On Non-Native Listeners’ Ability to Identify Prominence and Pitch Accents in English Monologic Speech

ABSTRACT

A comparative perceptual study involving two experimental groups with different native languages (Russian and Czech) shows that phonologically trained non-native speakers of English are good at identifying basic suprasegmental features of the English language, namely prominence (sentence stress) and accent types, which potentially makes it possible to use their prosodic annotations when validating cross-language intonation research. The occasional failure of both experimental groups to identify certain accent types is explained in the study by the annotators’ mother tongue’s prosodic interference: Czech and Russian speakers rely on different acoustic cues when identifying prosodic features in their native languages and transfer this habit to the discrimination of English prosodic characteristics. The study demonstrates that when a prosodic cue is not marked in the speaker’s mother tongue, it will likely be ignored in the foreign language.

Keywords: prominence, pitch accents, L1 prosodic interference, perceptual analysis

Sposobnost nerojenih poslušalcev angleščine prepoznavati naglase v angleškem monološkem govoru

IZVLEČEK

Primerjalna študija percepcije, ki je vključevala dve eksperimentalni skupini z različnima maternima jezikoma (ruščino in češčino), je pokazala, da so fonološko usposobljeni nerojeni govorci angleškega jezika dobro prepoznali osnovne suprasegmentalne značilnosti angleškega jezika, in sicer prominenco (stavčni poudarek) in vrste naglasa, kar potencialno omogoča uporabo prozodičnih zapisov pri potrjevanju medjezikovnih raziskav intonacije. Občasna neuspešnost obeh eksperimentalnih skupin pri prepoznavanju določenih naglasnih tipov je v študiji pojasnjena s prozodično interferenco maternega jezika testirancev: češki in ruski govorci se pri prepoznavanju prozodičnih značilnosti v svojih maternih jezikih zanašajo na različne akustične signale in to navado prenašajo na razločevanje angleških prozodičnih značilnosti. Študija dokazuje, da če prozodični namig ni označen v maternem jeziku govorca, bo verjetno prezrt tudi v tujem jeziku.

Ključne besede: prominencia, višinski poudarki, prozodična interferenca z L1, percepcijska analiza
1 Introduction

According to Rietveld and Chen (2006), the established procedure of conducting intonation research requires a mandatory validation of the collected prosodic data by native speakers’ perception. However, finding a suitable annotator for conducting perceptual analysis may be complicated, as it means considering a whole range of factors. For instance, the ability of the assessor to understand and annotate a recording has been proven to be significantly influenced by the depth of their previous exposure to a specific language variety or dialect (Mat Daud, Abu Kassim and Mat Daud 2011). Cole, Mahrt and Roy (2017, 322) prove that the most reliable prosodic annotations are made by people “who have strong familiarity with the language variety represented in the materials to be annotated”, which means in a time of “global Englishes” it makes sense to look for annotators “locally”.

Recruiting auditors “locally” may also be a more rational option as it puts many study parameters under the researcher’s immediate control, and allows for more flexibility in experimental designs. However, in non-English speaking countries researchers typically have difficulty recruiting enough phonetically trained native speakers of English to perform the perceptual verification of their results. To solve this problem, Snow et al. (2008) suggest looking into the potential of non-native-speaking annotators.

Although considering this suggestion a reasonable idea, Barry and Andreeva (2001) warn that non-native auditors inevitably attend to the prosodic cues dominant in their native languages, and listeners with native languages that are very different from English may come to unreliable conclusions. In contrast, studies of EFL speakers’ suprasegmental features (Riazantseva 2001; Trofimovich and Baker 2006) show that the amount of language contact positively correlates with the learner’s ability to recognize English prosodic patterns. McAndrews’ (2019) analysis of non-native listeners’ identification and discrimination abilities additionally shows that even brief periods of targeted suprasegmental instruction can help minimize foreign language comprehension difficulties and get near-native prosodic annotations.

Snow et al. (2008) believe that if the number of non-native annotators in a research group is big enough, it excludes many perceptual biases and makes the results compatible with the data acquired from native speakers. For instance, Cole, Mahrt and Roy (2017, 306) claim that 10–12 annotators are sufficient in crowd-sourcing tasks to “guarantee the replicability of the annotations at the group level, at least for American English”.

Taking the suggestions mentioned above into account, we decided to put non-native listeners’ ability to identify certain English suprasegmental features to the test. The present research aims to analyze the degree of correct perception of English prominence and pitch accents by two middle-sized groups of non-native auditors – Czech and Russian (L1) proficient speakers of English – before and after additional phonological training, with the following research questions:

(1) Do the results of the perceptual analysis of English intonation, namely prominence and accent types, performed by phonologically trained non-native speakers differ significantly from those performed by native speakers of English?
(2) Do the systemic prosodic differences between the Czech/Russian and English languages interfere with perceiving prominent words and accent types in English monologic speech?

(3) Is there a meaningful difference in the results of the perceptual analysis of a recorded English text performed by Czech/Russian speakers using different annotation techniques (RPT, ToBI, traditional (British))?

(4) Can the results of the perceptual analysis performed by non-native speakers be potentially used to validate the intonation of EFL (English as a foreign language) speakers?

2 English Prominence Perception and Marking

English prominence perception was chosen as the object of this study following Levis and Silpachai's (2017) research, which considers adequate prominence recognition a key parameter of oral speech perception. Research on prominence (Halliday 1967; Herman 2000; Trench 2011) has proved that phrasal accents in English convey a great deal of structural and discourse information indispensable for effective communication. For instance, the absence of clearly observable prominent words in a dialogue increases the listener's reaction time and deprives the utterances of certain pragmatic functions (Sanderman and Collier 1997).

Another factor influencing our choice was the multifaceted nature of prominence, which, according to Fougeron and Keating (1997), is believed to be a complex phenomenon influenced by both the word's phonetic realization and linguistic contrasts at the same time. Indeed, Cangemi and Grice (2016) show that although the acoustic marking of prominence in English may be rich and multilevel, the same acoustic parameters should be treated differently depending on the word's position in the utterance and its meaningful alignment with the whole text.

Cole, Mo and Hasegawa-Johnson (2010) posit that in most cases the underlying mechanisms of prominence perception are strongly tied to the listeners' expectations, which is illustrated by the fact that in most perceptual tasks of prominence identification the annotators do not rely solely on the acoustic information, but also their expectations of how the utterance will develop. In this way, prominence presents a perfect basis for studying the cumulative effects of various prosodic cues on its perception by groups with different native languages.

Although realizing the complexity of accent structures in English, we suppose that the variety of prominence types in English can be reduced to several intuitively understandable yet valid categories to fit in with our research findings' potential classroom application. Therefore, it is suggested that an intentionally simplified algorithm for annotators with different levels of annotation experience be established to guide them through the basic steps of prominence recognition. This algorithm should include Rapid Prosody Transcription (RPT), traditional (British) prominence marking, and ToBI annotation convention.

*The Rapid Prosody Transcription (RPT)* technique identifies prominent words as the ones “that the speaker has highlighted for the listener to make them stand out” (Cole and Shattuck-
Hufnagel 2016, 8). The annotator has several listenings (passes) of the recording to mark the prosodic boundaries of tone units (the “b-score”) and prominences (the “p-score”) of each word in the recording.¹

The RPT approach is easy to use, even with annotation novices. It allows sorting focus words irrespective of their actual accent type and other cue variations that make them appear prominent (Example 1).

Example 1. RPT annotation of a research sample (abstract) with calculated p-scores.

|in David Cameron (1) I follow in the footsteps (.75) of a great modern Prime Minister (1)| under David’s leadership (1)| the government stabilized the economy (1)| reduced the budget deficit (1)| and helped more (.75) people to work than ever before (1)| but David’s true (1) legacy| is not about the economy (1)| but about social justice (1)| from the introduction of same-sex marriage (1)| to taking people on low wages (1) out of income (1) tax| – The annotators marked perceived boundaries (|) and prominences (in bold) in a specially prepared transcript (all punctuation marks and capital letters at the start of the sentences were removed to avoid leading the perception) while listening. The calculated p-scores (in brackets) reflect the perception of the word as prominent in the research group.

The marked boundaries and prominences allow the researcher to divide the identified prominences into “nuclear accented” (preceding the boundary) and “prenuclear accented” (preceding another prominent word) (Example 2).

Example 2. Nuclear and prenuclear prominence in RPT annotation.

|a language | becomes a global language | because of the power of the people who speak it| – The utterance contains three tone units separated by perceived boundaries. Prominent words “language” and “global” are the only prominences in the respective units. Each of these words is followed by a boundary (although not immediately in the second case) and classified as a nuclear accent – an indispensable element of a tone unit carrying a contrastive focus. In the third tone group, “speak” is a nuclear accent, and “power” and “people” are prenuclear accents. Prenuclear accents, as a rule, mark new information in the text and play a rhythmical role, balancing the tone group.

Bishop, Kuo and Kim (2020) report that a combination of a distinct pitch movement on a nuclear-accented word with a following boundary tone typically creates a notable pitch change, reportedly accounting for nuclear accents’ easy identification in speech even by non-expert annotators. In contrast, prenuclear tones are more difficult to identify, especially in languages where acoustic marking of sentence prominence, like in Czech, differs from English. Still, Cole and Shattuck-Hufnagel (2016) report that the RPT technique can be a

¹ The b-score and the p-score are calculated as a ratio of annotators who marked the boundary/prominence in the transcript to the total number of auditors in the study. A figure close to 1 indicates the presence of the respective suprasegmental feature, and a figure close to 0 indicates the absence of the boundary/prominence (Cole and Shattuck-Hufnagel 2016, 9).
reliable means for crowdsourcing large amounts of prosodic information on prominence and boundaries due to its universality.

*Traditional (British) annotation:* Some practical teaching courses still follow O’Connor and Arnold’s (1973) traditional (British) annotation system, which presents melody as a directed continuum typically consisting of several stressed syllables in the “head” (at the beginning) of a prosodic phrase followed by a mandatory terminal tone at its end (Example 3).

**Example 3.** A prosodic phrase marked using traditional (British) terms.

<table>
<thead>
<tr>
<th>because of the</th>
<th>POwer of the PEople who</th>
<th>SPEAK</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-head</td>
<td>+ falling head</td>
<td>+ low fall as a terminal tone</td>
<td>+ tail</td>
</tr>
</tbody>
</table>

Brazil, Coulthard and Johns (1980) stress the importance of the discourse function of the tonic syllable (the terminal tone) in the intonation phrase, which establishes a basic opposition of finality (falling tones) and continuity (rising tones) of the interaction. Similarly, Wells (2006, 15) links the main pitch accent of each intonation phrase with the function of different nuclear tones from the perspective of grammatical or semantic-pragmatic meaning (e.g., “a definitive fall” or an “implicational fall-rise”), thus highlighting the opposition of “falling and non-falling tones” in the nucleus of an intonation phrase.

Concerning the annotation choices, Grabe, Nolan and Farrar (2000) note that at least in four varieties of English, including Standard Southern British English, terminal “falls” and “rises” make up the majority of nuclear tones, which, in our view, makes it possible to limit the inventory of nuclear markings to the fall vs. rise binary opposition. This choice of nuclear tones also seems most relevant for teaching as it reflects the fundamental differences in the syntactic meaning conveyed by nuclear prominence. However, in his recommendations, Brown (2014) emphasizes that nuclear tones do not have a simple one-to-one relationship with meaning, although some general conclusions (e.g., a fall equals finality, whereas a rise means continuity) can still be made.

Summing it up, a tone unit (prosodic phrase) in the traditional (British) terminology always has a terminal tone, viewed as the core of its structure. It may or may not have a preceding head, where the alteration of stressed and unstressed syllables can form different patterns (a “high head”, a “falling head”, a “stepping head,” and others). Traditional (British) annotation identifies prominences as stressed syllables in the “head” if they appear at the beginning of the tone unit and, very generally, subdivides them into falls vs. rises when they appear at its end (as a terminal/nuclear tone).

*ToBI annotation:* The ToBI annotation convention (Pierrehumbert 1980) classifies at least five different pitch accents varying in the placement of the high (H*) or low (L*) pitched stressed syllable. Although recognizing the advantages of the ToBI approach for accent identification using visual pitch contours, González-Ferreras et al. (2010) state that the annotators’ confusion about the ToBI typology is typically very high in the manual labelling of prominence.
Researchers believe that the ToBI typology works much better – with the agreement rate exceeding 87.7% – only when the annotators are given a binary choice. When presented with a multiclass prominence identification task, even annotation experts report doubts about discriminating different ToBI clusters (González-Ferreras et al. 2010). Therefore, picking the most typical and easily identifiable prenuclear and nuclear accents seems logical to help the annotators limit their confusion.

**Example 4.** A prosodic phrase marked in ToBI convention.

<table>
<thead>
<tr>
<th>4 because of the power of the people who speak it 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*</td>
</tr>
</tbody>
</table>

A series of high-pitch accents in a natural downstep followed by a boundary tone.

As for prenuclear accents in ToBI terms, studies prove that in most standard varieties of English modifications of the H* tone (mostly L+H*) come up most frequently (Cho and Keating 2009). González-Ferreras et al. (2010) also state that words with an H* accent are more strongly associated with perceived prominence among various annotator groups (Example 4), which allows us, following Bishop, Kuo and Kim (2020), to contrast the H*(L+H*) prenuclear accent to all other prenuclear types as a binary opposition.

Concerning the nuclear accents, there are again two distinct phrasal tones (L– and H–) and two final boundary tones (L% and H%) at the end of each tone unit, resulting in a range of potential combinations, with H–L% most frequently contrasted with L–L% in four varieties of British English (Grabe, Nolan, and Farrar 2000).

The complete prominence identification algorithm, considering RPT, traditional (British), and ToBI marking, is presented in Figure 1. It reflects the three main approaches to prominence identification presented as binary choices.

The apparent similarity in the approaches consists of dividing the stressed words in the tone unit into two broad categories depending on their closeness to the phrasal boundary. The stressed syllables in the “head” (traditional British annotation terminology) thus correspond to “prenuclear accents” (ToBI terms), and “terminal tones” (traditional (British) annotation terminology) correspond to “nuclear accents” (ToBI terms).

The main differences lie in the types of accents the two systems discriminate. For instance, the British annotation does not assess separate stressed syllables in the head, just marking the resulting contour. In contrast, the ToBI system views each accent separately and even separates a pitch accent and a phrasal/boundary tone, which are viewed as a unit in traditional (British) terms.

In conclusion, we believe that exposing learners to various annotation systems can turn classroom intonation practice (typically aided by traditional (British) terms), crowdsourcing for prosodic data (conducted using the RTP means), and computer-assisted acoustic analysis (recorded via the ToBI system) into a comprehensive FL intonation experience promoting better identification and comprehension of target language prosody.
3 Cross-Language Differences in Prominence Perception

The perception of many phonetic phenomena in a foreign language relies significantly on their comparative realization in the annotator’s/speaker’s native language (Vaissière 2005). According to the Markedness Differential Hypothesis (Eckman 1977), it is predicted that all segmental and suprasegmental features of the L2 (foreign language) phonological system can be roughly subdivided into several categories depending on the degree of their markedness in the speaker’s L1 (native language) phonological system.

If the marked features of the foreign language phonology differ from the corresponding patterns in the native language, they are also predicted to be more challenging to perceive (Eckman 1977). For instance, Czech learners of English typically have difficulties with English prominence identification, since stress is not a marked prosodic feature in their L1 (Weingartová, Poesová, and Volín 2014).

At the same time, those areas of the target language which are different from the native language but do not form phonological contrasts in L2 will not pose a severe problem to the students, primarily because their impact on the intelligibility of discourse is insignificant. Thus, a different fall/rise ratio in English and Czech does not affect a Czech speaker’s speech intelligibility, although it creates an impression of a strong foreign accent (Chamonikolasová 2007).

Since we deal with two groups of non-native speakers, we have to consider the differences in the phonetic realization of prominence in the Russian and Czech languages to point out the L1’s potential interference with English prominence perception.
The characteristics and functions of word stress. Generally, the nature of word stress in Russian is closer to English stress than Czech stress due to its contrastive nature and the absence of a fixed position in a word. With word stress being a marked feature in their native language, Russian L1 speakers are expected to be more sensitive to English stress placement and its contrastive role in a word or in a sentence than Czech speakers, who primarily rely on the delimitating function of the fixed on the first syllable word stress in their L1 (Skarnitzl and Eriksson 2017).

The acoustic marking of a stressed syllable: There is considerable experimental evidence that the prominent constituents of a Russian utterance (i.e., tonal units) are marked acoustically by a perceptible increase in vowel intensity, vowel length change, and higher F0 (Luchkina et al. 2015; Potapova and Potapov 2006). This combination makes Russian word prominence close to word prominence in English, marked by similar acoustic cues (Cruttenden 2014).

In contrast, the fixed on the first syllable of the Czech word stress is not acoustically marked (Skarnizl and Eriksson 2017). It is usually independent of vowel quality and length, and carries a lower F0 than the following syllables. Skarnitzl and Eriksson thus hypothesize that, unlike in Russian or English, word prominence in Czech is rarely achieved by the stressed syllable prominence, but rather by the acoustic configuration of the whole prosodic word (2017, 3221), with pitch going up and vowels being longer on the post-stressed syllables.

The acoustic marking of unstressed syllables (vowel reduction): Another critical characteristic that, in our view, may complicate both the perception and imitation of English prominence by Czech students is the absence of vowel quantitative and qualitative reduction (positional loss of certain characteristics) on unstressed syllables in the Czech language. In contrast to English and Russian, unstressed vowels in the Czech language do not become shorter and do not lose any of their characteristic features in unstressed positions, which, in Weingartová, Poesová and Volín’s opinion (2014), explains why Czech listeners demonstrate a weakened perceptual sensitivity to English vowel reductions and experience difficulties identifying distinctions between stressed and unstressed syllables.

The characteristics of information foci: Luchkina and Cole (2021, 3) argue that Russian prominence is mostly signal-driven, with positional changes being just an additional means of prominence expression: the acoustic-prosodic expression and the linear order of a sentence, in their opinion, only “reinforce” each other as cues of perceived prominence. In contrast, Skarnitzl and Hledíková (2022) claim that Czech prominence is mainly expectation-driven due to the above-mentioned communicative dynamism (Firbas 1996), which significantly limits the chances of an acoustically marked word being perceived by Czech native speakers as prominent if it comes outside of the expected context.

2 Although we realize that stress at a word level is different from prominence at a sentence level, Cho and Keating (2009, 468) insist that “stress and accent [i.e., sentence stress] represent different degrees along a single scale of prominence and are thus manifested in the same set of physical properties”. There is a definite overlap in the use of acoustic dimensions for lexical stress and phrasal accent; therefore, considering the established stress patterns in learners’ L1 may contribute to a better understanding of their perception of English prominence.
The range of terminal tones in L1: Despite certain apparent similarities in the pitch variation among all three languages, comparative studies of Russian/Czech and English intonation demonstrate a hugely different repertoire of tones employed by the speakers in their native languages, which, due to the L1 intonation transfer, can potentially influence English prominence perception. Leed’s (1965) account of Russian intonation states that many tones differ in distribution and meaning from the English tones. In particular, many pretonic rises, non-finite falls, and considerably higher pitch accents of Russian speakers are mentioned as impediments to mutual English-Russian intelligibility. As for the Czech tones, Chamonikolasová (2007) reports smaller melodic variability and a narrower pitch range of Czech speakers, which makes the Czech language sound rather “monotonous” to a foreign ear.

The boundary markers: Intonation studies of the Russian language demonstrate that the positional lengthening of the final stressed syllable, also present in English, significantly facilitates the listeners’ placement of boundary tones (Leed 1965). In contrast, although Volín and Skarnitzl (2007, 445) prove that in the Czech language “the temporal downtrend in the phrases is, to a large extent, driven by the final deceleration”, there is no reliable evidence that Czech speakers regard vowel lengthening as the principal boundary marker as “in languages with fixed stress where every word has stress on the first syllable, it is impossible to separate the correlates of stress from the word boundary effects” (Van Heuven 2018, 25).

To conclude, the comparative analysis of prominence expression means in Russian and Czech shows that the difference in the realization of various stress categories between Russian/Czech and English is significant. The acoustic nature of prominence in Russian seems closer to English prominence, with Czech prominence, on the other hand, being very different from that in English. The absence of salient acoustic markers and unstressed vowel reduction might seriously obstruct English prominence perception by Czech L1 speakers in cases when prominent words do not directly precede a phrasal boundary.

We believe the positive L1 intonation transfer will significantly facilitate the Russian speakers’ perception of English prominence. In contrast, Czech speakers of English might be confronted with specific difficulties due to the discrepancy between the stressed syllable’s acoustic properties in their L1 and English and the absence of unstressed vowel reduction in the Czech language. On the other hand, meaningful phonological instruction and extensive language exposure are expected to level the differences between the two groups of listeners and make them more sensitive to English phonological contrasts, neutralizing the adverse effects of L1 transfer.

4 Method

The annotation experiment was designed as a comparative study involving two groups of participants with different native languages (12 Czech speakers and 14 Russian speakers) but relatively similar in other ways (age, level of English, the study programme). Six British native speakers made up a control group that provided referential data (the so-called “ground truth”) for the similarity analysis. All participants signed an informed consent form allowing unrestricted use of their readings in scientific research and publications.
The study was designed as longitudinal research (Figure 2) that involved looking into the two groups’ prominence identification over an extended time: at the beginning of the English phonology course (an “amateur ear”, RPT annotation) and 12 weeks later, at the end of it (a “trained ear”, ToBI and traditional (British) annotation).

![Figure 2. Overview of the experiment’s structure and timing.](image)

4.1 Research Procedure

Stage 1 (an “amateur ear”): The first stage of the experiment was meant as a Rapid Prosody Transcription (RPT) annotation of four audio recordings (video footage excluded). The annotators were asked to circle/underline the words that, in their opinion, “stood out” in the speech. Each marked word in the transcript was then attributed p-scores identifying how many annotators marked it as prominent (e.g., [...] with a complete intolerance of failure [...] ) and b-scores marking the presence of a perceptual boundary after it. The results of the two annotator groups (Russian and Czech) were compared with the annotations marked by the native-speaker control group (Sub-group 1).

The RPT annotation technique was chosen for the first experimental stage as the most suitable for inexperienced assessors because it requires no special preparation. Annotations were performed after a short introduction of the task. The transcripts for this stage were edited: all punctuation marks and capital letters had been removed from the text to avoid punctuation-guided segmentation bias.

According to Bishop, Kuo and Kim (2020), RPT annotators make more correct prominence identifications when given more passes (listenings); however, the number of false identifications also increases proportionately. Since our annotator groups did not have any annotation experience at the beginning of the experiment, they were given three passes of the text: to familiarize themselves with the recording (first pass), to mark the prominences (second pass), and to mark the boundaries (third pass).³

Stage 2 (a “trained ear”): In the second stage of the experiment, the same groups of annotators had to identify the exact types of nuclear accents (Figure 1). All the participants had the transcripts with the prominent words pre-marked in bold (210 words) to limit the research corpus. The pre-marking was based on the groups’ Stage 1 collective performance and included only the words with the highest p-scores (.75–1). The annotators were given two passes of the recordings and were asked to make annotations employing two different annotation

³ Cole, Mahrt and Roy (2017, 305) report that prominence and boundary marking of the same recording can be performed in any order without any effect on the results.
conventions: ToBI annotations and traditional (British) marks. The texts from Stage 1 were intentionally divided into halves to avoid the annotators’ listening to the same recordings four times. While listening, additional short pauses (3–5 sec.) were made by a research assistant to allow some extra time for the annotators to make notes.

The experiment included annotation in ToBI and traditional (British) terms to check whether the groups were still following different annotation traditions after the training course. Following Cole, Mahrt and Roy (2017, 303), in ToBI annotation listeners are often guided by “down-up” pitch evidence from the accompanying pitch/soundwave contour, and a “consensus” annotation using an established inventory provides the best results. In traditional (British) terms, according to Wells (2006, 13), marking tones is more a question of “context and exercise”, with top-down factors such as the utterance’s discourse function leading the annotator. The aim of contrasting the two systems was to compare the reliability of the readings and to analyse the prevalence of acoustic vs. contextual factors.

**Statistical analysis:** Both groups’ traditional (British) annotation symbols and ToBI marks were then isolated from the transcripts, transferred to digital spreadsheets and contrasted with the control group’s annotations with the help of the Excel Compare 2.3 software tool. Cohen’s kappa ($k$) was computed to study the similarity of individual annotations with the control group’s data. Jaccard coefficients ($j$) were calculated to measure the reliability of small-to-medium groups’ annotations of eight sample texts. Finally, the intraclass similarity coefficients (ICC, $C$, $k$) for the two experimental groups (Russian, Czech) and three data sets (RPT, traditional (British), ToBI) were computed to determine the level of internal agreement within each experimental group.

### 4.2 Participants

The Russian annotator group comprised 14 second-year BA students of the Moscow Pedagogical State University (Moscow, Russia) majoring in Anglophone Studies and willing to participate in the experiment. The annotations were made in September and December of the 2022/23 winter term, at the beginning and end of the elective Practical Phonetics 3 course, aimed at exploring and practising English suprasegmentals. All the participants (four males, ten females between the ages of 19 and 22) were advanced speakers of English (C1–C2 level$^4$), had a final assessment grade above 74%. The declared level of language proficiency (C1–C2) was guaranteed by a computer-based level test taken by all the participants before the start of the course.

The Czech annotators comprised 12 second-year BA students of the Technical University of Liberec, studying English as one of their two majors. The annotations were conducted within the scope of their SPELB (Experimental Phonology) classes, an elective one-semester course aimed at developing essential analytical and experimental skills in suprasegmental phonology, in September–December 2022. The students of the Czech group (three males, nine females between the ages of 21 and 22) were advanced in English (C1–C2 level) and had successfully passed the final SPELB test (above 70%).

$^4$ According to the Common European Framework of Reference for Languages (CEFRL).
As the description above shows, both experimental groups are perfectly comparable in terms of participants’ age, gender composition, academic performance, and language experience, the only meaningful difference between the groups being the students’ mother tongue.

4.3 The Phonology Training Course

An attempt was made to modify the syllabi of the respective phonetic courses (Practical Phonetics 3/ SPELB (Experimental Phonology)) to minimize the differences between them. Both study courses followed the same syllabus (Appendix 1) and used the same recordings for annotation practice (five texts for traditional (British) annotation and five texts for ToBI practice), amounting to approximately the same amount of instruction time devoted to annotation practice (around four academic hours). The lecturers of both courses were in constant contact to ensure the minimal divergence of the participants’ learning experience.

As for the annotation experience, we include in this the extensive suprasegmental training during a 12-week period, including rigorous auditory recognition practice, contrastive differentiation of tones, working with Praat, and targeted annotation practice, sufficient enough to call the learners phonologically trained annotators in the second stage of the experiment (a “trained ear”). The students whose attendance (max. two absences) and academic performance (lower than 70% in the final test) did not meet the established criteria were excluded from the final counts, which explains a slight discrepancy in the number of participants in the experimental groups.

4.4 The Control Group

The data for non-native speakers’ annotation validation (regarded in this study as the “ground truth”) were acquired from the control group of English native speakers. As it was challenging to find a comparable group of English students doing a Phonology course including both traditional (British) and ToBI annotation practice, we compiled the control group from two different sub-groups, emulating the critical parameters of our experimental groups.

The “amateur ear” sub-group (Sub-group 1) comprised four Erasmus students from the University of Reading (UK), who had similar demographics to those of the experimental groups (three females, one male between 21–22 years old). This group comprised students from the BA Primary Education (QTS) programme, and none of the participants had any previous experience in prosody annotation. This “amateur ear” sub-group completely replicated the steps taken by the experimental groups at Stage 1 of the experiment (RPT annotation).

The “trained ear” sub-group (Sub-group 2) consisted of two experienced (PGCE, DipRSA TEFL, LTCL DipTESOL) British lecturers (one male, one female; 40–55 years old) familiar with both annotation systems (traditional (British) and ToBI), meaning that no additional phonological training was needed. Sub-group 2 provided “consensus” ToBI and traditional (British) annotations, meaning that the experts performed the annotations collectively and had an opportunity to discuss their findings.
4.5 Materials

Four abstracts of recorded public speeches by four British speakers (two male and two female speakers) were chosen for analysis.\(^5\) The recordings can be characterized as semi-spontaneous monologues and public addresses. Fragments of around two minutes (each up to 209–293 words, and 24–29 grammatical sentences long) were extracted from the video clips and converted into the WAV format. For the second experimental stage, the audio fragments were further divided into halves (each up to one minute long, 95–160 words, and 12–15 grammatical sentences). This was done to present the assessors with different texts by the same speakers to eliminate additional variability.

5 Results

5.1 Descriptive Statistics of RPT Prominence Identification

In the first stage of the experiment (with an “amateur ear”, RPT marking), a word was considered prominent and included in the final counts if at least 75% of the participants marked it as prominent (its p-score\(^6\) was > .75). The normalized data (Means and Standard Deviations) for all four texts are presented in Table 1.

Table 1. The number of identified prominent words (%,\(^7\) raw, standard deviation).

<table>
<thead>
<tr>
<th>Prominent Words</th>
<th>Text 1</th>
<th>Text 2</th>
<th>Text 3</th>
<th>Text 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Duration – 1:39</td>
<td>Duration – 1:42</td>
<td>Duration – 1:54</td>
<td>Duration – 1:33</td>
</tr>
<tr>
<td>Speaker</td>
<td>Speaker – JR, female</td>
<td>Speaker – DC, male</td>
<td>Speaker – TM, female</td>
<td>Speaker – JH, male</td>
</tr>
<tr>
<td>Word count</td>
<td>218</td>
<td>293</td>
<td>258</td>
<td>209</td>
</tr>
<tr>
<td>The Russian annotators</td>
<td>30% (67 words) SD – 8</td>
<td>18% (54 words) SD – 9</td>
<td>24% (61 words) SD – 7</td>
<td>21% (44 words) SD – 6</td>
</tr>
<tr>
<td>The Czech annotators</td>
<td>19% (42 words) SD – 11</td>
<td>15% (45 words) SD – 10</td>
<td>18% (47 words) SD – 11</td>
<td>14% (29 words) SD – 10</td>
</tr>
<tr>
<td>The control group</td>
<td>32% (70 words) SD – 4</td>
<td>26% (76 words) SD – 2</td>
<td>27% (69 words) SD – 5</td>
<td>28% (58 words) SD – 2</td>
</tr>
</tbody>
</table>

The RPT test’s results demonstrate that both groups of non-native annotators identified fewer prominent words than the native speakers (Sub-group 1). The Russian group identified, on average, 13% fewer prominent words than the control group, and the Czech group marked roughly 40% fewer.\(^8\)

\(^5\) The videos of the annotated recordings are accessible on YouTube:
- https://www.youtube.com/watch?v=wHGqp8lz36c (3:10–5:03),
- https://www.youtube.com/watch?v=WZI1EjxxXKw (0:19–1:58),
- https://www.youtube.com/watch?v=FDyZ8trge2E (1:30–3:37),
- https://www.youtube.com/watch?v=1N2yk0ubNai (4:00–6:08).

\(^6\) The p-score was calculated as the ratio of annotators who marked the word as prominent to the total number of annotators in a research group.

\(^7\) Percentage of the total number of words in the respective text.

\(^8\) Here and further in the text, the word count reflects the number of orthographic words, not the phonetic words (feet).
The typological (nuclear-accented words vs. prenuclear-accented words) data comparison of the groups’ annotations (Figure 3) revealed that the similarity in the numbers of spotted nuclear accents among all three groups was the highest (10% difference with the control group for the Russian annotators, 24% difference for the Czech annotators).

As for the prenuclear accents, the Czech annotators marked almost no prominent words (9%) in this category (34% difference with the control group), while the Russian annotators identified 21% of prenuclear accents (22% difference with the control group). The problems with prenuclear accent identification are illustrated in Example 5.

Example 5. Prominence identification (p-scores) by various annotator groups (excerpt).

| I cannot (1) remember (1) telling my parents (1) | that I was studying Classics (1) | The RPT annotation by British native speakers identified four prominent words (two prenuclear and two nuclear prominences) in two tone groups separated by a boundary. The high p-scores of all the words in the sample indicate a high level of annotator agreement. |
| I cannot (.86) remember (.26) telling my parents (1) | that I was studying Classics (1) | The RPT annotation by Russian speakers demonstrates a high degree of agreement about the nuclear prominences (“parents” and “classics”) and one prenuclear accent “cannot”. A low p-score for “remember” demonstrates the annotators’ uncertainty. |
| I cannot (.42) remember (.33) telling my parents (.83) | that I was studying Classics (1) | The RPT annotation by the Czech speakers shows their agreement only on marking nuclear prominences (“parents” and “Classics”) and weak/no agreement on marking prenuclear accents (“cannot,” “remember”). |

This sample from the research corpus provides a typical example of the difficulties faced by non-native listeners in the identification of prenuclear accents. All three groups (British native speakers, Russian listeners and Czech listeners) successfully identified nuclear prominences.
("parents," "Classics"), whose acoustic marking (a notable pitch change and increased length) is more or less similar in all the three languages and additionally supported by an expectation-driven shift of contextually important words to the end of the utterance.

In contrast, the marking of prenuclear prominences was problematic. The Russian listeners, who acoustically marked "accidental rises" (Leed 1965) in their mother tongue, noticed the acoustic marking of the first prenuclear prominence ("cannot") but failed to identify the following prenuclear accent ("remember") since its acoustic marking is expected to be different from the mentioned "accidental rise" due to the natural downstep in F0. The Czech listeners (an "amateur ear") ignored the acoustic marking of both prenuclear accents probably because they do not have such acoustic cues as high pitch or increased vowel length at the start of the utterance in their mother tongue.

As for individual annotator performances, two annotators in the Czech group (16% of the experimental group) marked considerably fewer prominences (12%) than the other group members, which can be attributed to the assessors’ misunderstanding of the task (e.g., marking only nuclear prominences), rather than their general inability to identify prominent words in English. In contrast, three assessors from the Russian group (20% of the annotators) and one in the Czech group (8%) marked a proportionately (5–7%) bigger number of accents in all four texts than the other assessors in the respective groups, which made their number of identified prominences closer to the control group’s. Still, although with some exceptions, which are reflected in high SD figures for both experimental groups, we believe there is a general trend in non-native prominence identification pointing towards a lower sensitivity to prominence perception, and prenuclear prominence perception in particular.

5.2 The Assessment of Individual Annotators’ Performances (ToBI vs. Traditional (British) Marking)

In the second stage, the individual performances of non-native phonetically trained annotators were evaluated with the help of Cohen’s kappa (k) by comparing (one-to-one) their annotations with the control group’s readings (Sub-group 2, expert annotators). Cohen’s kappa was then interpreted following McHugh’s (2012) recommendations (Table 2). The extremes of Cohen’s kappa scale are highlighted for the readers’ convenience by italics to indicate “none” to “minimal” agreement, and by bold to indicate “strong” to “almost perfect” agreement.

Table 2. Interpretation of Cohen’s kappa (McHugh 2012).

<table>
<thead>
<tr>
<th>Value of Kappa (k)</th>
<th>Level of Agreement</th>
<th>% of Data that are Reliable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–.20</td>
<td>None</td>
<td>0–4%</td>
</tr>
<tr>
<td>.21–.39</td>
<td>Minimal</td>
<td>4–15%</td>
</tr>
<tr>
<td>.40–.59</td>
<td>Weak</td>
<td>15–35%</td>
</tr>
<tr>
<td>.60–.79</td>
<td>Moderate</td>
<td>35–63%</td>
</tr>
<tr>
<td>.80–.90</td>
<td>Strong</td>
<td>64–81%</td>
</tr>
<tr>
<td>above .90</td>
<td>Almost perfect</td>
<td>82–100%</td>
</tr>
</tbody>
</table>
Since the annotators had an option not to mark the accent type if they were unsure of its type or they did not have a relevant category in the suggested decision-making algorithm (Figure 1) to reflect the type of accent they heard, the number of accents effectively identified in the second stage (a “trained ear”) was somewhat lower than the total number of prominent words pre-marked for analysis (210 words). The number of words eventually annotated by the participants was 94% (198 words) for the Russian group and 96% (202 words) for the Czech group, with annotators claiming they would need more passes to complete the task.

The Russian group’s performance results (Table 3) demonstrate that only one-fourth of the Russian group's Cohen's kappa coefficients (in bold) fit into the category of “strong agreement” (with the control group readings), with the mean agreement value for the whole group (.67) carrying “moderate” similarity.

Table 3. The Russian group’s Cohen’s kappa ($\kappa$).

<table>
<thead>
<tr>
<th>The Russian group</th>
<th>ToBI annotation</th>
<th>Traditional (British)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text 1</td>
<td>Text 2</td>
</tr>
<tr>
<td>Annot. 1</td>
<td>.67</td>
<td>.81</td>
</tr>
<tr>
<td>Annot. 2</td>
<td>.55</td>
<td>.42</td>
</tr>
<tr>
<td>Annot. 3</td>
<td>.44</td>
<td>.60</td>
</tr>
<tr>
<td>Annot. 4</td>
<td>.24</td>
<td>.30</td>
</tr>
<tr>
<td>Annot. 5</td>
<td>.60</td>
<td>.89</td>
</tr>
<tr>
<td>Annot. 6</td>
<td>.80</td>
<td>.77</td>
</tr>
<tr>
<td>Annot. 7</td>
<td>.68</td>
<td>.68</td>
</tr>
<tr>
<td>Annot. 8</td>
<td>.78</td>
<td>.80</td>
</tr>
<tr>
<td>Annot. 9</td>
<td>.32</td>
<td>.42</td>
</tr>
<tr>
<td>Annot. 10</td>
<td>.67</td>
<td>.81</td>
</tr>
<tr>
<td>Annot. 11</td>
<td>.60</td>
<td>.68</td>
</tr>
<tr>
<td>Annot. 12</td>
<td>.68</td>
<td>.81</td>
</tr>
<tr>
<td>Annot. 13</td>
<td>.32</td>
<td>.70</td>
</tr>
<tr>
<td>Annot. 14</td>
<td>.60</td>
<td>.81</td>
</tr>
</tbody>
</table>

There is an evident trend towards the Russian group’s traditional (British) annotations’ higher reliability (Mean: .77; Median: .78; Mode: .71) than their ToBI markings (Mean: .57; Median: .60; Mode: .60), with nearly one-third of the assessors (28%) providing firm prosodic markings in at least three out of four sample texts. The values for these selected assessors (Annotators 3, 6, 7 and 13; traditional (British) annotation) reach the following peaks: Mean: .84; Median: .83; Mode: .80. The ToBI markings of the Russian group show “moderate” reliability values (Mean: .57; Median: .60; Mode: .60) and contain some “weak” agreement coefficients (7%, in italics).

The irregular distribution of incorrect annotations (highlighted by italics in Table 3), in our view, does not reflect any systemic annotation problems and proves their occasional character, except for Annotator 4, whose low kappa in all four texts (ToBI) most likely signals a misunderstanding of the basic ToBI annotation principles.
McHugh’s (2012) interpretation of Cohen’s kappa (Table 2) allows us to conclude that the Russian group’s traditional (British) markings signal a relatively high degree of data reliability (60%–65% of data is fully reliable). For some annotators (roughly one-third of the group), the level of reliability goes as high as 70% for traditional (British) annotations (bold in Table 4). In contrast, the group’s ToBI marks are of worse quality (30%–33% of reliable data).

*The Czech group’s performance* (Table 4) reflects an almost even distribution of both high (29%, in bold) and low (14%, in italics) Cohen’s coefficients across the annotators in both annotation types. The readings demonstrate slightly higher kappa values of the group’s ToBI markings (Mean: .65, Median: .68, Mode: .68) than traditional (British) markings (Mean: .59, Median: .60, Mode: .60). The percentage of unreliable assessments is 14%, which is twice higher than in the Russian group.

**Table 4.** The Czech group’s Cohen’s kappa ($\kappa$).

<table>
<thead>
<tr>
<th>The Czech group</th>
<th>ToBI annotation</th>
<th>Traditional (British)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text 1</td>
<td>Text 2</td>
</tr>
<tr>
<td>Annot. 1</td>
<td>.78</td>
<td>.56</td>
</tr>
<tr>
<td>Annot. 2</td>
<td>.67</td>
<td>.70</td>
</tr>
<tr>
<td>Annot. 3</td>
<td>.80</td>
<td>.68</td>
</tr>
<tr>
<td>Annot. 4</td>
<td>.74</td>
<td>.88</td>
</tr>
<tr>
<td>Annot. 5</td>
<td>.20</td>
<td>.42</td>
</tr>
<tr>
<td>Annot. 6</td>
<td>.68</td>
<td>.92</td>
</tr>
<tr>
<td>Annot. 7</td>
<td>.56</td>
<td>.70</td>
</tr>
<tr>
<td>Annot. 8</td>
<td>.46</td>
<td>.36</td>
</tr>
<tr>
<td>Annot. 9</td>
<td>.82</td>
<td>.92</td>
</tr>
<tr>
<td>Annot. 10</td>
<td>.76</td>
<td>.70</td>
</tr>
<tr>
<td>Annot. 11</td>
<td>.17</td>
<td>.13</td>
</tr>
<tr>
<td>Annot. 12</td>
<td>.72</td>
<td>.86</td>
</tr>
</tbody>
</table>

The character of the invalid annotations’ distribution (highlighted by italics in Table 4) among the Czech assessors reflects their non-systematic occurrence, except for Annotators 8 and 11, whose consistently low similarity coefficients in seven out of eight samples may be the result of either limited annotation practice or the annotators’ insufficient language exposure, which, as it was mentioned (Weingartová, Poesová, and Volín 2014), can seriously impede Czech students’ English prominence perception.

In contrast, some annotators (e.g., Annotators 4, 9, and 12) demonstrated a sustained ability to mark accent types in both annotation systems, with the following averages for the selected group: Mean: .80, Median: .81, and Mode: .80. These annotators’ successful performance (as well as the selected reliable annotations of the Russian team), may prove the positive effects of preliminary phonological training on English prominence perception. The successful annotations can also be attributed to the students’ high English proficiency (C1–C2).
In conclusion, the recommended quantitative interpretation of Cohen’s kappa (McHugh, 2012) signals the moderate reliability of the Czech group’s ToBI annotations (48% – 55% of reliable data) and weak reliability of their traditional (British) assessments (30–35%).

5.3 The Assessment of Groups’ Collective Performances: Jaccard Similarity Indices

To assess the collective reliability of 12/14-member teams made up of non-native phonologically trained annotators, Jaccard similarity indices ($J$) were computed for eight texts (four were marked in ToBI and four using traditional (British) annotations). The experimental group’s most common annotation mark for each of the pre-marked 210 words was compared to the respective control group’s marking.

The computed Jaccard similarity indices for both annotator groups (Figures 4 and 5) were evaluated following recommendations for the similarity coefficients’ evaluation (Verma and Aggarwal 2020), with values of 1%–20% indicating “none” to “slight” similarity, 21%–40% – “fair,” 41%–60% – “moderate,” 61%–80% – “substantial,” and 81%–100% – “almost perfect” similarity.

The results of the Russian group’s comparison (Figure 4) demonstrate “moderate” to “substantial” levels of annotation similarity with the native speakers’ data, with a maximum of 80% achieved in one of the four texts annotated with the help of traditional (British) annotation symbols. The ToBI annotations (ranging from 45% to 65%) in three out of the four ToBI marked texts demonstrate a slightly lower level of annotation similarity than do the traditional (British) annotations (65% to 80%), which is quite in line with the group’s Cohen’s kappa values.

The Czech group’s Jaccard indices (Figure 4) demonstrate “moderate” to “substantial” similarity with the control group’s marking. The experimental group had a lower similarity

![Figure 4. The Jaccard ($J$) similarity indices for the Russian annotations.](image)

The Czech group’s Jaccard indices (Figure 4) demonstrate “moderate” to “substantial” similarity with the control group’s marking. The experimental group had a lower similarity in the annotations of two texts performed using traditional (British) symbols (43%–63%). At the same time, the ToBI annotations done by the Czech assessors showed a consistently better correlation with the native speakers’ data in all four texts (65%–82%). These findings also support the evidence of the group’s better ToBI performance calculated with the help of Cohen’s kappa values.
in the annotations of two texts performed using traditional (British) symbols (43%–63%). At the same time, the ToBI annotations done by the Czech assessors showed a consistently better correlation with the native speakers’ data in all four texts (65%–82%). These findings also support the evidence of the group’s better ToBI performance calculated with the help of Cohen’s kappa.

5.4 Intragroup Reliability Coefficients

Finally, a two-way, consistency, average-measure intraclass correlation coefficient (ICC (C, k), with k standing for the number of participants in the group) was computed to examine the internal agreement within two experimental groups of assessors. The number of identical assessments of a unit (an accent type) across the group in tree data sets (RPT, ToBI, traditional (British)) was referred to the total number of assessments of a particular unit (word) and statistically normalized by the number of group members (Table 5).

Table 5. ICC (C, k) reliability coefficients.

<table>
<thead>
<tr>
<th>Stage 1 (RPT annotation)</th>
<th>Stage 2 (traditional (British) marking)</th>
<th>Stage 2 (ToBI annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Russian group, k=14</strong></td>
<td>.88</td>
<td>.71</td>
</tr>
<tr>
<td><strong>The Czech group, k=12</strong></td>
<td>.76</td>
<td>.62</td>
</tr>
</tbody>
</table>


The groups’ RPT annotations, naturally, have higher intraclass correlation indices (.76 – the Czech group, .88 – the Russian group), as the annotators had a limited choice of options.
(“stressed” vs. “unstressed” word). At the same time, since the number of options doubled at the second stage (an “experienced” ear), the intragroup agreement coefficients went slightly down (min. = .62, max. = .71), while still remaining high enough to substantiate the validity of the groups’ results.

6 Discussion

6.1 The Performance of the Russian and Czech Experimental Groups vs. the Control Group

The results of the two-stage perceptual experiment demonstrate that Russian and Czech speakers are generally less sensitive to prominence perception in English monologues than native speakers of English. It is assumed that the fact that both experimental groups identified at least 20% fewer prominent words and had 7–14% unreliable readings can be explained by either their insufficient phonological training (Stage 1) or some cumulative effects of the prosodic differences between the speakers’ native languages and English that interfere with English prominence perception.

In line with earlier experiments (Bishop, Kuo, and Kim 2020), which prove that annotators identify specific types of prominence more easily than others, our findings demonstrate that nuclear-accented words are generally more noticeable for non-native annotators than prenuclear prominences. We believe that the relative similarity of terminal pitch movements on nuclear accents, the terminal tones’ delimitating function, and the similarity of their pragmatic load in all three languages lead to their better recognition.

In contrast, prenuclear accents in non-native speaker annotations demonstrate significantly lower recognition indices, probably because they carry primarily acoustic (local/signal-driven) cues, which are believed to be more difficult to perceive for non-native speakers than expectation-driven semi-global and global factors. The Russian annotators did not recognize one-fifth of prenuclear accents, and the Czech group failed to identify one-third. The results of the Russian group are believed to be better because of the similarity of certain local acoustic features in Russian and English syllables, facilitating prenuclear accent identification.

6.2 The Czech/Russian Languages’ Prosodic Interference in English Prominence Perception

We believe that the closeness of the acoustic marking of English and Russian accented words, as shown in some respects in earlier research – such as peak pitch values, noticeable stressed vowel’s positional lengthening, and higher loudness on the focus words – contributed to the Russian group’s better English prominence recognition, accounting for a higher number of successfully annotated words in both experimental stages. The reported wider pitch range inherent in Russian intonation contours (another marked prosodic feature in Russian) may also assist in the Russian annotators’ better recognition of English accents. Additionally, phonological contrasts that result from the reduction of unstressed vowels, present in English and Russian, are generally thought to facilitate the Russian group’s accent recognition.
In contrast, the Czech annotators, who, according to the aggregated data, have neither peak pitch values nor positional vowel lengthening on the stressed syllables in their L1, found English prominence perception a bit more challenging. In our opinion the main difficulty, which may generally serve as an impediment to more successful prominence identification for the Czech speakers of English, lies in the absence of vowel reduction in unstressed syllables as a marked feature in the Czech language. This accounts for the Czech speakers’ lower sensitivity to vowel-length contrasts, although it is a valuable acoustic cue for English native speakers. The narrower pitch range in the Czech language reported in earlier studies may also significantly complicate the perception of English prenuclear accents at the beginning and, particularly, in the middle of the intonation phrase, with high pitch, although significant for the native speakers of English in marking the prenuclear accents, serving as just an irrelevant distractor for the Czech speakers.

6.3 The Effects of Using Different Annotation Techniques

The decision to compare and contrast different methods of prominence marking in the present study was made because earlier comparative studies of annotation types in the phonology-teaching context showed no significant difference in their classroom application, and left the choice of annotation system to the theoretical preferences of the teacher (Toivanen 2005). However, we believe that students’ L1 may also influence the researcher’s potential choice of annotation technique. The chosen annotation method may better reflect the non-native annotators’ dominant L1 prosodic cues and, therefore, may provide higher annotation precision in the target language.

The use of RPT brought highly reliable results with untrained assessors. In line with earlier studies (Cole and Shattuck-Hufnagel 2016), we thus recommend its wider application with annotation novices, mainly because RPT does not require much preparatory practice.

As for ToBI and traditional (British) annotation methods, the acquired readings show mixed results. Noticeable discrepancies in the ToBI vs. traditional (British) annotations’ similarity indices demonstrate the Russian assessors’ better performance with the traditional (British) annotation technique, which, as expected, better reflects “top-down” contextual cues. In contrast, the Czech group had consistently better results in the ToBI marking, likely due to the instilled practice of “down-up” acoustic cue perception. Nevertheless, it seems too early to draw conclusions about the students’ preferences for one or the other system, as more factors, such as their annotation experience and the quality of annotation teaching, should also be considered.

6.4 Non-native Speakers’ Potential in EFL Intonation Research Validation

Our results demonstrate that the experimental groups’ annotations can be characterized by “moderate” (60%) to “substantial” (70%) similarity with native speakers’ assessments, with the number of incorrect annotations (7%–14%) being only slightly higher than the results seen in earlier studies (80% reliability, 5% of incorrect assessments, for the Russian assessors in Potapova and Potapov (2006)).

The computed intraclass correlation coefficients show “good” to “excellent” data homogeneity of the 12/14-member groups in all three types of tasks (RPT, traditional (British), ToBI).
These alignment coefficients are in line with earlier findings (Cole, Mahrt, and Roy 2017), which means that attracting medium-sized annotator groups to data validation is reasonably practical.

Attracting 12 to 14 annotators to data validation allows the researcher to overcome the effects of individual perceptual differences and makes it possible for small to medium-sized groups of non-native speakers to be employed for cross-language validation of prominent words in various experimental studies involving EFL speakers. However, it should be noted that prior phonological training, a general level of foreign language proficiency, and a suitable annotation method, should also be considered when forming a non-native annotation team.

7 Limitations

The present research had the following limitations that were addressed with great care in the present paper and should be addressed in further studies.

The experimental design of annotation tasks needs to consider the issue of the annotators’ English language exposure and annotation experience, as closer attention to these parameters may provide explanations for several annotation mistakes. Our study addressed the issue of language proficiency by conducting a computer-based level test in both research groups. Students’ attendance scores also served as a criterion for their participation in the final research stage.

Giving annotators more time to think (while keeping the same number of passes), excluding pre-marked words from the transcripts, and departing from the pre-set algorithm of prominence marking will potentially increase the variability of identified accents and prosodic cues the listeners rely on. On the other hand, using binary choices as we did in our study (the suggested binary-choice algorithm in Figure 1) ensures the homogeneity of research standards for the experimental groups and the materials to be annotated.

Finally, the occasionally mixed record of ToBI vs. traditional (British) annotations evident from our study calls for more attention to the choice of annotation technique in experimental papers. Despite the apparent preference of certain pronunciation instructors for the use of traditional (British) annotation principles, the vast majority of current intonation research uses the ToBI annotation method, and students should have enough in-class annotation practice to be able to confidently use this system in their research papers.

8 Conclusion

The statistical data acquired in this study, namely the sufficient similarity and reliability coefficients, demonstrate that language proficient and phonologically trained non-native speakers of English can effectively recognize and identify word prominence and basic accent types in English monological speech. As suggested in earlier papers (McAndrews 2019; Snow et al. (2008)), the results of phonologically trained non-native listeners’ perceptual analysis, although with certain reservations, can be successfully used to validate English intonation research in non-English speaking countries where English is predominantly spoken as a second or foreign language.
The study results demonstrate that trained non-native auditors can use various annotation techniques, such as Rapid Prosody Transcription, ToBI marking, and traditional (British) annotation, to mark key prosodic phenomena in English. Although neither annotator group in our study expressed a clear preference for one or another annotation system, it can be assumed that the choice of annotation technique for educational or research purposes should be linked to the annotator’s prior phonological training.

Our findings also show that positive intonation transfer (i.e., easier recognition of similar, often marked features) may contribute to a better perception of the prosodic cues similar in the annotators’ native (Russian and Czech) and target (English) languages. Despite the negative L1 interference, evident from the annotators’ more problematic recognition of the prosodic features unmarked in their native languages (e.g., prenuclear prominence), the experimental groups of Czech and Russian speakers still demonstrate a sustained ability to mark English prominence, which, presumably, can be further improved through additional language and annotation practice.

In conclusion, we believe that more cross-language perceptual studies involving annotators with different native languages are needed to specify the cumulative role of various prosodic cues in English learners’ intonation recognition, as their results can help EFL teachers focus prosodic instruction on the aspects crucial for language intelligibility.

Acknowledgements

The author expresses profound gratitude to the colleagues who helped collect data and conduct the present experiment.

References


### Appendix 1. The Phonology Course Syllabus

<table>
<thead>
<tr>
<th>Study Week</th>
<th>The topics discussed:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td>Introduction to experimental phonology. Methods of phonetic and phonological analysis. Phonetic and phonological mistakes. Acoustic and auditory (perceptual) analysis of speech.</td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td>Components of pronunciation. Suprasegmentals. Syllabic structure of a language. Falling and rising tones. Fall/rise perception and production practice.</td>
</tr>
<tr>
<td><strong>Week 4</strong></td>
<td>Stress in English and in Czech/Russian. Perceptual and acoustic differences. Stress identification practice.</td>
</tr>
<tr>
<td><strong>Week 8</strong></td>
<td>Traditional (British) annotation. Speech melody. Rhythm. Differences between Czech/Russian and English intonation. Annotation practice.</td>
</tr>
<tr>
<td><strong>Week 10</strong></td>
<td>Introduction to the computer analysis of speech. Praat analysis. Acoustic correlates of perceptual parameters. Annotation practice.</td>
</tr>
<tr>
<td><strong>Week 11</strong></td>
<td>Autosegmental phonology. Introduction of ToBI. Pitch accents and boundary tones. Annotation practice.</td>
</tr>
<tr>
<td><strong>Week 12</strong></td>
<td>Regional intonation variation. ToBI annotation practice. Comparison of ToBI and traditional (British) system. Annotation practice.</td>
</tr>
</tbody>
</table>