosodically natural speech in terms of phrasing. Hence, this process is not primarily focused on investigating the nature of prosodic phrases but the results can help identify them. The aspects of semantic accent annotation are analogous to the following aspects of prosodic phrase annotation.

2.1 Speech synthesis requirements

The first aspect of the prosodic phrase annotation is determined by various demands posed by speech synthesis techniques – namely by the unit-selection concatenative synthesis algorithm.

This algorithm is based on selection of speech units according to their classification into various relevant, mostly structural categories. A synthesized utterance is represented as a target sequence of units and their categories, and the synthesis algorithm tries to find units from the speech segment database matching the target sequence as closely as possible.

This means that every unit in the database must be described (annotated) by values of all the categories used in the algorithm.

Since one of the categories is based on prosodic phrasing (e.g. position of a unit in a prosodic phrase, etc.), prosodic phrases must be determined for every sentence of the source speech corpus whose units appear in the speech segment database. Most current unit-selection TTS systems try to make use of as much data as possible (so as to achieve the most natural sounding speech), thus speech segment databases usually consist of thousands of sentences. In the case of our TTS system, the database consists of 10 000 Czech declarative sentences (taken from newspaper texts) recorded in a studio by a male professional speaker. All of the sentences were annotated with prosodic phrase boundaries.

It should be mentioned that the difficulty of prosodic phrase placement strongly depends on the type of the text in the corpus and on the speaker. If the text is uttered very rhythmically and affectively (such as heard in a good actor's performance), it is easier to find prosodic phrases than in a neutral and intonationally "flat" speech. Due to the constraints posed by the TTS techniques our speech corpus is of the latter kind. Moreover, prosodic phrase boundaries in Czech speech often tend to be more vague and ambiguous than in English.

2.2 Automatic annotation

Due to the large amount of data to annotate (for a single voice – and most TTS systems offer many voices) it seems to be inevitable for machines to replace human annotators. Another, and perhaps more important reason is consistency in transcription. It is extremely difficult (or even impossible) for a human annotator to maintain consistent perception and annotation of such phenomena throughout corpora including thousands or tens of thousands of sentences.

The idea is to manually (i.e., by humans) designate prosodic phrase boundaries and semantic accents in a reasonable sub-part of the whole corpus (250 sentences in our case) and then use ML pattern recognition techniques to automatically and consistently extend this annotation to the whole corpus (10 000 sentences in our case).

This method, however, imposes even stronger demands on consistency of the manual annotations. Should there be discrepancies between acoustic/textual cues and phrase boundary judgements in training and testing data, reliability of the ML classifier may seriously decrease.

2.3 Prosodic phrases as theoretical entities

Prosodic phrases are what we define them to be. Their ontological status is the same as that of other theory-based entities. Only through theories do we know what a prosodic phrase is.

“Theories are nets cast to catch what we call ‘the world’: to rationalize, to explain, and to master it. We endeavour to make the mesh ever finer and finer.” (Popper, 2002, p. 38)

Some theories define prosodic phrases by their boundaries, realised as particular f0 or duration movements. These definitions are undoubtedly useful for certain purposes, but due to their reductionist form they lack one important attribute – *function*. I believe that defining prosodic phrases (and not only them) through their function is much more epistemologically valuable, and that the goal of a TTP system is to generate prosody with a proper function (from the point of view of a listener), no matter its actual form.

We can say that the function of a prosodic phrase is: 1) to create one layer of the rhythmical structure of an utterance (this rhythmical layer is hierarchically and structurally higher than the level of prosodic words and lower than the level of utterances); 2) to help a listener reconstruct the underlying nonlinear structure of the utterance. This means that a prosodic phrase corresponds to a continuous segment of an utterance where a single instance of the function (1) and a disposition for the function (2) are prosodically realised (i.e. realised by means of a prosodic form). On one hand this definition brings “epistemological sense” into looking for prosodic phrases, on the other hand it introduces subjectivity and hence inconsistencies.

This brings us to the following problem: the TTP system should generate functionally proper prosody, thus the whole corpus must annotate prosodic phrases in the aforementioned functional-perceptual sense. The training/testing datasets also must be manually annotated in the same sense. Such manual annotation is likely to be inconsistent – and this is indeed unwanted.

What is and what is not a prosodic phrase (or a prosodic phrase boundary) from a scientific point of view can be formulated and decided by an empirical theory. However, it seems that this theory can only rely on either subjective judgements about perception, rhythm, syntactical disambiguation, etc., or reductively on non-functional cues, such as f0 movements, etc. The former case is scientifically problematic because empirical facts are substituted by subjective beliefs, the latter lacks functional relevancy.

If a trained phonetician manually annotates a part of a speech corpus, he might be consistent in his subjective judgements (because of his training) and might strictly obey the principles of a particular theory about prosodic phrases, but one can immediately find a group of people (non-specialists) who will disagree with a significant number of phrase boundaries placed by this expert in phonetics. This means that such a prosodic phrase annotation does not represent prosodic phrases where they really are, but where one person thinks they are. This situation actually looks like the particular phonetic theory “knew” where the prosodic phrases really are and the phonetician either “hits” or “misses” them. We could also consider three or five or more phoneticians doing this job as a team – they would try to make best of their experience and knowledge of the theory, they would discuss where the theory posits phrase boundaries and they would eventually settle on some mutually agreeable decision. Still, there would be no way to see how close to the “real” phrase placement their decision is. So, we must ask ourselves a question: what is the nature of the empirical statements (about prosodic phrases) of such a theory?

Of course if we want to discover the nature of prosodic phrases from within the language system, it is perfectly correct to posit theoretical features of prosodic phrases a priori and then test them on real speech data. The only problem in this particular case is that the testability of these units is rather questionable, as I have already argued.

However, if our primary goal is to know objectively and exactly where the prosodic phrases in a particular speech corpus are, we can describe the prosodic reality by a different, more pragmatic theory. After all, such objective knowledge can be of great usefulness for testing other prosodic theories.

The theory I am proposing here to define what we call a "prosodic phrase" (at least in the Czech language and at least for the sake of what I have discussed above) comprises following assumptions:

1. Every normal speaker/listener (native speaker/listener) has an intuitive sense of rhythm in speech. The purpose of this rhythm is to help perceive and structure utterances.
2. We can suppose that speech rhythm is constituted by specific units, which are, on a certain structural level, called prosodic phrases. This is an important piece of knowledge we are borrowing from other theories.
3. There is a probabilistic causal relationship between the presence of the boundary of a prosodic phrase and the intuitively (subjectively) conditioned conscious designation of this boundary by a listener.
4. Empirical facts are statements about behaviour of listeners – a listener either asserts that there is a prosodic phrase boundary at a certain place in speech or asserts that there is not.
5. If there is a statistically relevant number of empirical facts from independent listeners describing the same portion of speech, a model of an *objective annotator* can be created. The objective annotator is a maximum likelihood estimation over the empirical facts.
6. A prosodic phrase is what is designated by the objective annotator.

The nature of prosodic phrases based on these assumptions is entirely clear, testable and reproducible. It is quite likely that there would be differences between the "opinion" of the objective annotator and the opinion (perhaps collective) of the aforementioned phonetic experts. Although it may be interesting to analyse such differences, one must keep in mind that, metaphorically speaking, it is comparing two different (theoretical) worlds without clear bridging links or principles.

It might seem rather vague to use the term "statistically relevant number", but we can define this number more precisely as the number of listeners which satisfies the condition that an objective annotation created over this set of listeners equals to the objective annotation created over this set extended by one more arbitrary listener.

We can go even further in exploiting the aspects of our task described in the sections 2.1 and 2.2 and take into account the automatic annotation by these assumptions:

7. The *objective machine annotator* is such a classifier set up by ML techniques which achieves the highest possible classification performance on the testing data prepared by the objective annotator.
8. Prosodic phrase is what is designated by the objective machine annotator.

This allows us to acquire stable and consistent prosodic phrase annotation of speech corpora of an arbitrary size without doubts (towards or in the sense of the theory) about its objectiveness. There is one more question: can we do all these things also for semantic accents?

3.0 Experiments

The annotation process with the aspects described in Section 2.0 has been based on two large-scale listening tests – I will further denote them as Test 1 and Test 2.

3.1 Listening tests

The listening tests were organised on the client-server basis using a specially developed web application. We used the speech corpus which the text-to-speech system ARTIC (Matoušek – Romportl, 2007) is based on. The corpus was very carefully recorded in a studio by an experienced male speaker (the choice of the speaker was made in consultation with two experts from the Institute of Phonetics, Charles University in Prague) who had been instructed to read isolated sentences naturally, yet avoiding any expressiveness. The speaker did not know that the recorded sentences also would be used for the phrasing analysis. The way the corpus has been recorded (i.e., the type of recorded speech) obviously influences the scope of linguistically relevant findings of the research – therefore the relevance of the quantitative results presented here is limited to the aforementioned speech domain; however, the methods we have used are definitely not limited to this data.

3.1.1 Test 1

In Test 1, we randomly selected 100 sentences from the corpus and loaded them together with their orthographic transcriptions into the web application. Potential test participants were selected among university students from all faculties (with a special focus on students of linguistics). When they finished the listening tests, they were financially rewarded (so as to increase their motivation). The participants could do all of the work from their homes without any personal contact with the test organisers – we have thus undertaken various measures to detect possible cheating, carelessness or misunderstandings.

The participants were instructed to listen to the sentence recordings very carefully and subsequently designate words where they are sure there is a phrase boundary and words where they feel there might be a phrase boundary (i.e., these two cases were carefully distinguished). Prior to the test itself the participants had been briefly familiarised with the background of the problem and in this tutorial they listened to several training samples which showed possible phrasing demonstrations. It is, however, very important to note that we intentionally did not want to make almost any a priori assumptions about phrase boundary qualities or behaviour. We wanted to create a "notion of prosodic phrase" in the participants and let them designate whatever subjectively fulfilled this notion (cf. Section 2.3).

We eventually received correctly finished tests from 103 participants (the total number of students who took part in these tests was 174, some of the students had not finished their tests, some of them had not even started, and there were also several apparent cheating attempts), which provided a robust observation set for further evaluation. Several quantitative facts about the Test 1 are in Table 1.

Table 1. Quantitative facts about Test 1 and Test 2

	Test 1	Test 2 (Part 2)
Finished tests	103	99
Participants with phonetic education	25	19
Average time spent on one test	92 min	168 min
Average number of sentence replays	2.33	2.25
Average number of sessions per user	3.10	5.14
Total number of sentences	100	150
Total number of word tokens	1063	1531
Total length of speech	≈ 508 s	≈ 741 s

3.1.2 Test 2

Test 2 (which was carried out 3 months after Test 1) consisted of two parts (hereafter denoted as Part 1 and Part 2). Part 1 focused on finding the semantic accents in sentences where the prosodic phrase boundaries were already given. The same sentences from Test 1 were used again and the participants had been instructed to listen to these sentences very carefully and subsequently designate words where they perceived semantic accent. The textual form of the sentences was displayed together with the a priori prosodic phrases acquired from the objective annotation based on Test 1. The participants had to accept this phrasing and adapt their semantic accent assignment accordingly. Part 1 also served as a "tutorial" for Part 2 since the participants could infer how to annotate the prosodic phrases in Test 1.

Part 2 was actually a combination of Part 1 and Test 1: we selected another 150 sentences from our corpus and the participants were again instructed to listen to the sentence recordings and designate the semantic accents. However, in this part, the task was also to designate words with perceived prosodic phrase boundaries (cf. Section 3.1.1).

The quantitative facts about Test 2 can be compared with Test 1 again in Table 1.

3.2 Objective annotation

We can now describe the problem of modelling annotation based on many independent observations on a more abstract and formal level:

Let X be a random process defined as $X = \{X_t : t \in T\}$, where $T = \{1, 2, \dots n\}$ is a set of time points respective to the ordinal numbering of words in the test sentences (i.e., the first word in the first sentence has $t = 1$, the second word in the first sentence has $t = 2$, and so on), and X_t are random variables which hold $X_t = 1$ if and only if the t-th word finishes a prosodic phrase, and $X_t = 0$ otherwise. Exactly the same can be done for the semantic accents, such a random process is analogous to X and will be denoted as Y. We assume that the random processes X and Y are mutually independent.

Now let the test participants be numbered by the set $J = \{1, 2, \dots m\}$, i.e., the first participant has $j = 1$, the last one has $j = m$. We can define m random processes $O^{(1)}, \dots O^{(m)}$ representing the participants' responses (observations, empirical facts) such that $O^{(j)} = \{O^{(j)}{}_t : t \in T\}$, where t has the same meaning as for the process X, and $O^{(j)}{}_t$ are random variables which hold $O^{(j)}{}_t = 1$ if and only if the j-th participant asserts that the t-th word finishes a prosodic phrase, and $O^{(j)}{}_t = 0$ if and only if the j-th participant does not assert that the t-th word finishes a prosodic phrase.

Our goal can now be re-formulated as follows: knowing the observations $O^{(1)}, \dots O^{(m)}$, we want to estimate the hidden trajectory of the process X which best satisfies the given observations. This can be determined analogically for the process Y. For the sake of clarity, I will speak further in the text only about the process X, assuming that everything which holds for it, also holds for the process Y. It is supported by the fact that the two variants of the answers on the phrase boundary presence/absence (i.e., "boundary certain" and "boundary maybe") were treated equally – this was based on the assumption that if the "statistically relevant" number of participants think that there *might be* a phrase boundary at the given place, it *really is* there. The reason for allowing two levels of certainty from the participants' view was mainly due to the experience that if a listener is really not sure, he answers randomly – and this can be avoided by the "maybe" variant. The difference between these two variants is utilised in the participants' agreement calculation (see Section 3.3.2).

The aforementioned goal of the hidden trajectory estimation can be transformed into the problem of finding the most likely model parameters given the observed data – a *maximum likelihood* approach (cf. Section 2.3). I will not describe this method here because it involves some mathematics, and I have described it elsewhere (Romportl, 2008). In any case, the result of this method is the objective annotation of 250 sentences with both prosodic phrases and semantic accents

3.3 Validity of the objective annotation

The validity of the objective annotation can be interpreted as a quantitative measure of inter-participant agreement. This way we can test the validity of the assumptions listed in Section 2.3. If the measure of inter-participant agreement is too low, it will suggest that one or more assumptions should be modified.

3.3.1 Kappa measures

The agreement between two test participants can be measured by means of a statistical correlation. However, if one of the answers in the test is significantly more frequent that the other, two participants can often agree just by chance and the correlation is thus relatively high, therefore misleading. For example, if the first participant designates phrase boundaries where they really are and the second participant asserts there is no phrase boundary in the corpus, their correlation still will be relatively high because they will often "agree" on the non-boundary words which are more frequent than the boundary ones.

Such influence of the agreement by chance can be eliminated by using Cohen's and Fleiss' kappa measures (κ_C, κ_F). Cohen's kappa (Cohen, 1960) is a scalar value measuring agreement between two test participants; Fleiss' kappa (Fleiss, 1971) expresses agreement among more participants at once.

We calculated κ_F for Test 1 and Test 2 separately and then for the whole set of 250 sentences in two variants – including and excluding words followed by a pause (a phrase boundary with a pause is much easier to detect). This was calculated for semantic accents too. As κ_C measures only mutual agreement between two participants, we calculated it for every pair of the participants and then presented it as the average value. The results are displayed in Table 2. Moreover, κ_C can be also calculated for every participant paired with the objective annotator – this is summarised in Table 3.

Table 2. Values of Fleiss' and Cohen's kappa. E{κ_C} is the average value for all pairs of the participants, D{κ_C} is the variance.

	Prosodic phrases				Semantic accents		
	Whole set	Whole set excl. pauses	Test 1	Test 2	Whole set	Test 1	Test 2
κ_F	0.5790	0.4171	0.4542	0.6636	0.1283	0.1325	0.1201
$E\{\kappa_C\}$	0.5837	0.4293	0.4632	0.6710	0.1271	0.1417	0.1259
$D\{\kappa_C\}$	0.0068	0.0154	0.0180	0.0083	0.0052	0.0048	0.0069
max κ_C	0.7669	0.8179	0.8538	0.8929	0.7690	0.7910	0.7875
min κ_C	0.1718	0.0927	0.0801	0.1978	0.0081	0.0072	0.0010

Table 3. Values of Cohen's kappa for the participants paired with the objective annotator. E{κ_C} is the average value, D{κ_C} is the variance.

	Prosodic phrases				Semantic accents		
	Whole set	Whole set excl. pauses	Test 1	Test 2	Whole set	Test 1	Test 2
$E\{\kappa_C\}$	0.7100	0.5637	0.6182	0.7729	0.2596	0.2977	0.2574
$D\{\kappa_C\}$	0.0061	0.0086	0.0192	0.0058	0.0069	0.0061	0.0078
max κ_C	0.8242	0.8419	0.8173	0.9013	0.4572	0.4610	0.4638
min κ_C	0.3854	0.1301	0.0000	0.3780	0.0894	0.0613	0.0000

3.3.2 Heuristically modified relative agreement

In spite of the kappa measures (both Cohen's and Fleiss') being the chance-corrected measure of agreement, their usage and more importantly their interpretation is often rather problematic (Maclure – Willett, 1987). It is thus advisable to supplement them with another, more informed quantitative criteria.

Although it might seem that mere relative agreement between two participants is not a very good choice, I am convinced that if this simple measure is slightly heuristically modified, it can provide a statistical tool which takes into account information about behaviour of the participants and relevancy of various types of answers.

I therefore propose two types of relative agreements for prosodic phrase boundaries:

- The agreement of a pair of participants is calculated as the number of cases in which both participants chose the same answer, divided by the total number of answered cases. More formally: the agreement $A_1(i, j)$ between the participants i and j is defined as

  $$A_1(i,j) = \frac{\sum_{t\in T} f_{ij}(t)}{n},$$

  and

  $$f_{ij}(t) = \begin{cases} 1 \leftrightarrow \left(\rho\left(o_t^{(i)}\right) = \rho\left(o_t^{(j)}\right)\right) \\ 0 \leftrightarrow \left(\rho\left(o_t^{(i)}\right) \neq \rho\left(o_t^{(j)}\right)\right) \end{cases},$$

  where ρ(x) is integer rounding of x (will be explained later). The overall average agreement of this type is then given as

  $$A_1 = \frac{\sum_{i,j\in J, j>i} A_1(i,j)}{\frac{1}{2}m^2 - m}.$$

- The agreement of a pair of participants is calculated as the number of all cases in which *both* participants chose the *positive* answer, divided by the number of cases in which *at least one* of these two participants chose the *positive* answer. In this way, the agreement calculation is motivated by the heuristic knowledge that a vast majority of cases agreeing by chance involve negative answers; moreover, the agreement on *absence* of a prosodic phrase has epistemologically “lower” modality than the agreement on its *presence*. Again, formally we can write

$$A_2(i,j) = \frac{\sum_{t \in T} \left(f_{ij}(t) \cdot c_{ij}(t) \right)}{\sum_{t \in T} c_{ij}(t)},$$

where

$$c_{ij}(t) = \begin{cases} 1 \Leftrightarrow \left(O_t^{(i)} = 1 \vee O_t^{(j)} = 1 \right) \\ 0 \Leftrightarrow \left(O_t^{(i)} = 0 \wedge O_t^{(j)} = 0 \right) \end{cases}.$$

I have mentioned in Section 3.2 that the participants actually had three choices when answering whether a particular word bears a phrase boundary: “yes”, “no” and “maybe”. Although the “maybe” variant is treated as “yes” in the process of the objective annotation estimation, the way how it is interpreted in the process of evaluation can influence the result of the evaluation. We can introduce three methods of interpretation of the “maybe” variant:

- ***M1***: There is no difference between the “maybe” and the “yes” variant. This means that $O^{(j)}_t = 1$ in both cases: the *j*-th participant designates the *t*-th word as “certainly with phrase boundary” or he designates it as “maybe with phrase boundary”.
- ***M2***: The “maybe” variant is ignored, i.e. $O^{(j)}_t = 0$ anytime the *j*-th participant designates the *t*-th word as “maybe with phrase boundary”.
- ***M3***: The “maybe” variant has a different value than the “yes” variant: $O^{(j)}_t = 1$ in case the *j*-th participant designates the *t*-th word as “certainly with phrase boundary” and $O^{(j)}_t = 0.6$ in case the *j*-th participant designates the *t*-th word as “maybe with phrase boundary”. It is the “maybe” variant here for which I have defined the operator ρ(*x*) for integer rounding. For this method it is also necessary to slightly modify the following equation:

$$c_{ij}(t) = \begin{cases} 1 \Leftrightarrow \left(O_t^{(i)} = 1 \vee O_t^{(j)} = 1 \right) \\ 1 \Leftrightarrow \left(O_t^{(i)} = 0.6 \wedge O_t^{(j)} = 0.6 \right) \\ 0 \Leftrightarrow \text{otherwise} \end{cases}.$$

The combination of these three methods with the functions A_1 and A_2 gives us six ways how to compare a pair of participants. Table 4 summarises values of these heuristically modified relative agreements.

Table 4. Heuristically modified relative agreement between the test participants. The most relevant and informative values are bold.

	M1		***M2***		***M3***	
	A_1	A_2	A_1	A_2	A_1	A_2
Test 1	0.81	0.41	0.86	0.41	0.81	**0.56**
Test 2	0.93	0.43	0.95	0.43	0.93	**0.74**
Whole set excl. pauses	0.73	0.37	0.76	0.37	0.73	**0.53**

4.0 Discussion and conclusion

On the basis of the values presented in Table 2 and 3 it is clear that the validity of prosodic phrases based on the assumptions from Section 2.3 is well supported because the inter-participant agreement is relatively high. The Fleiss' kappa value is very similar to the values for English presented in recent studies (Mo et al., 2008).

The average agreement between the participants and the objective annotator is also high and it suggests that prosodic phrases defined via the objective annotator are not a mere formal construct, but they maintain a very strong link with human perception. This is a very important conclusion.

Semantic accents, on the other hand, are significantly more difficult to test. Thus, their existence in terms of their definition from Section 1.0 is very questionable. Therefore, the estimation of the process *Y* should probably not be called the objective annotation (of semantic accents). It is the most stable and objective annotation on which we can base the responses of the test participants (i.e., on the empirical facts about their assessment of semantic accents), but the actual responses acquired in Test 2 are apparently rather chaotic and too inconsistent. Although we can subjectively agree that there really is "something" in many utterances that "sounds emphasised", we will need a different theory to be able to capture such phenomena objectively.

Although the kappa measures are a good quantitative indicator, the heuristically modified relative agreement is easier to interpret when creating overall judgement about the acquired data: high agreement calculated by the function A_1 with the method *M1* significantly decreases when using A_2, which implicates that the participants evidence strong agreement on the absence of phrase boundaries. The most informative value about the agreement on the presence of boundaries, taking the "maybe" variant in consideration is given by the combination of A_2 and *M3*.

I would also point out the differences between Test 1 and Test 2. The different values of the agreement measures are most likely caused by two reasons: 1) in Part 2 of Test 2, the participants had already passed the annotation of semantic accents from Part 1, so they had already been familiarised with the phrase objective annotation from Test 1. This means they could acquire better implicit understanding of the phenomenon of prosodic phrases; and 2) it probably makes a difference whether the participants designate phrases and semantic accents separately or at the same time. This will be the focus of future investigations designed to evaluate the extent to which these two reasons could have influenced the results.

The current state of development and performance of the objective machine annotator is presented elsewhere (Romportl, 2010). The classifier is based on artificial neural networks and is able to designate prosodic phrases in the whole speech corpus significantly better than an average human annotator.

Considering these results, I think the proposed pragmatic theory of prosodic phrases can be characterised as acceptable. There are still many questions to be answered, such as the appropriateness of the criterion for the "statistical relevance" of the number of empirical facts, or possible differences in the objective annotation given different sets of test participants. Answering these questions, however, will demand data based on a new, specially and carefully designed listening test.

Acknowledgements

Support for this work was provided by the Ministry of Education of the Czech Republic, project LC536.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76.

Maclure, M. & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161–169.

Matoušek, J. & Romportl, J. (2007). Recording and annotation of speech corpus for Czech unit selection speech synthesis. In *Proceedings of TSD 2007*, Lecture Notes in Artificial Intelligence, vol. 4629 (pp. 326–333). Berlin–Heidelberg: Springer.

Mo, Y. & Cole, J. & Lee, E.-K. (2008). Naive listeners' prominence and boundary perception. In *Proceedings of Speech Prosody 2008* (pp. 735–738). Campinas, Brazil.

Palková, Z. (1974). *Rytmická výstavba prozaického textu* (with English resume: The rhythmical potential of prose). Prague: Academia.

Popper, K. R. (2002). *The Logic of Scientific Discovery*. London: Routledge.

Romportl, J. (2008). *Zvyšování přirozenosti strojově vytvářené řeči v oblasti supra-segmentálních zvukových jevů* (Improving Naturalness of Machine-Generated Speech on the Suprasegmental Level). Ph.D. dissertation, Department of Cybernetics, University of West Bohemia, Pilsen.

Romportl, J. (2010). Automatic prosodic phrase annotation in corpus for speech synthesis. In *Proceedings of Speech Prosody 2010* (to be published in May 2010). Chicago IL, USA.