UNIVERZITA KARLOVA

Filozofická fakulta

Ústav anglického jazyka a didaktiky

DIPLOMOVÁ PRÁCE

Bc. Michaela Banýrová

The Correlations between Perceived Fluency and Productive Fluency in the Speech of Advanced Czech Speakers of English

Korelace mezi percepční plynulostí a verbální plynulostí v projevu pokročilých českých mluvčích angličtiny

Praha 2019 Vedoucí práce: PhDr. Tomáš Gráf Ph.D.

Poděkování

Ráda bych poděkovala vedoucímu práce, PhDr. Tomáši Gráfovi, Ph.D., za cenné podněty, vstřícnost, trpělivost a podporu.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 5.srpna 2019 ……………………………....

Michaela Banýrová

Abstrakt

Diplomová práce se zabývá tématem plynulosti žákovského jazyka, konkrétněji verbální a vnímanou plynulostí. Plynulost žákovského jazyka, plynulost českých žáků angličtiny nevyjímaje, nebyla dosud dostatečně prozkoumána. Cílem práce je určit, zda a do jaké míry korelují verbální plynulost, reprezentovaná tempem řeči, a vnímaná plynulost, reprezentovaná hodnocením posluchačů, a lépe porozumět procesu hodnocení plynulosti posluchačem. K analýze byly použity vzorky nahrávek z korpusu mluveného žákovského jazyka LINDSEI, pro něž bylo spočítáno tempo mluvy ve slovech za minutu, hodnocení plynulosti těchto vzorků rodilými mluvčími angličtiny na 7stupňové škále a komentáře hodnotitelů k procesu hodnocení. Analýza ověřuje hypotézu, že tempo řeči je jednou z několika složek, které ovlivňují vnímanou plynulost mluvy. Výsledky ukazují, že tempo řeči ovlivňuje vnímanou plynulost, ale v menší míře, než ukazuje předešlý výzkum a jednotliví posluchači se ve svých hodnoceních výrazně liší. Korelace pak byly nalezeny jen v případě některých hodnotitelů. To ukazuje, že plynulost je velice subjektivní, komplikovaný pojem, a další výzkum vnímané plynulosti, respektive jejích složek, je zásadní pro výuku jazyka a plynulosti jako takové.

Klíčová slova

Plynulost, verbální plynulost, vnímaná plynulost, tempo řeči, mluvený jazyk, žákovský korpus

Abstract

The present thesis is concerned with the topic of fluency in learner language, more precisely of two types of fluency - perceived and productive. Little is known about L2 fluency, especially about the fluency of Czech learners of English. The main aim of the thesis is to establish whether there is a correlation between productive fluency, represented by speech rate, and perceived fluency, represented by native speakers' evaluations. In addition, it aims at better understanding the process of evaluation of perceived fluency by native speakers of English. The material for the analysis were samples of recordings from the LINDSEI corpus, for which speech rate in WPM was calculated, evaluations of fluency of these samples by native speakers of English on a 7-point scale and the raters' commentaries on the evaluation process. The analysis tries to prove or disprove the hypothesis that speech rate is one of the features which influence perceived fluency. The results show medium correlations for two raters, low or no correlations for the rest of the raters, showing together with the commentaries, that there is a relation between perceived fluency and speech rate, but it is not as strong as previous research suggests. The results show that fluency is a complicated, highly subjective phenomenon, and further research of perceived fluency is essential for ELT and for teaching fluency.

Keywords

Fluency, productive fluency, perceived fluency, speech rate, spoken language, learner corpus

Table of contents

List of abbreviations

ALP = average length of pause

AR = articulation rate

CEFR = common European framework of reference

EFL = English as a foreign language

ELT = English language teaching

ESL = English as a second language

L1 = first language

L2 = second language

LINDSEI = Louvain International Database of Spoken English Interlanguage

MLR = mean length of run

NNS = non-native speaker

NS = native speaker

phw = per hundred words

PSR = pruned speech rate

PTR = pause-time ratio

SD = standard deviation

SR = speech rate

WPM = words per minute

List of figures

1. Introduction

Fluency is a key component of the mastery of a language. "To speak a language fluently" is a common expression. As frequent as the term *fluency* is, neither the general public, nor the academics agree on what is meant by the term. Fluency can be used as an equivalent to overall spoken proficiency, as well as a more specific term, used for example in the model of proficiency consisting of complexity, accuracy and fluency, e.g. Skehan (1998). To be able to teach fluency or to improve students' fluency, it is necessary to understand the phenomenon, to know what its components are and what it is influenced by. Fluency is also one of the categories in which students are evaluated in language tests, based on the perception of the examiner. This gives even more reason to study fluency, to be able to provide an objective measure of fluency for language testing, so that students are evaluated based on clearly given, precise and objective measures.

The present thesis examines two types of fluency, productive and perceived. Productive fluency is viewed from the point of view of the act of speech production, it can be measured using a wide variety of measures, analysing the speed of speech, numbers and distribution of repairs or speech breakdowns. Perceived fluency is concentrated at the point of view of the listener, it is concerned with how fluent the speaker is perceived by the listener. More precisely, the thesis aims at establishing, whether there is a correlation between productive fluency and perceived fluency. Previous research has shown several aspects of productive fluency to be predictors of perceived fluency, speech rate being one of the most prominent, e.g. Kormos & Dénes (2004) or Derwing et al. (2004). The thesis aims at verifying the hypothesis that speech rate is one of the most prominent predictors of perceived fluency and at learning more about perceived fluency in general and about the process of evaluation of fluency.

In chapter 2, the theoretical background for the phenomenon of fluency is provided, mainly for the two types studied in the thesis. However, influential authors' views on fluency in general are given, their division of fluency into types as well as their definitions. In addition, ways of operationalizing productive and perceived fluency are given, showing positives and negatives of different measures. The material and method used in the thesis are described in chapter 3. Samples are taken from the Czech part of the LINDSEI corpus, which means that the speakers are advanced Czech learners of English. Speech rate is calculated for the samples (in WPM) and the same samples are evaluated by five native speakers of English with some experience in teaching English as a foreign language. The raters are also asked to comment on the process of evaluation and on prominent features for ten samples. The total of 35 samples are evaluated by 5 raters, giving a total of 50 commentaries and 175 numerical evaluations. Chapter 4 contains the research questions, the results and their analyses are presented in chapter 5. The data are analysed using qualitative as well as quantitative method, giving not only the Pearson correlation coefficient, but also an insight into the evaluation process. The results are discussed and their consequences are outlined in the discussion in chapter 6 of the thesis, together with implications for teaching and limitations of the thesis and suggestions for further research.

2. Theoretical background

2.1 Research on fluency and its definitions

Fluency has been a problematic concept in terms of its definition and identifying its components. A number of works by various authors have been devoted to the topic of fluency. However, their opinions on what aspects are a part of the phenomenon and how it can be categorized and measured, differ considerably. In this chapter, we will attempt to define fluency as it will be viewed in the present work, referring to authors who had defined it before.

One of the difficulties of defining fluency lies in the fact that the term itself is metaphoric and many of the definitions provided in literature (especially in older works) draw on the metaphoric expression and do not give any clear description of what is meant by the term. Segalowitz (2010) addresses this issue, pointing out the positives of thinking of language as motion, such as the metaphor helping us to imagine fluency and its aspects, although he also warns against such descriptions of fluency, as they cannot be sufficient and to fully understand a concept, we need to be able to describe its aspects with precision, in objective measures: "Ultimately, if fluency is to be fully understood, notions like "fluidity," "smoothness," "coordination" will have to be operationalized" (Segalowitz, 2010, p. 179)

One of the first and most influential authors to have studied fluency is Lennon (1990), he distinguishes two types of fluency: fluency in a broad sense and a narrow sense. Fluency in a broad sense according to Lennon is equivalent to overall language proficiency. In this view, a fluent speaker of a language has perfect control of the language, its grammar, lexicon, etc., being fluent in a language in the broad sense means being perfectly capable of speaking the language. Fluency in its narrow sense is defined by Lennon as "one, presumably isolatable, component of oral proficiency" (Lennon, 1990, p. 389) and he describes it as a component of language frequently used in oral language examinations, together with categories such as correctness, pronunciation, lexical range, which in the broad sense would be subcategories of fluency, while in the narrow sense these are aspects of language proficiency, coexisting with fluency at the same level.

As Witton-Davies (2014) mentions, Lennon (2000) later complicates things by renaming the categories to higher order and lower order fluency, corresponding to broad sense and narrow sense of fluency respectively, and naming fluency in the narrow sense "false fluency" (Lennon, 2000, p. 28), explaining that this fluency is based on automatization of simple phrases. This corresponds to Schmid's (1983) case of Wes, who managed to speak fluently in the narrow sense of fluency, but his language was characterized by very simple, incorrect grammar (e.g. the use of present continuous tense for expressing most temporal relations, past and future included). However, Schmid (1983) considers fluency as distinguishable from accuracy and complexity, pointing out that even a person capable of using only simple phrases with many mistakes can be fluent, while Lennon (2000) seems to consider this kind of fluency as inferior to fluency in the broad sense.

In addition, Witton-Davies (2014) describes Chambers's (1997) line of reasoning as similar to Lennon's, as she turns from distinguishing between fluency and overall oral proficiency towards the opinion that syntactic complexity has to be considered a feature of fluency. However, Witton-Davies (2014) finds an argument against such understanding of fluency, supporting it by a different interpretation of a study by Towell et al. (1996), concluding, unlike Chambers, that fluency needs to be studied in context, with regard to genre and subject matter of the utterance, as it is more difficult to reach the same fluency with more complex structures and the same speaker will show different levels of fluency in speeches of different complexity. Therefore, it is not necessary to consider complexity a part of fluency, but it is necessary to take into account the complexity of the utterance the fluency of which is being studied.

Another important author to have studied fluency is Fillmore (1979), who concentrated on native speaker fluency. He distinguishes four types of fluency, the first of which is "the ability to fill time with talk" (Fillmore, 1979, p. 93), and the three following types include coherence, semantic density, having appropriate things to say and creative language use. It is not clear, whether Fillmore's categories can be useful for studying second language fluency and therefore whether they can be useful for this thesis, however, other definitions and categorizations seem to be more relevant.

Witton-Davies (2014) comments on Fillmore's categories in disagreement by stating that calling a fast speaker fluent is reasonable, even if the speaker lacks content density or originality. On the contrary, no matter how original and dense the utterances, if spoken slowly and hesitantly, their speaker would not be called fluent. This suggests that Witton-Davies considers the speed and lack of hesitation or pauses as a more important part of fluency than sophistication, density, creativity or any other aspect connected more with the knowledge than the production of speech. A more useful categorisation of fluency for the study of fluency of learner language is that presented by Segalowitz (2010), where he distinguishes between three types of fluency: cognitive fluency, utterance fluency and perceived fluency.

2.1.1 Cognitive fluency

Segalowitz explains that it is impossible to understand which features of oral performance are a part of fluency, when we are looking at fluency as one phenomenon. We have to look at different types of fluency separately and study its different aspects to be able to understand what influences speakers' fluency. The first type of fluency he looks at is cognitive fluency, he defines it as the "ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances" (Segalowitz, 2010, p. 48). Several processes need to be at play for a speaker to produce an utterance, cognitive fluency is the ability to coordinate such cognitive processes efficiently, so that utterances can be produced smoothly without too much hesitation. Such processes involve planning what we want to say, retrieving the appropriate lexis, putting it into grammatical form, activating the articulatory system, etc. Although we include this type of fluency to provide a complete overview, it will not be a subject of the present study.

2.1.2 Utterance fluency

Utterance fluency can be defined in terms of the features or characteristics of an utterance (Segalowitz, 2010). The number of features, which influence utterance fluency, is still unclear and researchers are trying to establish, which do and which do not influence fluency considerably. As the number of features can be rather large, it is important to examine the relative importance of particular features and find those features that are crucial. So far, researchers mostly seem to agree on the importance of certain measures (e.g. speech rate), but to disagree on others. Even the ways of operationalizing a particular feature can vary, e.g. speech rate can be measured in words per minute, syllables per second, etc. According to Segalowitz (2010, p. 48), utterance fluency "refers to the temporal, pausing, hesitation, and repair characteristics," he describes them as "actual properties of the utterance, not just impressions a listener might have" to contrast utterance fluency with perceived fluency.

Skehan (2003, 2009) introduces a categorization of fluency based on the components which need to be distinguished in order to obtain effective measures. Although Skehan does not use the term utterance fluency, we place his distinction under utterance fluency as it clearly serves for measuring this fluency type. He argues that to measure fluency correctly, we need to take measures in the three following areas: breakdown fluency, repair fluency and speed fluency. Breakdown fluency means to measure the amount of silence in the utterance, the number and length of interruptions, repair fluency stands for measuring the number of repetitions, corrections, false starts, etc. in the utterance and speed fluency for measuring the speech rate.

Another term, by which some authors (e.g. Götz) refer to this type of fluency is productive fluency. She defines it as "features that relate to speech production" (Götz, 2013a, p. 13) and in order to describe productive fluency, she describes "features that establish fluency on the part of the speaker" (Götz, 2013a, p. 13). It could be argued that Götz's productive fluency covers not only Segalowitz's utterance fluency but also his cognitive fluency, as Götz distinguishes between fluency from the part of the speaker and the listener but does not distinguish between the process underlying the production of speech and the product and its features. She mainly concentrates on the features by which fluency can be described and through which it can be examined, not on the processes, therefore we could argue that Segalowitz's cognitive fluency is implicitly included in Götz's productive fluency but does not play a significant part in it.

Götz introduces the term "fluencemes of production" (Götz, 2013a, p. 14), which refers to the features that enable a thorough description of productive fluency. Such fluencemes include temporal variables (such as speech rate, mean length of run, etc.) and strategies that native speakers use to reduce the pressure of producing an utterance, e.g. formulaic sequences and performance phenomena. Formulaic sequences are chunks of language, stored and retrieved as single units and automatized, so that the speaker does not need to retrieve them one word by another and devote part of the brain capacity to the grammatical relations between the words. Performance phenomena are features of unplanned speech, dysfluencies such as filled pauses, repetitions, self-corrections. There is a tendency to regard such features as negative but they should not be regarded so, given that they contribute to the natural sound of speech.

Götz (2013a) divides the fluencemes into two groups: those which always occur in speech production (e.g. speech rate – an utterance must always be characterized by its speech rate) are called primary variables, those which do not have to occur in an utterance (e.g. discourse markers as it is possible to have an utterance without discourse markers) are called secondary variables. She also points out that Lennon (1990, p. 388) uses the same distinction but calls these variables "core and peripheral fluency variables".

2.1.3 Perceived fluency

According to Segalowitz (2010, p. 48), perceived fluency is defined as the "inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency," which corresponds to Lennon's (1990) view of fluency as "impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (Lennon, 1990, p. 391). Both quotes show the relation of perceived fluency to the listener, and to other types of fluency. Segalowitz's quote shows the interconnection between all three types of fluency, cognitive fluency being defined as the processes behind the creation of an utterance, utterance fluency as the features of thus produced utterance and perceived fluency as the listeners impressions about the processes that result in the utterance he/she hears. Similarly, Lennon mentions the psycholinguistic processes, which correspond to Segalowitz's cognitive fluency. He does not use utterance fluency in his definition (nor its equivalent, as he does not use the term utterance fluency at all), however, later in the same paragraph, he speaks about "a finished product" (Lennon, 1990, p. 391) by which he refers to an utterance without disfluencies which enables the listener to concentrate on the message and not the form, and with that he brings the third component of fluency as described by Segalowitz and we can therefore say that their views of components of fluency are very much in accordance.

While Götz (2013a) agrees with Lennon's (1990, p. 391) definition of fluency (her term is perceptive fluency) as "an impression on the listener's part," she disagrees with both Lennon (2000) and Segalowitz (2010) on which features actually influence the listener. She states that "listeners' judgements on productive fluency performance, for instance, the number and positions of temporal variables like unfilled pauses" (Götz, 2013a, p. 45) are not easy to detect by listeners and she introduces the term "fluencemes of perception" (Götz, 2013a, p. 45), by which she labels the features that in her opinion contribute much stronger to the perception of the fluency of a speaker. In addition, she calls this type of fluency perceptive, a term similar but not identical to Segalowitz's term. Her fluencemes of perception include accuracy, idiomaticity, intonation, accent, pragmatic features, lexical diversity and sentence structure. To be able to judge which of the researchers is right, we will look at more studies on perceived fluency in the following chapter.

Unlike productive (or utterance) fluency, most authors agree on the definition of perceived fluency – the definitions used in the majority of research on perceived fluency are those by Lennon and Segalowitz (stated above). What the researchers do not agree upon are the components of perceived fluency, the features of speech which make the listener consider the speech fluent or disfluent. The possible phenomena which may to different extent influence perceptions of fluency include speed of speech, pauses, the length of runs, repetitions, ease, naturalness and appropriateness, pronunciation, grammar, lexical variety, etc. (E.g. Riggenbach, 1991; Rossiter, 2006; Préfontaine & Kormos, 2016). For more detail see section 2.3 Operationalization of perceived fluency.

2.1.4 Summary

The terms which will be used in the present thesis are productive fluency and perceived fluency. As productive fluency, we will consider both Götz's definition and features of productive fluency and Segalowitz's definition of utterance fluency. We will also keep in mind Skehan's distribution of the features of productive fluency. Perceived fluency will be regarded as the listeners impressions about the speech they hear and the processes underlying the production of such speech (complying with the definitions by Segalowitz, Götz and Lennon). The reason for using the term "productive fluency" instead of "utterance fluency", which has been used by more authors, is that the thesis focuses on fluency in L2 speech and implications of fluency research for ELT. The word "productive" keeps the learner in the picture, it is the fluency with which the learner produces speech, while the term "utterance fluency" seems to exclude the learner and concentrate uniquely on the product he or she produces.

2.2. Operationalization of productive fluency

Having explored the definitions and categorizations of fluency, it is equally important to look at different ways of operationalizing fluency, to explore how fluency has been studied and measured. In this section we will explore the quantitative aspects of fluency, i.e. the ways of measuring productive fluency – which aspects can be measured, what units can be used and how the measurements can be combined. Although the authors we will be referring to do not usually use the term "productive fluency" but rather "utterance fluency" or simply "fluency", by all of these terms the same type of fluency is meant, and that is what we call "productive fluency" and what has been defined in the previous chapter.

There are several aspects to productive fluency, which can be studied separately, or in combination. To categorize these aspects, we will use Skehan's (2003) division of aspects of fluency into three groups: speed, breakdown, and repair. Speed fluency refers to the rate of speech, breakdown fluency refers to the amount of silence in an utterance, the number of pauses, filled as well as unfilled, and repair fluency refers to the number of repetitions, false starts, self-corrections, etc.

One of the dangers of fluency measures, which several authors warn against, is the intercollinearity of measures (e.g. Witton-Davies, 2014). Different fluency measures can overlap or even measure the same aspect. For example, the measure of speech rate is related to the measure of silent pauses – large amount of silent pauses and/or their long duration cause the speech rate to be lower (de Jong, 2016). Therefore, if different measures are used in combination, it is necessary to be aware of their relations. De Jong (2016, p. 211) calls them "confounded" measures and she warns against using these measures especially in research aiming at establishing which aspects of speech are related to fluency ratings. Similarly, Bosker et al. (2013) suggest that for the sake of interpretability of the results, a combination of measures should be avoided.

2.2.1 Speed fluency

Following the example of Witton-Davies (2014), "rate of speech" will be used in a general sense to describe the speed of speech, while the term "speech rate" will be reserved for the particular fluency measure defined below. Three measures are most frequently used to quantify rate of speech: articulation rate (AR), speech rate (SR) and pruned speech rate (PSR). AR divides the number of words or syllables by the total phonation time (the time spent articulating those words/syllables), excluding silent pauses. This means it only takes into account the time when speech of any kind was being uttered. Witton-Davies (2014) quotes Goldman-Eisler (1968) saying, that AR is a high order skill and its measures are stable and therefore the variation in rate of speech is in fact caused by variation in pausing. However, AR can change within a longer period of time, it can be increased by practice.

SR or unpruned speech rate is a more general measure, used for example in Riggenbach (1991), acquired by dividing the number of words or syllables by phonation time and pause time. The resulting measure can be words/syllables per minute/second. The preference of particular units often depends on the field of study, Witton-Davies (2014) observes a tendency of psycholinguistic and pausological researchers to use syllables per minute (e.g. Derwing et al., 2004; Kormos and Dénes, 2004) while researchers in the field of ELT tend to prefer words per minute (e.g. Lennon, 1990; Riggenbach, 1991). As Gráf (2015) suggests, counting syllables per minute would provide more accurate results, but the calculations are rather time-consuming. De Jong (2016) also points out that counting syllables based on transcript is problematic, i.e. number of canonical syllables does not correspond to the number of syllables actually uttered by the speaker, as speakers (especially native speakers) have a tendency to reduce some syllables. This results in the count showing more syllables per minute than were produced by the speaker. Therefore, to obtain a precise count of syllables, the researcher would need to use a program such as PRAAT to analyse the sound properties of the utterance. Another problem some authors warn about (e.g. Gráf, 2015) are the missing definitions of word or syllable in some studies, and it is therefore up to the reader of such study to assume what the author meant. Similarly, it is not always clear whether filled pauses are included, or whether repetitions and repairs are counted, etc.

Pruned speech rate (PSR) is another measure encompassing multiple aspects. It was used e.g. in Lennon (1990) or Derwing et al. (2004). The method of calculating PSR is very similar to SR, the only difference is that in this case we count "pruned" syllables or words, i.e. all the words/syllables that remain after repair phenomena have been removed. That means the repetitions and reparanda are deleted, all the rest is used for the count, including reparata[1] . Witton-Davies (2014) expresses his surprise that PSR is not a more frequently used measure in the studies of fluency, as it can easily be calculated, is comprehensive and combines "the three main aspects of fluency – rate of speech, pause time and repair – making it the most global of fluency measures" (Witton-Davies, 2014, p. 72). De Jong (2016) corroborates this view, saying that if a researcher needs only one measure to encompass all aspects of fluency at the same time, PSR is the one to be used, although we then lose the ability to see what influence the subcomponents have. She also calls PSR "the king of confounded measures" (De Jong, 2016, p. 211), warning against using it in combination with other measures.

Another possible measure is pace, i.e. the number of stressed words per minute. In their study, Kormos and Dénes (2004) found pace to be a good predictor of fluency, which was a novel discovery, and they consider it relatively simple to calculate. Many studies use a combination

[1] By reparandum, we mean the part of an utterance that is changed, by reparatum, the part that replaces the reparandum and by repetition the part which is repeated (if a word or expression is pronounced twice, only the second instance is included in the calculation).

of measures, such as AR and SR (e.g. Kormos and Dénes, 2004; Towell et al., 1996), or SR and PSR (e.g. Lennon, 1990), others only include one measure. From the preceding paragraphs, we can see that PSR is the ideal measure if we wish to use only one measure encompassing as many aspects as possible, while if we prefer to distinguish between different aspects of fluency and combine more measures, AR is the most convenient option for speed fluency.

2.2.2 Breakdown fluency

The research in unfilled pauses is complicated in that pauses have multiple functions (Lennon, 1990), they can have a rhetorical function, be physiological (for breathing), or mark disfluency. Different kinds of pauses are present in every utterance and the researcher needs to decide which pauses to include in his/her analysis and which to ignore. Most researchers base the distinction on length of the pause, not counting pauses shorter than 0.2 seconds (e.g. Lennon, 1990), 0.25 seconds (e.g. Bosker et al., 2013), or even 0.4 seconds (e.g. Derwing et al., 2004). Some authors also exclude longer pauses, e.g. Riggenbach (1991) excludes all pauses above 3 seconds, as she does not consider them standard and does not think they should be included in measures such as speech rate. Another way of distinguishing between different kinds of pauses is based on their location – in some places the pauses sound more natural than in others, e.g. Chambers (1997) distinguishes between natural pauses (occurring at structural junctures) and unnatural pauses (occurring elsewhere, in the middle of semantic or structural units), being characteristic of fluent speakers and non-fluent speakers respectively.

To measure pausing, there are several options available to the researcher. One of them is pause-time ratio (used e.g. by Lennon, 1990), which measures what proportion of the overall speaking time is taken up by unfilled pauses. The inverse measure is phonation-time ratio, which is calculated as a proportion of total articulation time and total speaking time, but as Gráf (2015, p. 35-36) states, "[phonation/time ratio] provides a rather crude measure which is hard to interpret as it provides no indication as to the location and explanation of the pauses used" and this applies to both of these measures. In spite of that, several authors have used these measures (e.g. Kormos and Dénes, 2004; Towell et al., 1996).

It might be a better option to calculate pause frequency. There are several possible calculations, such as number of pauses per minute, number of pauses per number of words or syllables, e.g. pauses per 100 words, number of pauses per clause or per unit, and the number of words per pause. The last of the calculations gives the mean length of run (MLR), which is "the most common pause frequency measure" (Witton-Davies, 2014, p. 82), it is the amount of speech uttered between two pauses. However, Witton-Davies (2014) also points out that MLR is affected by length of turns, which may be problematic in measuring pause frequency in dialogues. Gráf (2015) adds that the measure is not reached simply, it needs to be identified clearly which runs will be included and length of pauses which mark the runs' boundaries needs to be specified. He suggests that number of pauses per 100 words, i.e. pause rate, might be a better indicator of fluency and easier to calculate. Another measure that can be used is the average length of pause (ALP), which is calculated as the total pause time divided by number of pauses. It was used e.g. by Kormos and Dénes (2004) or Towel et al. (1996).

Research on pauses is quite inconclusive, with different authors coming to different conclusions or acquiring data which can be interpreted in different ways. However, e.g. Kormos and Dénes (2004) showed relation between quantitative measures of pausing and fluency as assessed by raters, their findings showed that MLR, ALP and PTR were predictors of fluency. Riggenbach's (1991) theory about disfluency chunks might be one of the reasons for that – she claims that markers of disfluency, such as pauses (unfilled or filled), repetitions, etc. do not give the impression of disfluency if they stand alone, but if they are cumulated into groups, i.e. disfluency chunks, they give an impression of non-fluency.

The research in filled pauses is probably even more complicated than in unfilled pauses. Filled pauses are typical for native speakers as well as non-native, they have several functions, e.g. to signal a pause or a repair, or to show that the speaker has the intention to continue with his/her turn, however they can also function as dysfluency markers. Witton-Davies (2014, p. 92) claims that correlations between measures of filled pauses and fluency are rarely found, probably due to the variability between speakers and the necessity to analyse filled pauses in combination with other hesitation phenomena, such as unfilled pauses or repetitions. This lack of conclusive results led to some researchers not including filled pauses in their studies, e.g. Derwing et al. (2004), Towell et al. (1996). The authors who did include filled pauses in their research used various methods of measuring them. E.g. Lennon (1990) uses the ratio of total duration of silent pauses and total speaking time, and the number of filled pauses per T-unit and their location. Kormos and Dénes (2004) count filled pauses per minute and Götz (2013a) counts the number of filled pauses per hundred words. Some authors studied filled pauses in combination with other phenomena, Riggenbach (1991) studies "clusters of disfluencies", such as the combination of filled and silent pauses, which seem to have more significant influence on fluency than either type of pauses studied separately. Witton-Davies (2014) suggests that examining filled and silent pauses together while keeping a separate count of both is a sensible research option, he therefore argues for including filled pauses in pausing measures.

2.2.3 Repair fluency

The concept of repair fluency consists of several phenomena: self-corrections, false starts and repetitions. Self-corrections are the result of the speaker's monitoring his/her speech (e.g. Levelt, 1999) and finding it incorrect. In a self-correction, an utterance is interrupted and the erroneous part is uttered again, correctly. As Gráf (2015) states, a self-correction is only classified as such if it is the correction of an error. Otherwise (if it is not an error or cannot be determined) the term to be used is a reformulation or a false start.

A false start differs from a self-correction in that the original utterance is abandoned completely. In addition, a false start is not limited in reasons for the interruption. Self-corrections and false starts are rather similar in their nature and can be difficult to distinguish. A different kind of phenomenon are repetitions. Gráf (2015) draws upon Götz's (2007; 2013a) findings that L2 speakers tend to underuse repetitions, finds the opposite to be the case in his data and suggests that L2 speakers also have a different distribution of repeats, as they tend to use them more within clauses, which seems to correspond to L2 speakers' use of pauses. However, repeats are not considered markers of disfluency, but natural components of speech. Gráf (2015) even points out that by calling repair phenomena speech management strategies, we acknowledge their being highly natural and functional components of speech and Götz (2013a) suggests teaching these strategies to L2 learners, as she thinks they would help the learners' fluency. Therefore, it could be said that Götz rather considers them markers of fluency than disfluency. Witton-Davies (2014) also states that repairs are not indicators of lack of fluency, based on research by Freed (1995), who found L2 speakers who stayed in the target country to use repair phenomena more than those L2 speakers who did not participate in any such stay.

The measures of repair phenomena include for example Tavakoli and Skehan (2005) who measured the frequencies of repetitions, reformulations, false starts and substitutions (i.e. reformulations where only lexical items are changed). Their results show that in terms of proficiency, there is no difference in number of these phenomena, but there are differences in their character, e.g. less proficient speakers correct basic grammar while more proficient speakers correct style etc. Witton-Davies (2014) states that repair phenomena can be measured in combination or in isolation, but it is necessary to consider both frequency and extent. He suggests PSR (pruned speech rate) as the ideal measure, as it takes repairs into account.

2.3 Operationalization of perceived fluency

Perceived fluency has been studied by several authors, such as Kormos & Dénes (2004), Derwing et al. (2004), Bosker et al. (2013) or Préfontaine, Kormos, & Johnson (2016). In the following part, we will look at the studies of perceived fluency more closely, especially focusing on the methodology that has been used. An overview of the results will also be provided in this section, as the studies quite often aim at determining, which aspects of productive fluency are best predictors of perceived fluency.

One of the first authors to have studied perceived fluency is Riggenbach (1991). Her aim was to compare the speech of 6 speakers, 3 fluent and 3 non-fluent and to examine the differences. The fluent and non-fluent speakers were chosen by 12 ESL instructors who rated a number of recordings on a basis of a 7-point open-ended scale. The interrater reliability was not particularly high, which Riggenbach (1991) attributes to the use of open-ended scale and the possibility of different interpretations of fluency by raters – they were not given detailed information about fluency or guidelines for ratings. Even though the material to be studied was chosen on a basis of perception, the following microanalysis was based purely on measures and analyses of the utterances themselves, sometimes with regard to the raters' commentaries. The conclusion drawn from the commentaries is that raters considered other aspects of speech than just speed, pause phenomena and repair phenomena, such as grammatical structures and accuracy.

Lennon (1990) was the first of a number of researchers to study the relations between perceived fluency and productive fluency. The aim of his work was to establish, which aspects of productive fluency are related to perceived fluency in order to establish a way of assessing fluency without raters. He recorded four subjects with English as L2 before and after a stay in England. He had the recordings rated for fluency by 9 native-speaker teachers of EFL. The judges were provided with "a brief gloss on the term fluency as comprising: (1) a temporal element (speed of delivery, for example) and (2) a degree of freedom from various dysfluency markers (such as repetitions, self-corrections, filled pauses, and the like)." (Lennon, 1990, p. 403) In addition, 12 measures were taken to quantify the different aspects of fluency. However, the judges provided global ratings, without regard for the 12 different aspects of fluency.

Lennon (1990) suspected that the teachers may be influenced by more factors than just those provided in the gloss. The results show that improvements in perceived fluency are associated with reduction of filled pauses and repetitions, faster speech rate and reduction of pause time (increased MLR).

Derwing et al. (2004) also studied associations between productive and perceived fluency, but they were using untrained judges for the rating. Their aim was to determine, whether untrained judges' assessment corresponds to temporal measures of fluency and whether they stay consistent throughout different tasks. The material used were recordings Mandarin speakers speaking English (their L2) in three speaking tasks – a picture-based narrative, a monologue on a given topic and a dialogue in which the speaker was instructed to ask the researcher questions. From each recording, a sample was taken, 30 seconds from the picture story and monologue, 90 seconds from the dialogue, giving a total of 60 samples from 20 non-native speakers (40 samples of 30 seconds and 20 samples of 90 seconds). The raters were 28 native speakers of English, enrolled in an undergraduate ESL course at the University of Alberta, they had no prior experience with Mandarin speakers.

The listeners were told to listen for temporal variables, such as pauses, false starts and self-repetitions, they were informed that the researchers are interested in "fluency in terms of the flow and smoothness of speech rather than in terms of overall proficiency" (Derwing et al., 2004, p. 664). The pictures on which the storytelling was based and the topic of the monologue were provided in order to avoid the familiarity bias. The listeners assessed each recording on a numbered response sheet using a 9-point scale where 1 is extremely fluent and 9 is extremely disfluent. The authors state that they avoided Fulcher's (1996) descriptors, expecting them to overwhelm untrained listeners, as they were designed for trained raters. The listeners were also asked to rate comprehensibility and accentedness, both on a 9-point scale. The temporal measures taken were PSR (in syllables per second), MLR and silent pause frequency. The results show pruned speech rate and pause frequency to be a good predictor of raters' judgements. Derwing et al. (2004) also point out that more than just an interview should be used in proficiency exams, as fluency varies through different tasks.

Similarly, Zhang & Elder (2011) studied perceived fluency of Chinese speakers of English and had teachers (native and non-native) evaluate their speech. Unlike most authors, they did not compare perceived fluency with utterance fluency measure, their goal was to compare the rating of NS and NNS raters. They conclude that there are qualitative and quantitative differences between the ratings, which may have implications for the debate of native norm for language learners. For the actual ratings, the performance of students from CET-SET test was used, which provided ten 20-minute recordings with three candidates in each recording. The raters were provided with a scale from 1 to 5, the points being described as very poor, poor, good, very good, excellent. No further information on fluency rating was provided to the raters.

Kormos & Dénes (2004) also studied perceived fluency using teachers (but not trained raters) as judges. The aim, similarly to Derwing et al. (2004), was to establish, which variables predict the raters' perception of fluency and distinguish fluent learners from non-fluent. The temporal measures analysed 10 variables, from which SR (unpruned, measured in syllables per second), MLR, PTR and pace were found to be most influential. They also found accuracy to have impact on fluency judgements. Unlike Derwing et al. (2004) and several other researchers, Kormos and Dénes (2004) did not find breakdown phenomena (the number of filled and unfilled pauses) to have impact on fluency perceptions. Perceived fluency was rated by 6 judges – 3 native and 3 non-native speakers of English, they rated the recordings of 16 participants (with Hungarian as L1) on a 5-point semantic differential scale, where 1 was least fluent and 5 was most fluent. The raters were not provided with descriptors on the 5 categories in order to make intuitive judgements, however, they were asked for comments on the scores they gave each participant. The speech samples were 2-3 minutes long, which means they were longer than most of the samples used in other perception studies. From the raters' commentaries it seems that speed of delivery was important for all raters, hesitation phenomena were considered by several of them, but they varied in other aspects (e.g. in importance of lexical variety or accuracy). Interrater reliability was higher for non-native speaker assessors than for the native speakers.

A different approach can be observed in a study by Götz (2013b), in which she investigated fluency in the broad sense (i.e. overall oral proficiency) of German speakers of L2 English. She selected five "learner reference types" (Götz, 2013b, p. 1): the most accurate one, the least accurate one, one with very good temporal fluency, one with very poor temporal fluency and one with average performance in both aspects. The speakers were then judged by 50 native-speaker raters in order to assess the speakers' overall oral proficiency and six variables, which are central to perceptive fluency according to Götz (2013b): idiomaticity, register, lexical diversity, sentence structure, accent and pragmatic features. Temporal fluency score and errors (phw) were also measured. The raters, the majority of which were speakers of Australian English (20% were speakers of other varieties of English), were the staff and PhD students of

Macqurie University Sydney, they consisted of linguists as well as non-linguists, which made it possible to account for possible differences between NS and NNS judges' perception.

The raters were asked to listen to each interview once, then rate overall proficiency on a 10-point scale, where 1 "sounds like an absolute beginner" and 10 "sounds like a native speaker" (Götz, 2013b, p. 5), then they were asked to listen to the recording again and rate the six variables on the same 10-point scale. All the variables had been briefly explained in the questionnaire. Another difference from the aforementioned perceived fluency studies is that the rating process was performed in an online survey, in which the judges were able to listen to each recording as many times as they wished and they were also able to go back and change their ratings. The judges also had the option to comment on each learner and on the whole survey if they wished. The results showed that the variable with least impact on overall ratings is accuracy (the number of errors per hundred words), temporal fluency has a higher, but still insignificant correlation. From the six variables, only accent and pragmatic features have significant correlations. Götz (2013b) therefore concluded that above some proficiency level, accuracy no longer plays a role and other aspects, like accent or pragmatic features become more prominent.

Rossiter (2009) examined the ratings of expert NSs, novice NSs and NNSs of English. She studied how different the ratings of such judges are and how they correlate with objective measures of fluency. The material was a picture story narrated by 24 adult ESL learners at two points in time, from which 1-minute excerpts were taken. The judges were instructed to judge the excerpts for temporal fluency and were provided with a list of features commonly associated with the phenomenon: "speech rate, hesitation phenomena (e.g., unfilled or non-lexical filled pauses, repetitions, self-corrections), and formulaic sequences or 'chunks.'" (Rossiter, 2009, p. 401). They were first instructed to write their general impressions and then rate the recording on a 9-point scale where 1 is extremely dysfluent and 9 is very fluent. The recordings were presented to the judges in pairs with the instruction to assign a different number to each recording. The results showed that the ratings were all inter-correlated and that they correlated with measures of pause per second and pruned syllables per second. The results also showed non-temporal features such as pronunciation, grammar and vocabulary to have influence on perception of fluency.

Bosker et al. (2013) performed four experiments to investigate the impact of three fluency aspects (i.e. pauses, speed and repair) on perceived fluency of L2 Dutch speakers. In the first experiment, untrained raters assessed oral fluency of learners of Dutch, analyses were then performed which showed that pause and speed measures were the best predictors of perceived (i.e. subjective) fluency ratings. The three following experiments used a new set of untrained raters to assess the same recordings for the use of pauses, speed and repairs respectively. The total of 80 raters, all Dutch native speakers without training in language rating, participated in the study. The recordings which were rated, were of a group of 15 L1 English speakers, 15 L1 Turkish speakers and 8 Dutch native speakers who functioned as a reference point for the raters to compare the non-native speakers to. The speakers performed a wide variety of speaking tasks, from which three were selected for the experiments. From each task, a sample was chosen to be rated. Therefore, the material counted 114 items of approximately 20 seconds recorded from 38 speakers. Each sample started at a phrase boundary and ended in a pause.

The acoustic measures calculated for each recording were mean length of syllables, number of silent pauses, number of filled pauses, mean length of silent pauses, number of repetitions and number of corrections. The raters were instructed not to rate the items based on the broad definition of fluency, but rather on use of pauses, speed of delivery and hesitations and corrections, but nor grammar, for example. Six practise items were provided for the raters before the beginning of the experiment. The scale used was a 9-point Equal Appearing Interval Scale, where the extremes were "not fluent at all" and "very fluent" (Bosker et al., 2013, p. 166). The results show that the complex rating model best predicted fluency, and that raters were sensitive to all three aspects. Repair fluency was the weakest predictor of fluency.

Another study of perceived fluency in Dutch is Cucchiarini, Strik, & Boves (2002), who examined the relations between objective properties of speech and perceived fluency in spontaneous and read speech. They concluded that speakers are more fluent in read speech and that raters base their ratings on different properties for different kinds of speech. The recordings were rated by multiple groups of experts (phoneticians and speech therapists) in the first experiment and 10 ESL teachers in the second experiment. The speakers were non-native speakers of Dutch of various levels and the material consisted of two 5-sentence sets read out loud (ca. 1 minute of speech per speaker) in the first experiment and answers from a language proficiency test in the second experiment. The evaluation consisted of a 10-point scale and the set of 5 sentences was evaluated as a whole, no specific information on fluency assessment was provided. In the second experiment, the raters gave each participant a score as in the test and then fluency score on the same 10-point scale as in experiment 1.

Another paper which examined perceived fluency in a different language than English is Préfontaine et al. (2016), which compared perceived and utterance fluency in L2 French. The study followed a similar structure to most of the previous papers – 11 untrained raters evaluated the recordings of 40 learners of French. To calculate utterance fluency, four measures were taken – mean length of run, articulation rate, frequency of pauses and length of pauses. The results showed MLR and AR to be the most influential factors. A novel finding was that length of pauses was positively related to fluency scores, i.e. longer pauses were assigned to more fluent learners, which is the opposite of the findings in English.

The raters were all French language instructors, as they are used to evaluating learners' speech and their results are expected to be more consistent. No training was given to avoid influencing the raters with the authors' interpretations of fluency. The study tried to imitate real-life or testing contexts, therefore the raters were asked to assess the whole recording (three speaking tasks), as they would assess in an exam situation. The raters were asked to evaluate the recordings on a 6-point scale based on the CEFR (Common European Framework of Reference) with each point consisting of a can-do statement and corresponding to a CEFR level (A1-C1). Another assessment on the Fluency Perception Semantic Scale (designed for the study specifically) consisted of rating pauses and speed, corresponding to breakdown fluency and speed fluency. The raters first listened to the whole recording, giving their overall impressions of fluency, and then to each task separately, with an interval of several days/weeks. The results of the study support results of most previous studies – MLR being one of the stronger predictors, together with AR and average pause time. Surprisingly, pause frequency was found to be the weakest predictor of pause behaviour ratings in two of the three tasks.

Another paper on perceived fluency in French, which brings forward the qualitative perspective, is Préfontaine & Kormos (2016). They had 30 adult learners of French record 3 speech tasks, which were assessed by 3 native speaker teachers of French with no previous experience as fluency raters. The raters were not given information on fluency or its assessment and were asked for justifications of their fluency ratings. The main features that influenced the raters' perception of fluency were "speed, rhythm, pause phenomena, self-correction, efficiency/effortlessness in word choice and target-like rhythm and prosody." (Préfontaine & Kormos, 2016, p. 151) What differentiates this research from others is the conclusion that rhythm plays an important role in fluency ratings in syllable-timed languages.

Another language in which perceived fluency has been examined is German, in the study by Dressler and O’Brien (2017). They used 48 speech samples, each 20 seconds in length, produced by native and non-native German speakers. What makes this study unique is that the samples were rated not only by native and non-native speakers of German, but also by non-speakers of German. The authors also suspected a difference between the terms fluency and fluidity, which is why half of the judges in each group were told to evaluate fluency and the other half fluidity (with all the other information provided identical), the term in both cases was defined as “how smoothly and rapidly an utterance is spoken” (Isaacs and Trofimovich, 2012 in Dressler & O’Brien, 2017, pp. 7–8). The survey was performed online, the raters were given the instruction to place themselves in a quiet room and complete the experiment without help of others. They completed a background questionnaire, underwent a practise rating session and then rated the samples. The results show that raters in all groups were able to distinguish native from non-native speakers in their ratings, there were no significant differences between the ratings for fluency and fluidity, but the measures on which the raters relied were different, leading to the authors’ suggestion that “fluidity” may be a more fitting term to use for the perceived fluency scale. Native speakers relied more on narrow definition of fluency, non-native raters took into account other aspects, such as grammatical correctness.

2.3.1 Summary

Based on the overview of previous empirical research on perceived fluency, we can see a shift from smaller number of expert raters to larger numbers of novice, NNS raters or even raters who do not speak the language at all. The research varies in the language fluency is studied in, number of speakers that are evaluated and number of raters. It can be observed that the number of speakers and raters tends to rise in the more recent studies (although there are exceptions, e.g. Préfontaine & Kormos, 2016 only have 3 raters assess the recordings). Another variable is the length of the samples, the shorter samples being from 20 seconds (Dressler and O’Brien, 2017) to one minute (e.g. Rossiter, 2009), and longer samples from 2-3 minutes (e.g. Kormos and Dénes, 2004) to 20-minute recordings (Zhang & Elder, 2011), which included three speakers, making the average 6.3 minutes per speaker. The researchers choosing longer recordings generally aimed at conditions typical of proficiency examination, where longer stretches of speech are rated.

Another varying factor is the amount of information provided to the raters – the most common procedures are the following two. Either the researcher wants to avoid influencing the raters’ idea of fluency and gives no information at all (and in that case, raters are usually asked for commentary) or a definition is given, even a list of features to listen for or a list of features to ignore (in that case, commentary is not always asked for). The situation of the evaluation also varied, some researchers were present to the listening, making sure conditions were the same for everyone, others did an internet survey. Some judges were allowed a limited number of listenings, others could listen as many times as they wished and some could even go back and change their ratings. The larger numbers of participants in the recent studies can be explained by the availability of the Internet for the evaluations. However, it can be seen from Préfontaine & Kormos (2016) that even qualitative research can yield interesting results.

3. Material and method

3.1 Material

The material used for the research consists of two parts. The first part of the data are samples taken from the recordings, which come from the Czech subsection of LINDSEI corpus (Gilquin et al., 2010), LINDSEI_CZ (Gráf, 2017). The second part of the data are the perceived fluency ratings of these samples by native speakers of English.

3.1.1 Data from LINDSEI_CZ

The LINDSEI corpus is a database of recordings of non-native speakers of English, the speakers' profiles and transcriptions of the recordings. For the thesis, the recordings of Czech speakers of English were used. The speakers recorded for the corpus were all students of the English and American Studies bachelor programme in the third or further year of their studies. This choice ensures the proficiency of the speakers – they are all advanced speakers of English (CEFR level B2 and higher, based on Huang et al., 2018). Each recording consists of three parts: a monologue on a set topic, a free interview and a picture description. For the present research, only a sample of the first part – a monologue on a set topic was chosen. The recordings were cut and some of them modified in Audacity® recording and editing software, version 2.3.0.

For the monologue, the speakers were given a choice of three topics and time to decide and prepare. The possible topics were (Gráf, 2017):

1) Important life experience
2) Important film or play
3) Important travelling experience

The first set of data used for the analysis are the speech rates calculated from the monologue part of the recording, more precisely from the exact part which was used as a sample in the evaluation task. The second set of data comes from two evaluation tasks.

3.1.2 Data from evaluation tasks

The material for the evaluation consists of two sets of samples. The first set is a pilot study aiming to establish inter-rater and intra-rater reliability. (For more information see section 3.2.3. Reliability of judges.) In the first set, there are ten 60-second samples from 5 speakers. For each speaker, at least one of the samples has a modified pitch so that the listener would not recognize the two recordings as coming from the same speaker. For the same reason, the parts of the monologues were chosen which could not be easily connected based on their content. The choice of recordings for evaluation included several factors. In the first phase, based on transcriptions, the recordings where the interviewer had to pose questions even in the monologue part were excluded. In the second phase, the recordings were ordered based on speech rate. In the third phase, five speakers along the speech rate continuum were chosen (one with the lowest speech rate, one with the highest, three in between with approximately equal differences among them). In this phase, some speakers were excluded from the first set of recordings based on the topic, namely because of the specificity of the topic, which might indicate to the listener that the two recordings come from one story and therefore from the same speaker. Another possible criterion would be a choice of prominent features, such as a high frequency of filled/unfilled pauses, pronunciation, complexity, accuracy of speech, to see which of the criteria would influence the listeners and which would not. However, this criterion was omitted, as it was more important that the listeners would not recognize that each speaker occurs twice in the set and the distribution of speakers based on speech rates still ensures a variety of features occurring in the samples.

For the second set of recordings, 25 samples were chosen in order to cover a wide range of speech rates. Each sample was approximately 90 seconds in duration. For this set, recordings were excluded for three reasons. First, the same as in the first phase described above, the recordings which were not actually monologues. Second, the recordings which had been used in the first set. Third, some recordings were excluded because of the quality of sound – the volume was considerably different and/or there was an echo or background noise, which might influence the listeners' evaluation.

These two sets of recordings were evaluated by native speakers of English on a 7-point scale, for the first set of recordings, the listeners were also asked to comment on the evaluation process, to describe which features of speech caught their attention and influenced their rating. Consequently, the material from the evaluation tasks consists of two sets of numeric evaluations for the quantitative analysis and one set of written commentaries for the qualitative analysis.

3.2. Method

3.2.1 Speakers and the speaking task

As the speakers and the speaking task are both part of the already compiled LINDSEI_CZ corpus, information about them have been provided in section 3.1 Material above.

The number of speakers chosen for evaluation is partly based on the fact, that with the LINDSEI_CZ corpus and its 50 speakers at hand, it is considerably less complicated to have a small number of raters evaluate a larger number of recording samples than vice versa. However, to evaluate all 50 speakers would mean either very short samples or evaluation process too lengthy for a person to concentrate throughout. It was also convenient to exclude some recordings for the aforementioned reasons. The length of the samples was chosen with the same aim. It is essential for the length of the sample to provide enough material for the listener to decide, while restricting the duration of the whole evaluation process so that a listener would be able to concentrate the whole time and would be willing to undertake such evaluation process.

Moreover, the monologue part of the recording has various lengths, with the shortest recording lasting less than 3 minutes. For the first listening task, it was necessary to extract two samples, which could not be recognized as belonging to the same monologue, which resulted in 1 minute being the ideal length of the sample. For the second listening task, there was no such restriction, which means the length was only influenced by the total duration of the task, which resulted in 90 seconds being chosen as the ideal length.

The choice of the monologue task as the material for the listening task was mainly due to practical reasons. Although several authors (e.g. Segalowitz, 2010) have proved that different speaking tasks have influence on speech fluency and some works compare fluency in different tasks, including the interview and the picture description in the evaluation would mean extensive length of the listening task and complicated analysis, as in an interview, the interviewer can influence the pace, and it can be complicated to decide where turns begin. The picture description task, on the other hand, has the lowest speech rate of the three tasks (for the majority of the speakers) (Gráf, 2015, pp. 131–132) and might sound unnatural to the raters. In addition, this task would not provide enough listening material on its own, as it is the shortest for most speakers, lasting less than one minute in some cases. For these reasons, choosing the monologue task was the most practical option.

3.2.2 Listeners

There are five native-speaker listeners, all of them volunteers, without linguistic education, but with some experience in teaching English as a second/foreign language. Two listeners have teaching experience shorter than one year, two have 1-3 years of experience and one has more than three years. All five listeners are native speakers of American English, all have the USA as country of origin. Two of them are women, three are men, their ages ranging from 22 to 45.

Even though recent studies mostly use a higher number of listeners, this study is based on only five raters' evaluations. One of the reasons is the number of samples evaluated (e.g. Götz (2013) has a large number of raters, but only 5 speakers being rated). With the total duration of the two sets of samples being 47.5 minutes, each of the two evaluation processes takes approximately 40 minutes depending on the thoroughness and speed of the rater. With an evaluation process of such length, most raters would demand a reward. For the same reason, non-professional raters were chosen, with regard to the fact that previous research found that the differences between professional and non-professional raters were insignificant. Another reason for this number of listeners is the qualitative analysis. With higher number of commentaries for each sample, the qualitative analysis would become extremely time-consuming, which seems to be unnecessary, as Préfontaine & Kormos (2016) yield interesting results with only three raters.

3.2.3 Reliability of judges

To account for the reliability of the judges, two separate listening tasks were designed. The second task is aimed at receiving the data for the main analysis of correlations between speech rate and perceived fluency, but before such analysis can be performed, it is necessary to establish how reliable the raters are. That is one of the aims of the pilot study, the other being to give insight into the process of evaluation, as the listeners are asked for commentary on this process, especially for features on which they based their decisions. To make sure of the judges' reliability, there are 10 samples of recordings in the first listening task, presented as 10 speakers, even though they actually come from 5 speakers. To determine intra-rater reliability, it is evaluated to which extent and in what number of cases each rater evaluates the two samples coming from the same speaker in the same way. To determine inter-rater agreement, we calculate to what extent the raters evaluated the same sample in the same way. To calculate inter-rater agreement, the Agreement Calculator from the Lancaster Stats Tools online (Brezina, 2018) was used. This test is not only performed on the pilot study data, but also on the data from the second listening task.

Based on the pilot study, raters 3 and 4 are the most internally reliable. Out of the five speakers, they evaluated four consistently, and in one case they differed by one figure. Raters 1 and 5 are less reliable, having evaluated two speakers consistently and differing by 1 figure in the case of three speakers. The least internally reliable rater is rater 2, who evaluated two speakers consistently, with another two he/she differed by 1 and in the case of one speaker, the evaluations differed by 2 figures. In general, we can consider the raters reasonably consistent, especially the two, who differed the least in their evaluations. It is necessary to take into consideration that the samples come from the same speaker and the same monologue but are not identical. They differ in terms of speech rates, and particular features. This means that a rater can be consistent in his/her evaluations and still differ in the particular figure. However, such difference should probably be restricted to 1 figure. It is quite unlikely that the same speaker within one monologue can objectively differ by 2 points on a 7-point scale, especially if the other raters evaluated this speaker within a 1-point difference and with regard to the fact that most raters only used 4-5 points of the scale.

The inter-rater reliability calculations show that in the pilot study, a considerably bigger agreement between the raters occurred than in the main evaluation task. According to the agreement calculator (Brezina, 2018), in the pilot research, the raw agreement was 94.67 % ($p < 0.001$), while in the main research, the raw agreement was only 84.44 % ($p < 0.001$).

3.2.4 Listening task preparation

For the first listening task, the chosen recordings were cut and modified using the Audacity (version 2.3.0) program. The resulting samples were uploaded onto Google Drive and inserted into Google Forms, which was used for the evaluation. The privacy of the form and the recordings was set so that only a person with a link can access the recording and fill in the form. Each form begins with a set of instructions for the listener and a number of personal questions, such as the country of origin, length of teaching experience, verification of English being their native language and a question about the variety of English they speak. There is a voluntary field asking for the first name, which means the listeners can remain anonymous if they wish to. In the actual evaluation part, there are 10 sections, named Speaker 1-10, and there is a link to the recording in each section, which can be played as many times as the listener needs. The form also enables the listeners to go back and forth, which means they could change their evaluation if needed. However, the movement is limited, it can only happen in the given order of the sections, one section forward or one section back at a time. The listener can never view two sections at the same time.

The samples are not ordered randomly, as it is not desirable for two samples from the same speaker to follow each other. Therefore, a semi-random ordering was performed manually, making sure that at least two other recordings are placed in between two samples from the same speaker and that the samples are not following the same pattern (i.e. the speakers do not follow each other twice in the same order).

Each sample is accompanied by a 7-point scale where 1 is extremely disfluent and 7 is extremely fluent. No tags are provided for the points in between. In each section, the listener is asked to evaluate the recording on the scale and fill in a field with a commentary. The choice of the scale was based on previous research, which varied from a 5-point scale to 10-point scale (see section 2.3 Operationalization of perceived fluency for more information). As Isaacs & Thomson (2013) mention in their research on rating scales, with a 9-point scale, the raters tend to have trouble differentiating between steps in the middle of the scale. However, as the scale is used to rate 10 samples in the first evaluation and 25 samples in the second, a 5-point scale offers very little space for differentiating between 30 speakers in total. Therefore, a 7-point scale, which was used e.g. by Riggenbach (1991), seems like a reasonable compromise between the demands put on the raters and the range available to them.

The second listening task has a similar structure to the first one, the very first section contains a similar set of instructions (modified to the needs of the task) and identical set of personal questions. Following this introductory part are 25 sections named Speaker 1-25, each including a link to the recording and the same 7-point scale as in the first listening task.

3.2.5 Procedure/Task

The listeners receive a link with the listening task via e-mail, they are free to choose the time and place which is the most convenient for them. Although it would provide more control over the quality of sound and equality of conditions to organize an evaluation session, it would be more time consuming for the raters and therefore considerably more complicated to organize. Furthermore, with the listening task performed over the internet, the raters are able to take a break if they have trouble concentrating, they can perform the task at their own pace without time pressure or stress and it can be done anywhere – even in a different city or state. To raise the probability of the rating procedure being performed in good conditions, the raters are asked in the instructions to make sure they have a good sound quality, ideally headphones, and calm environment before they begin rating. For the instructions as were provided to the raters, see the Appendix.

The link with the listening task has been sent to 10 Czech-based native speakers, 5 of whom agreed to participate. As a part of the initial set of instructions, the raters were given quite limited information on fluency. Authors vary on the information they provide to the raters. Some give no information at all, as they wish to avoid influencing the raters with their own views of fluency (e.g. Cucchiarini et al., 2002), others give definitions of fluency and a list of features to look for (e.g. Bosker et al., 2013). To be able to understand how the evaluation process works or have control over it, it is necessary either to give enough guidelines, or to ask the raters how they proceeded. As we ask the raters for commentary on the evaluation process, we do not need to provide the raters with too much data. What was made clear in the instructions was that fluency should not be regarded as overall speaking proficiency, that we regard fluency as one of the components of such proficiency. They were asked not to base their decision on grammar or number or mistakes, for example, but to base their evaluation on temporal features, pauses, repetitions, etc.

The raters were further instructed to listen to each sample as many times as they need and to go back and change their evaluation if they later feel they misjudged a sample. They were also given the approximate length of the whole task and throughout the task, the form provided them with information about how far in the process they are (based on the number of sections completed/remaining).

3.2.6 Data analysis

The data collected in the procedures described above is analysed using mixed methods. Qualitative method is used to analyse the commentaries from the first listening task, quantitative for the analysis of the points on the scale. Although recent research shows preference for the quantitative method, Préfontaine & Kormos (2016) argue for qualitative method not being omitted, as it can bring forward aspects of fluency and fluency evaluation, which quantitative methods do not find or do not show as prominent. As some authors disagree on how precisely the raters follow instructions - whether they follow exactly the instructions they are given or let themselves evaluate fluency based on their view of fluency and subconsciously ignore the instructions, it seems best to combine an approach where the raters are asked about the rating process and an approach where the numeric results they provide are analysed.

Using the qualitative method, we look at the features mentioned by the raters in their commentaries, especially at the frequency – which features occurred in the commentaries the most frequently or which were typical for a particular rater. Then we look at the commentaries of each rater separately, concentrating on his/her evaluation – the number of features the rater concentrated on, which occurred repeatedly and which were used uniquely for a particular speaker, etc.

In the quantitative analysis, we try to find correlations between the speech rate and the rating from the listeners, with the aim to show whether speed of speech has an influence on listener's perception of fluency and to what extent. For the quantitative analysis, the Pearson correlation test is used.

4. Research questions

The present study aims to examine the relation between perceived fluency as rated by non-expert native speaker judges and productive fluency in the form of one of the objective temporal measures, namely speech rate. Even though this relation has already been examined by several researchers, their results are not entirely consistent. The results of most studies show speech rate to be a strong predictor of perceived fluency (e.g. Kormos & Dénes, 2004; Derwing et al., 2004), but Cucchiarini et al. (2002) found speech rate to be a strong predictor for beginners' perceived fluency, but MLR was found to be a stronger predictor for intermediate students. Even more surprisingly, Götz (2013b) finds temporal fluency to have an insignificant correlation with perceived fluency. The first research question therefore tries to determine the extent of the relation between perceived fluency and speech rate.

RQ1: To what extent does perceived fluency correlate with speech rate in the speech of Czech L2 learners of English?

Most research uses only or mostly quantitative methods (e.g. Cucchiarini et al., 2002; Bosker et al., 2013 etc.), studying a wide range of variables, trying to establish which of the objective measures are stronger predictors of perceived fluency and which are insignificant. However, some researchers, such as Préfontaine & Kormos (2016) argue for qualitative method of research, as it can yield interesting results which would not be reached through quantitative method. As perceived fluency is based on the judgments of the listener, it seems logical not to restrict the material to numeric assessments, but to ask the listeners for more detailed input, to ask what they based their decisions on, which aspects of speech they concentrated on or which features in the particular utterance they noticed. The second research question concentrates on this aspect of research.

RQ2: What can the raters' commentaries tell us about the rating process and the perception of fluency in the speech of Czech L2 learners of English?

5. Results and analysis

In this chapter the results of the two listening tasks and their analyses will be shown. In the first part, the results of the qualitative analysis will be presented based on the raters' commentaries, in the second part, the results of the quantitative analysis of the evaluation will be presented in order to establish, whether there are correlations between the speech rate measurements and raters' evaluations.

5.1 Qualitative analysis

The aim of this section is to get a better understanding of the process of evaluation and the features of speech that the raters took into account when evaluating the speech samples. It is necessary to point out that the raters differ in how detailed their commentaries are, from the most concise commentaries consisting of two to three words, to commentaries consisting of approximately 50 words. Similarly, the total number of features mentioned in all evaluations by one rater differs considerably, with the rater giving the lowest number of concepts mentioning seven features in total, and the rater giving the highest number of concepts mentioning 21 features.

5.1.1 The commentaries

The features most raters mentioned include pauses, pronunciation, filled pauses or fillers, flow or fluidity (mentioned by 4 of 5 raters), speed, stresses, nativeness and accent (mentioned by 3 of 5 raters). Most raters also mentioned some sort of distraction - either the accent, or the pronunciation of a particular word, or misuse of a word cause the listener to get distracted from the contents of the utterance, or on the other hand, e.g. with the accent, they evaluated it as being detectable but not distractive. Similarly, the effort on the part of the listener was evaluated several times, comprehensibility, amount of concentration needed, etc. For the complete commentaries see the Appendix.

As for the pauses, the listeners evaluated their number as well as placement. Pronunciation was mentioned as a concept, but more frequently the pronunciation of a certain sounds was judged as incorrect. Some raters concentrated more on pronunciation in general, mentioning it quite frequently, others only mentioned it in relation to a particular speaker. Similarly, some raters concentrate more on filled pauses and mention them with several speakers, others only mention them with one or two speakers, who seem to be using them in a distracting way or amount.

However, only one of the raters commented on fillers in general, others pointed out a particular filler which they noticed in the sample. The flow or fluidity of speech was also noticed by most raters, commenting on its naturalness or interruptions. In one of the commentaries, it was differentiated between natural speed and natural flow, in another between flow and rhythm. Unfortunately, it is not clear whether the raters consider flow and fluidity to be equivalent terms, but it seems reasonable to suppose so.

In terms of speed, the raters usually commented on the speaker being slower than a native speaker or simply slow, quite frequently the evaluation of speed goes together with a commentary on hesitation or pauses. In the evaluation of rater R1, the difference between two speakers evaluated as 6 and 7 was in the commentary on speed, pauses and accent. Stress was mentioned when irregular or unnatural. Some speakers mention irregular or unnatural stresses, only rater R5 specified that he/she is commenting on word stresses. He/she also mentioned unnatural emphasis, by which sentence stress was probably meant. Native-like speech was mentioned by three raters, two of whom were describing the same speaker, one referred to the speaker himself/herself, the other referred to the accent. The third rater described the use of pauses as common in American speaking, which we understand as meaning native-like. The concept of nativeness was used uniquely in positive meaning, it seems that to express the negative meaning, the raters use the term natural, e.g. *the stresses were not so natural sounding*. However, several other features were described as being or sounding natural. Three raters noticed, whether the accent was present but negligible, noticeable, native sounding or consistent. Each commented on the accent of several speakers, meaning that accent was one of the most prominent features, commented on frequently.

Other features occurring less frequently, but still recurring, include ease of formulating thoughts, choppiness or interruptions, rhythm and word choice. In the evaluation of rater R1, the ease of formulating thoughts was used in connection with speaker evaluated as extremely fluent, stating that they seemed to have no problem formulating ideas, or on the other hand with a less fluent speaker (evaluated as 5), where short pauses for formulation were noticed. Similarly, rater R4 suspected that a number of filled pauses produced by a less fluent speaker (evaluated as 4) may be due to thinking of a word or trying to phrase something. Two raters described samples as choppy, R4 used the term three times, R5 once. Out of the four occurrences in total, three were used to describe the samples of the same speaker. Both speakers described as choppy belong to the part of speakers evaluated as less fluent (their ratings range between 2-5 in one case and 3-5 in the other). Although only two raters evaluated the samples as choppy, other speakers commented on the interruptions, by which the same quality of speech is meant probably. Rhythm was only mentioned twice, both times in an evaluation of a speaker belonging to the more fluent part, in the first case, the evaluation of rhythm is positive, in the other, the rater comments on the rhythm being strange, but still gives the sample the second best rating (6). Word choice was commented on multiple times, in a positive as well as negative way. It was used in general (e.g. *awkward word choice*), or in reference to a particular word being used (e.g. *the speaker used "like" the same way a native would*).

Some features or descriptions were only used by one of the raters, these are examined in the following part, in which we look at the evaluations of each rater separately.

5.1.2 The raters

The evaluations of rater R1 were quite detailed, the minimum of three features were given for each sample. The rater mainly concentrated on speed/pace, pauses (filled and unfilled), he/she also commented on native-like sound or natural sounding of the sample or a particular feature. For one speaker, he/she had the feeling that the speaker is thinking in L1 and quickly translating into English. Interestingly, almost the same commentary was given for both samples from this speaker. Furthermore, although the speaker gave the rater the impression of translating from L1, they did not receive the lowest evaluation. The evaluations given by rater R1 ranged from 3 to 7, showing that the rater did not consider any of the speakers extremely disfluent.

Rater R2 had a tendency to concentrate on pronunciation, accent, speed and grammar. The minimum of features given for a sample were two, and the features repeated more than with the other raters. Pronunciation, for example, was mentioned in every evaluation. This rater seems to have followed a slightly different procedure than the other raters – based on the repetition of features, it seems that he/she followed an internal list of features to concentrate on. He/she seems less flexible than the other raters, who often only mention a feature once, presumably with a sample where this feature was prominent. This does not happen with rater R2. In addition, this rater supports the theory that raters cannot be trusted to follow instructions, as he/she evaluated grammar several times, even though it was specifically stated in the instructions not to base the decision on grammar or number of mistakes. It seems that the instructions were overshadowed by the rater's internal view of fluency. The evaluations range from 4 to 7, leaving out three lowest points.

Rater R3 concentrated on use of pauses (filled and unfilled), flow or fluidity of speech, his/her evaluations were rather concise, with the minimum of two features given. He/she quite frequently mentions a feature, but does not state whether it is positive or negative, or to what extent it is used (e.g. the commentary *meter and lilt of speech, pronunciation* could be seen as positive as well as negative, we are therefore obliged to infer from the numeric evaluation what was meant). He/she notices the filler "um" several times, comments on word choice, repetition. In one evaluation, "few to no mistakes" were mentioned. Unfortunately, we do not know whether grammatical mistakes were meant, or self-corrections, or another kind of mistakes. As the whole evaluation states "speech is fluid, with few to no mistakes", the two previously mentioned kinds are the most plausible. The evaluations range from 3 to 7.

Rater R4 used the widest variety of features of all the raters (approximately twice as many as any other rater), his/her commentaries were the most detailed. The minimum of features mentioned is 3, however, most evaluations contained considerably more. He/she concentrated on speed or pacing, pauses (filled and unfilled), and natural or native-like sound of speech. He/she tends to mention many features just with one particular speaker, such as simpler language and repetition of the same words, shortening of words which resulted in the sample sounding choppy/staccato, etc. He/she also comments on the speaker sounding calm or stressed, and the natural sound of their inflections. For the least fluent speaker (according to the evaluations), he/she also points out the need to focus on what was being said in order to understand, which was not the case with the other samples. The evaluations range from 4 to 7, similarly to rater R2.

Rater R5 offered probably the least information, giving almost no information in some cases, such as *extremely fluent* or *overall not very fluent*, giving rather a commentary on the result of the evaluation than the process. He/she concentrated mainly on pronunciation features (such as word stress, linking) and pauses (mainly filled). He/she was the only rater to concentrate on lack of linking and one of the two who commented on one of the speakers sounding nervous. He/she also noticed the unnatural usage of *well* as a filler in one of the samples. Interestingly, the range of evaluations is the widest for this rater, ranging from 2 to 7.

5.1.3 Summary

It is interesting that none of the features of speech was mentioned by all of the raters. Although in some cases, it was not clear whether to categorize a description as a feature in itself or a part of a more general category, e.g. pronunciation and aspects of pronunciation such as word stress, linking, etc., it seems that each rater omitted at least one of the features mentioned by the others. Together with the fact that some raters wrote a thorough commentary, while others barely mentioned any features and some raters concentrated on grammar, which was given as an example of what not to base their decision on in the instructions, it shows that we should not view the raters as a mass. They are individuals, noticing different aspects of speech, using different procedures to evaluate it, differing in range of evaluations, thoroughness of the commentary, and the extent to which they succeed at following the instructions.

However, we can see quite clearly that temporal features played role in the evaluation process, the most frequently mentioned features include speed, filled pauses and unfilled pauses, flow/fluidity of speech. For most raters, it was also important whether the speaker sounds native-like or whether the speech sounds natural. Pronunciation was equally an important aspect for most raters. In addition, a high number of features only occurred in the ratings of one rater or occurred very infrequently, some features were also only mentioned with relation to a certain speaker. This supports the idea that the evaluation process differs not only with different raters, but also with different speakers/speech samples.

5.2 Quantitative analysis

Utterance fluency, represented by speech rate was calculated as the number of words uttered in the sample divided by the length of the sample in seconds times 60, giving the number of words per minute. The results (ranging from 120 to 191 WPM, mean 149 WPM, SD = 20) can be seen in Figure 1.

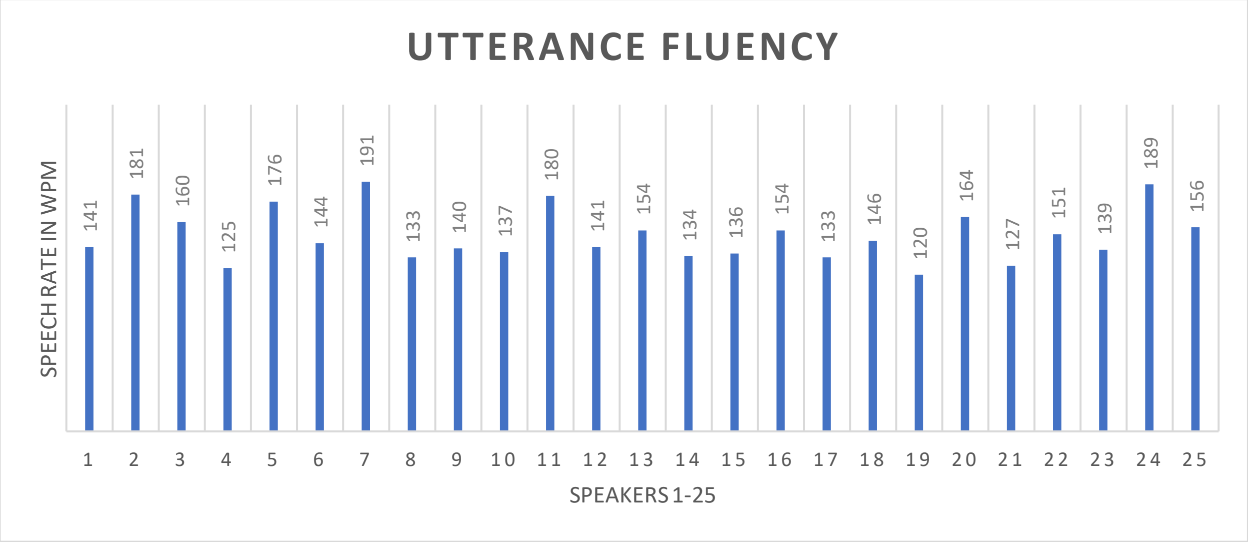
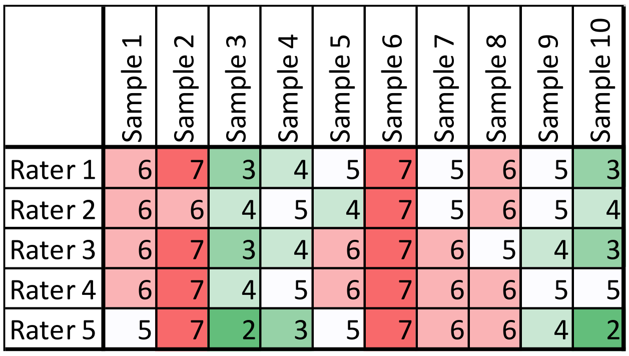


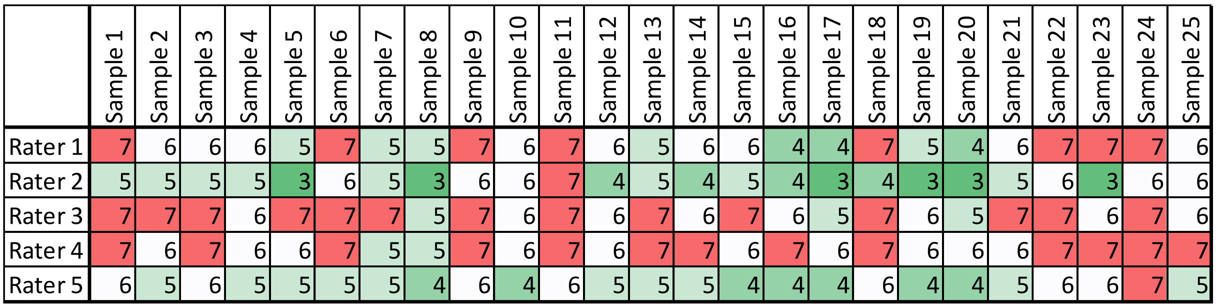
Figure 1: Speech rates in WPM

Perceived fluency, acquired from five raters who evaluated the samples on a scale from 1 to 7, can be seen in Figure 2 for the pilot study and Figure 3 for the main study.



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
Rater 1	6	7	3	4	5	7	5	6	5	3
Rater 2	6	6	4	5	4	7	5	6	5	4
Rater 3	6	7	3	4	6	7	6	5	4	3
Rater 4	6	7	4	5	6	7	6	6	5	5
Rater 5	5	7	2	3	5	7	6	6	4	2

Figure 2: Evaluations by raters on a scale from 1 to 7 (in the pilot research)



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12	Sample 13	Sample 14	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21	Sample 22	Sample 23	Sample 24	Sample 25
Rater 1	7	6	6	6	5	7	5	5	7	6	7	6	5	6	6	4	4	7	5	4	6	7	7	7	6
Rater 2	5	5	5	5	3	6	5	3	6	6	7	4	5	4	5	4	3	4	3	3	5	6	3	6	6
Rater 3	7	7	7	6	7	7	7	5	7	6	7	6	7	6	7	6	5	7	6	5	7	7	6	7	6
Rater 4	7	6	7	6	6	7	5	5	7	6	7	6	7	7	6	7	6	7	6	6	6	7	7	7	7
Rater 5	6	5	6	5	5	5	5	4	6	4	6	5	5	5	4	4	4	6	4	4	5	6	6	7	5

Figure 3: Evaluations by raters on a scale from 1 to 7 (in the main research)

As can be seen in Figures 2 and 3, the raters used a wider range of evaluations in the pilot study than in the main study. While in the pilot study, 3 was used by three raters, in the main study, it was only used by one rater (interestingly by the rater whose lowest evaluation in the pilot study was 4). In the pilot study, a higher inter-rater agreement can be found than in the main study. There is one sample which all raters evaluate with the same value and several samples, where the raters differed by one point. In the main study, on the other hand, the samples where the raters only differ by one point are quite rare, while the samples where the raters differed by 2 or 3 points are rather frequent and there is even one sample where the raters differed by 4 points. In the main stuudy, there is also a visible tendency of some raters to use consistently higher evaluations (5-7), e.g. raters R3 and R4, while other raters have a tendency to use consistently lower evaluations (4-6 for R5 and even 3-6 for R2) and the highest point is used only once in their evaluations. Rater R1 has relatively even distribution of the points from 4 to 7 (however, evaluation 6 and 7 are used more than 4 and 5).

The relationship between productive fluency (operationalized as speech rate measured in words per minute) and perceived fluency (operationalized as an evaluation on a scale from 1 to 7) was measured using Pearson product-moment correlation coefficient. Preliminary analysis was performed to measure intra-rater and inter-rater reliability, finding that raw inter-rater reliability was relatively low (84.44 %). This was supported by the fact that a correlation test for mode value of the 5 raters showed no correlation. Furthermore, when we looked at the correlations between speech rates and evaluations by raters separately for each rater, significant correlations were found only for raters R3 and R5. More precisely, for rater R1, correlation was not found, with Pearson's $r = 0.009$ and $p > 0.05$, for rater R2 there is a weak correlation with $r = 0.293$ and $p = 0.078$, for rater R3 there is medium correlation with $r = 0.392$ and $p < 0.05$ (therefore the result is statistically significant), for rater R4 there is no correlation with $r = 0.042$ and $p > 0.05$ and for rater 5 there is medium correlation with $r = 0.378$ and $p < 0.05$ (statistically significant).

To sum up, a medium correlation was found for two out of five raters with the result statistically significant, a weak correlation was found for one rater, but the result was not significant, and for two raters no correlation was found and the result was not significant. The possible reasons for such results and implications will be considered in the following chapter.

6. Discussion

The results show that utterance fluency (represented by speech rate) and perceived fluency (represented by raters' evaluations) correlate to a limited extent in the case of a part of the raters. This means that there clearly is a relationship between the two variables, however it is not very strong and it does not apply to all raters. One possible reason for such a result can be the fact that only some raters pay attention to speech rate or temporal features in general. To examine this theory, we cross-referenced the results of the Pearson correlation test with the raters' commentaries on fluency. Rater R1, whose evaluations did not correlate with speech rate, mentioned speed of speech, the fact that the speaker speaks fast or slow, several times. R2 (weak correlation) mentioned speed of speech in 4 out of 10 commentaries. R3 (medium correlation) evaluated fluidity, pauses, but did not mention speed in particular. R4 (no correlation) mentioned speed several times, however it was one of many features he/she concentrated on. R5 (medium correlation) did not mention speed, only pauses and flow of speech.

The cross-referencing of correlations and commentaries on evaluations seems to be giving contradictory results. However, it needs to be taken into account that the two raters who showed medium correlation, were unfortunately also the raters who gave a rather limited commentary, while the two raters who did not show correlation gave the most extensive commentaries, mentioning a wide variety of features, which may mean that other features were given more prominence in their evaluation process. Moreover, it is necessary to point out that the commentaries were given in the pilot research, which means they comment on the process of evaluation of certain samples, while the correlations were calculated based on the evaluations from the main research. Although the commentaries provide insight on the raters' process of evaluation in general, it does not provide information about the process of evaluation of the particular samples which were examined for correlations with speech rate.

The results are especially interesting in comparison with previous research, which shows speech rate as a strong predictor of perceived fluency, e.g. Kormos & Dénes (2004), Derwing et al. (2004). A possible cause for such discrepancy is in differing methodology. The studies in perceived fluency vary considerably in terms of number of raters, number of speakers/samples, the proficiency of speakers, etc. It is therefore possible that the results differ because of such differences. It is possible that speech rate is a good predictor of perceived fluency at lower proficiency levels, but it is not as influential in the evaluation of advanced speakers. To examine such an explanation, we look at the number and proficiency level of speakers, whose speech samples were used in the studies in which speech rate was shown to be a strong predictor of perceived fluency. In Kormos & Dénes (2004), the number of speakers evaluated is 16, namely 8 advanced learners and 8 low-intermediate learners. In Dressler & O'Brien (2017), 48 speakers in total were evaluated, 24 native speakers and 24 non-native, out of whom 12 speakers were described as B1 and 12 as C1 or C2 level. In Préfontaine, Kormos, & Johnson (2016), 40 speakers were evaluated, their proficiency levels varied from beginner to intermediate and advanced (the proportion is not given). In Cucchiarini, Strik, & Boves (2002), 30 beginner speakers and 30 intermediate speakers are evaluated, however, they conclude that speech rate (as well as other objectively measurable features) are stronger predictors of read speech than spontaneous speech. In Derwing et al. (2004), 32 beginner speakers are evaluated.

Having taken a closer look at the proficiency level of speakers used as independent variables in fluency studies, we can see that at least two different proficiency levels occurred in most of these studies. In the study where speakers of a single proficiency level participated, the level was beginner. We can therefore consider plausible the explanation that speech rate is a more prominent predictor of perceived fluency at beginner level or if more proficiency levels are compared, but it is not such a prominent predictor if only advanced speakers are evaluated.

Another factor possibly contributing to the lack of correlations between speech rate and raters' evaluations is the fact that the evaluation of larger numbers of speakers of similar proficiency can be very demanding for the raters, resulting in their evaluations being imprecise. Not only does the whole evaluation process take at least 40 minutes, but most of the speakers are at a similar proficiency level, and in some cases the differences in speech rates are very little. This is also supported by the fact, that the raters' reliability is much higher in the pilot research, where they were asked to evaluate fewer samples and give commentary, which means they needed to concentrate more on the task, while in the main research, there were 25 samples, each sample was 30 seconds longer and no commentary was required. Moreover, the samples in the pilot study were chosen with regard to speech rate, so that there would be regular intervals between the speakers. As a result, the speech rates of the samples in the pilot research ranged from 105 WPM to 217 WPM, giving a difference of 112 WPM, distributed among 10 samples and only 5 speakers. In the main research, on the other hand, the speech rates only ranged from 120 WPM to 191 WPM, giving a difference of 71 WPM per 25 samples. It is therefore possible that with such limited variation for such an extensive number of samples, raters have to rely on other features than speech rate in their evaluations. From this point of view, Götz's (2013) method, where she chooses a restricted number of prototypical speakers and has them evaluated by a large number of raters seems to be beneficial. However, by choosing the speakers with the prototypical features, we would create an ideal situation for the rater, which they would not encounter if, for example, evaluating fluency in a language testing environment.

The result clearly show that the perception of fluency is a highly subjective matter. Perceived fluency is very difficult to operationalize, which leads (together with differences in methodology) to complications in comparing different studies on perceived fluency. From the commentaries acquired in the pilot research we can see that although all raters were given the same instructions, they concentrated on different features, they differed in how closely they followed the instructions or to what extent their personal views on fluency influenced their judgement. The quantitative research then yields similar results in that each rater evaluates the recordings differently, some of the evaluations correlate with speech rate to some extent, others do not correlate at all, the same sample can be evaluated as 3 by one rater and 6 by another, which shows that the raters probably did not base their decisions on the same features, their evaluation process was not the same. As even researchers studying the topic of fluency vary in their definitions, it is no surprise that raters differ in what they understand by fluency and in their perceptions of fluency.

6.1 Implications for teaching

The present study has confirmed that different raters evaluate fluency differently, varying in the numerical evaluations as well as the description of the features which influence them. This shows fluency to be a complicated, multidimensional concept, which is consistent with the results of previous research, which differ depending on the dimension the researchers concentrate on. Perceived fluency is a highly subjective phenomenon, difficult to operationalize. The study has shown that different raters differ in their rating procedure, in the features they concentrate on, in the extent to which their rating correlates with speech rate. If we look at these results with regard to learner language and ELT, it is clear that this phenomenon should not be used as a component of language tests and proficiency evaluations. As can be seen in the evaluations of the raters, one speaker can be perceived as extremely fluent by one rater and more disfluent than fluent by another. Although it needs to be taken into account that the raters in the present study are not trained examiners, they are ELT teachers with some experience in evaluating learner language, if not in official examinations. Therefore, it is very likely that equal subjectivity and equal differences in evaluations occur in language testing environment and learners' fluency is not evaluated objectively. That is why it cannot be recommended to evaluate fluency in language tests until researchers come to a better understanding of how to measure fluency and identify a method which can be used for objective measuring of fluency in language testing and examiners are trained to use that method. The results suggest that any reports on fluency by examiners who have not received thorough and specific training are bound to be subjective and the same performance may be rated differently by different examiners.

6.2 Limitations and further research

A phenomenon as complex as fluency is difficult to examine and in studying fluency, we come across several challenges. First, to be able to compare productive fluency and perceived fluency, a number of different measures would need to be applied. By choosing one aspect (speech rate) as a representation of productive fluency, we omit several others, such as MLR, pause frequency, frequency of repeats and corrections, etc. and therefore we examine only one component of the phenomenon. Another possible limitation lies in having chosen WPM for measuring speech rate, which is the most common unit, best understandable for teachers, but it is problematic in that a speaker who uses more complex, longer words will have lower speech rate than a speaker who uses simpler, shorter words. In such case, a measure in syllables per minute would prove more precise. However, in the situation where all speakers have a given choice of topic and are therefore likely to speak with a similar level of complexity, the advantages of WPM seem to outweigh the limitations. Another challenge lies in the methodology, by choosing to enable the raters to evaluate the recordings over the internet, we lose control over the evaluation process – we cannot control how much attention the raters pay to the rating, how much time they spend doing it, whether they listen to the whole sample, what quality of sound they have and whether they make sure nothing distracts them during the rating process. It is possible that some raters listen to each recording several times while others only listen to it once, some pay more attention than others. Another limitation is the extent of the study. With 5 raters providing their evaluations, limited conclusions can be drawn, for further research, it would be beneficial to have more raters evaluate the recordings and comment on the process of evaluation so that more information can be gathered and patterns can be observed among the raters.

Further research will be needed to identify the other variables which influence the perceptions of fluency, operationalize them and measure them.

7. Conclusion

The aim of the present study was to determine whether and to what extent perceived fluency as represented by raters' evaluations correlates with productive fluency as represented by speech rate and to examine the process of evaluation based on commentaries provided by the raters. To establish the relationship between productive and perceived fluency, 5 speakers were chosen for the pilot study and 25 for the main research, and for each sample the speech rate was calculated and compared with the evaluation given by each of five native-speaker raters who evaluated the samples on a 7-point scale. The aim of the pilot study was two-fold, first to establish inter-rater and intra-rater reliability and second to provide insight into the process of evaluation as perceived by the raters themselves. The aim of the main research was to establish to which extent the evaluations of 25 Czech advanced speakers of English correlate with speech rate calculated in WPM.

The goal was partly achieved. The study shows that raters' perception of fluency correlates with speech rate to a limited extent and only in the case of a part of the raters. The commentaries show that the raters concentrate on different features when evaluating fluency – no feature was mentioned by all five raters – the features named the most frequently include pauses (filled and unfilled), pronunciation, flow or fluidity, less frequently also speed, word stress, native-likeness and accent. From these results, it can be seen that the relationship between productive fluency and perceived fluency is not as clear as it seems from the previous studies, most of which conclude that speech rate is one of the strongest predictors of perceived fluency. From the Pearson correlation test as well as from the raters' commentaries it is clear that in the case of advanced Czech learners of English, there are other factors which influence the raters more, such as the frequency and distribution of pauses, pronunciation, and flow of speech. To be able to acquire reliable measures of perceived fluency, it is therefore necessary to find all the features which influence the listeners, operationalize them and measure them. Thus, rather than with Kormos & Dénes (2004) or Derwing et al. (2004) the study is more in line with Götz's results (2013) and her conclusion that speech rate is not one of the most influential predictors of perceived fluency, and supports her theory about fluencemes of perception – features that influence the listeners' perceptions of fluency.

The results obtained in the present study show clearly that perceived fluency is a complex phenomenon which is difficult to study not only because there is a wide variety of features to consider, but also because it is highly subjective. Different listeners may take into account different features and thus differ in their evaluations. Besides, as fluency is not a concept with one commonly known definition, they also differ in their views on fluency, which may (possibly subconsciously) influence their evaluations.

8.Bibliography and sources

Audacity Team (2019). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 2.3.2 retrieved Jan. 22nd 2019 from https://audacityteam.org/

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535–544. https://doi.org/10.1016/S0346-251X(97)00046-8

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873. https://doi.org/10.1121/1.1471894

de Jong, N. H. (2016). Fluency in second language assessment. *Handbook of Second Language Assessment*, (March), 203–218. https://doi.org/10.1207/s15327752jpa8502

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Dressler, A. M., & O'Brien, M. G. (2017). Rethinking perceptions of fluency. *Applied Linguistics Review*, 1–22.

Fillmore, C. J. (1979). On fluency. In D. Kempler, and W. S. Y. Wang (Eds.), *Individual differences in language ability and language behavior*, 85-102. New York: Academic Press.

Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), Second language acquisition in a study abroad context (pp. 123–148). Philadelphia: John Benjamins

Fulcher, G. 1996. Does thick description lead to smart tests? A data-based approach to rating-scale development. *Language Testing* 13(2): 208-238.

Gilquin, G., De Cock, S. & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Goldman-Eisler, F. (1968). *Psycholinguistics experiments in spontaneous speech*. London; New York: Academic Press.

Götz, S. (2013a). *Fluency in Native and Nonnative English Speech*. *International Journal of Learner Corpus Research* (Vol. 53). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/ijlcr.1.1.08gra

Götz, S. (2013b). How fluent are advanced German learners of English (perceived to be)? Corpus findings vs. native-speaker perception. In *Huber, M., & Mukherjee, J. (Eds.). (2013). Corpus Linguistics and Variation in English: Focus on non-native Englishes. Helsinki: University of Helsinki Varieng Electronic Series.*

Gráf, T. (2015). Accuracy and fluency in the speech of the advanced learner of English.

Gráf, T. (2017). LINDSEI_CZ: korpus spontánní mluvené angličtiny pokročilých mluvčích. Ústav Českého národního korpusu FF UK, Praha 2017. Available from: http://www.korpus.cz

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia*, *8*(1), 14. https://doi.org/10.1186/s40468-018-0069-0

Isaacs, T., & Thomson, R. I. (2013). Rater Experience , Rating Scale Length , and Judgments of L2 Pronunciation : Revisiting Research Conventions Rater Experience , Rating Scale Length , and Judgments of L2 Pronunciation : Revisiting Research Conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: identifying the linguistic influences on listeners' L2 comprehensibility ratings. Studies in Second Language Acquisition, 34(3), 475-505. DOI: 10.1017/S0272263112000150

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach*. *Language Learning*, *40*(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levelt, W. J. M. (1999). Producing spoken language: a blueprint of the speaker. In *The neurocognition of language* (pp. 83–122). Oxford Press.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *IRAL - International Review of Applied Linguistics in Language Teaching*, *54*(2), 151–169. https://doi.org/10.1515/iral-2016-9995

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73. https://doi.org/10.1177/0265532215579530

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441. https://doi.org/10.1080/01638539109544795

Rossiter, M. J. (2009). Perceptions of L2 Fluency by Native and Non-native Speakers of English. *Canadian Modern Language Review*, *65*(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Schmidt, R. (1983). Interaction, acculturation and the acquisition of communicative competence. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition*, 137–174. Rowley, Mass: Newbury House.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. *Cognitive Bases of Second Language Fluency*. New York: Routledge. https://doi.org/10.4324/9780203851357

Skehan, P. (2003). Task-based instruction. Language Teaching, 36(01), 1–14. doi:10.1017/S026144480200188X

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), Planning and task performance in a second language (pp. 239–273). John Benjamins.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. Applied Linguistics, 17(1), 84–119. https://doi:10.1093/applin/17.1.84

Witton-Davies, G. (2014). The study of fluency and its development in monologue and dialogue. *Unpublished Doctoral Dissertation). University of …*. Retrieved from http://www.forex.ntu.edu.tw/en/files/writing/4092_dc0088cd.pdf

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31–50. https://doi.org/10.1177/0265532209360671

9. Résumé

Předkládaná práce se zabývá tématem plynulosti řeči v souvislosti s žákovským jazykem, konkrétněji porovnáním dvou typů plynulosti – vnímané a verbální – u pokročilých českých mluvčích angličtiny. Práce porovnává verbální plynulost zastoupenou tempem řeči vyjádřeným ve slovech za minutu (WPM) a vnímanou plynulost zastoupenou hodnotou na stupnici od 1 do 7, získanou od pěti rodilých mluvčích, kteří na základě nahrávek hodnotili plynulost zkoumaných vzorků tak, jak ji subjektivně vnímají. Práce vychází z předpokladu, že jednotlivé aspekty verbální plynulosti, které jsou přesně měřitelné, mohou v různé míře předurčovat hodnoty vnímané plynulosti. Jedním z aspektů, který podle předchozího výzkumu, např. Kormos a Dénes (2004), Derwing et. al. (2004) nejsilněji předurčuje vnímanou plynulost, je právě tempo řeči. Jiní autoři, např. Götz (2013), naopak dochází k závěru, že vnímaná plynulost je ve větší míře ovlivňována jinými aspekty, jako jsou akcent či pragmatické prvky. Předkládaná práce tedy zkoumá, do jaké míry je vnímaná plynulost předurčována právě tempem řeči.

Teoretická část zasazuje pojem plynulosti řeči do kontextu žákovského jazyka, poskytuje přehled nejvýznamnějších autorů, kteří se plynulostí zabývali a jejich pojetí plynulosti – to, jak ji definovali, na jaké typy ji dělili, z jakého úhlu na ni nahlíželi (např. zda studovali plynulost u rodilých mluvčích či v žákovském jazyce) apod. V této části se ukazuje, že nejčastější dělení plynulosti je na tři typy, a to kognitivní, verbální a vnímaná. Ačkoli někteří autoři používají pro jednotlivé typy jiné názvosloví, na rozdělení jako takovém se shoduje většina autorů. Následně teoretická část také prezentuje možné způsoby operacionalizace dvou typů plynulosti, a to verbální a vnímané. Uvádí způsoby, jakými plynulost měřili autoři v předchozích výzkumech, a představuje studie, které se zabývají vztahem vnímané a verbální plynulosti. U verbální plynulosti práce představuje aspekty měření verbální plynulosti, tedy přerušení řeči, opravy a rychlost. V případě prvních dvou lze měřit frekvenci výskytu jevů, jejich rozložení v promluvě, či jejich absolutní výskyt, rychlost je měřena jako tempo mluvy, lišit se může jednotkami či množstvím aspektů řeči, které jsou do tempa započítávány. V případě vnímané plynulosti poskytuje práce přehled metod, které byly v předchozích výzkumech použity k měření vnímané plynulosti (velikost vzorku mluvčích či hodnotitelů, délka vzorku, šíře škály, na níž byly vzorky hodnoceny apod.).

Následující metodologická část popisuje, jaký materiál a metodologie byly použity k vypracování analýzy. Nejprve představuje mluvčí, tedy korpus mluveného žákovského jazyka LINDSEI, z něhož byl čerpán materiál pro vzorky, jejichž plynulost je v práci zkoumána. V rámci představení korpusu jsou uvedeny základní informace o nahraných mluvčích, struktura nahrávek, témata, z nichž si mluvčí vybírali apod. Z korpusu nahrávek bylo vyňato celkem 35 vzorků od 30 mluvčích, 10 vzorků o délce jedné minuty pro pilotní výzkum a 25 devadesátisekundových pro hlavní výzkum. Všechny vzorky byly vybrány z monologické části. Druhá část materiálu byla získána na základě těchto vzorků, a to ohodnocením plynulosti vzorků pěti rodilými mluvčími angličtiny, pocházejícími z USA, z nichž všichni měli praxi v učitelství angličtiny jako cizího jazyka. Druhý segment metodologické části sestává z představení těchto hodnotitelů. Je zde uvedeno věkové rozpětí hodnotitelů, délka praxe v učitelství angličtiny a poměr mužů a žen.

Dále je popsána metodologie, která byla použita k získání hodnocení plynulosti, tedy jak byly sestavovány dotazníky, v nichž měli hodnotitelé vzorky poslouchat a hodnotit, jaké informace byly hodnotitelům poskytnuty, jaké instrukce jim byly předány, za jakých podmínek hodnocení probíhalo, zdůvodnění použití 7stupňové škály apod. Dále bylo vysvětleno, jakým způsobem bylo voleno pořadí jednotlivých vzorků a zdůvodněno médium, a to porovnáním výhod a nevýhod provedení výzkumu za fyzické přítomnosti hodnotitelů a společného poslechu a hodnocení nahrávek v předem daném časovém úseku a provedení výzkumu pomocí internetového formuláře, kdy může každý z hodnotitelů věnovat úkolu tolik času, kolik považuje za vhodné a může úkol provádět v čase, který mu vyhovuje nejlépe.

Metodologická část dále vysvětluje použití pilotního výzkumu, v němž bylo vybráno 5 mluvčích, od každého mluvčího dva vzorky, z nichž alespoň jeden byl upraven tak, aby nebylo možno určit, že se jedná o téhož mluvčího, což umožňuje určit, do jaké míry jsou jednotliví hodnotitelé ve svých hodnoceních konzistentní. V pilotním výzkumu byli dále hodnotitelé požádáni o slovní komentář, tedy o vysvětlení, na základě kterých prvků vzorky řeči hodnotili, co je v kterém vzorku nejvíce zaujalo či ovlivnilo. Tento komentář poskytuje materiál pro kvalitativní část analýzy. Hlavní výzkum pak poskytuje materiál pro kvantitativní analýzu, tedy pro určení, zda existuje korelace mezi verbální a vnímanou plynulostí. Tato část dále popisuje měření tempa mluvy, které bylo měřeno pro všechny použité vzorky, poskytuje informace o použitém výpočtu i zvolené jednotce.

Kromě vnitřní shody každého z hodnotitelů se metodologická část zabývá také shodou mezi hodnotiteli. Míra shody jak interní v rámci jednoho hodnotitele, tak mezi hodnotiteli, určuje, jakou váhu lze přikládat výsledkům získaným v následné analýze. Samotná analýza je v této části rovněž popsána. K analýze získaných dat je použita jak kvalitativní, tak kvantitativní metoda. Kvalitativní analýza je založená na komentářích hodnotitelů a zabývá se zejména jednotlivými rysy řeči, které jednotliví hodnotitelé uvedli, poskytuje jak přehled rysů podle toho, jak často byly hodnotiteli zmíněny, tak podrobnější pohled na komentáře jednotlivých hodnotitelů a jejich specifika. Kvantitativní analýza určuje pomocí Pearsonova korelačního koeficientu, zda existuje korelace mezi verbální plynulostí, reprezentovanou tempem mluvy, a vnímanou plynulostí, reprezentovanou hodnotami na stupnici, poskytnutými pěti hodnotiteli.

Následující část představuje výsledky analýzy. Nejprve jsou představeny výsledky kvalitativní analýzy – jsou zde popsány nejčastější rysy řeči, tedy takové, které zaujaly nejvíce hodnotitelů. Tato část ukazuje, že žádný rys není zmíněn všemi hodnotiteli, avšak některé rysy zjevně převažovaly na jinými. Z této analýzy je také vidět, že jednotliví hodnotitelé přistupovali k hodnocení samotnému i komentáři různě. Komentáře se liší jak v délce, tak v celkovém počtu uvedených rysů a míře opakování určitého aspektu. Z této části analýzy jasně vyplývá, že na hodnotitele je nutno nahlížet jako na jednotlivce, jejich komentáře jasně ukazují, že proces hodnocení plynulosti nelze generalizovat. Kvantitativní analýza se zabývá numerickými hodnotami, ukazuje, že korelace mezi tempem mluvy a vnímanou plynulostí existuje pouze u dvou hodnotitelů a jedná se pouze o střední korelaci, u zbylých tří hodnotitelů je korelace slabá nebo žádná. Tento výsledek potvrzuje, že jednotliví hodnotitelé se liší v tom, na co se při hodnocení zaměřují. Výsledek také ukazuje, že ačkoli u daných vzorků a hodnotitelů existuje korelace mezi vnímanou plynulostí a tempem řeči, jiné charakteristiky jednoznačně vnímanou plynulost také ovlivňují, a to velmi pravděpodobně do větší míry, zejména u těch mluvčích, tempo řeči s hodnocením nekorelovalo.

Další část nazvaná diskuze se pokouší interpretovat výsledky analýzy. V této části je nejprve nahlíženo na výsledky Pearsonova korelačního testu s ohledem na komentáře hodnotitelů, ty však neukazují jasnou souvislost mezi korelacemi a tím, zda hodnotitel zmiňuje ve svém komentáři temp řeči či nikoli. Dále jsou výsledky interpretovány s ohledem na předchozí výzkum, zejména na úroveň pokročilosti mluvčích, jejichž nahrávky byly použity jako vzorek. Následně je zohledněna komplikovanost hodnocení, vzhledem k srovnatelné úrovni mluvčích. Tato část dále pojednává o implikacích výsledků pro výuku angličtiny, to zejména pro testy jazykové úrovně, které často testují plynulost na základě vnímané plynulosti hodnotitele, která, jak se ukázalo v předkládané práci, je velmi subjektivní a může se lišit od objektivních měření plynulosti. Následně jsou popsána omezení práce a návrhy na další výzkum. Mezi omezeními je uvedeno zejména to, že práce zkoumá pouze korelace mezi tempem mluvy a vnímanou plynulostí, pro plné pochopení vnímané plynulosti je však třeba zkoumat i další aspekty verbální plynulosti na různých vzorcích a za pomoci různých hodnotitelů. Dále jsou popsány některé limity metodologie, např. nedostatek kontroly nad hodnotiteli během hodnocení, limitovaný počet hodnotitelů apod.

Poslední část, závěr, shrnuje, co bylo cílem práce, tedy určit, jaký je vztah mezi verbální plynulostí vyjádřenou tempem mluvy a vnímanou plynulostí vyjádřenou hodnocením rodilých mluvčích, konkrétněji míra korelace mezi nimi. Dále je zde připomenuto, jakým způsobem byla získána data k analýze, jak byla analýza vypracována a její výsledky. Následně je zhodnoceno, do jaké míry bylo cíle dosaženo a co z práce vyplývá jak s ohledem na předchozí výzkum, tak pro výzkum budoucí. Finálně je připomenuta komplikovanost konceptu plynulosti, subjektivnost jejího hodnocení a subjektivnost toho, co který mluvčí vnímá pod pojmem samotným.

10. Appendix

The appendix contains the instructions provided to the raters in each of the rating tasks, one example of a question from the pilot study and the raters' commentaries.

1. Instructions for the pilot study survey

Fluency judgements with commentaries

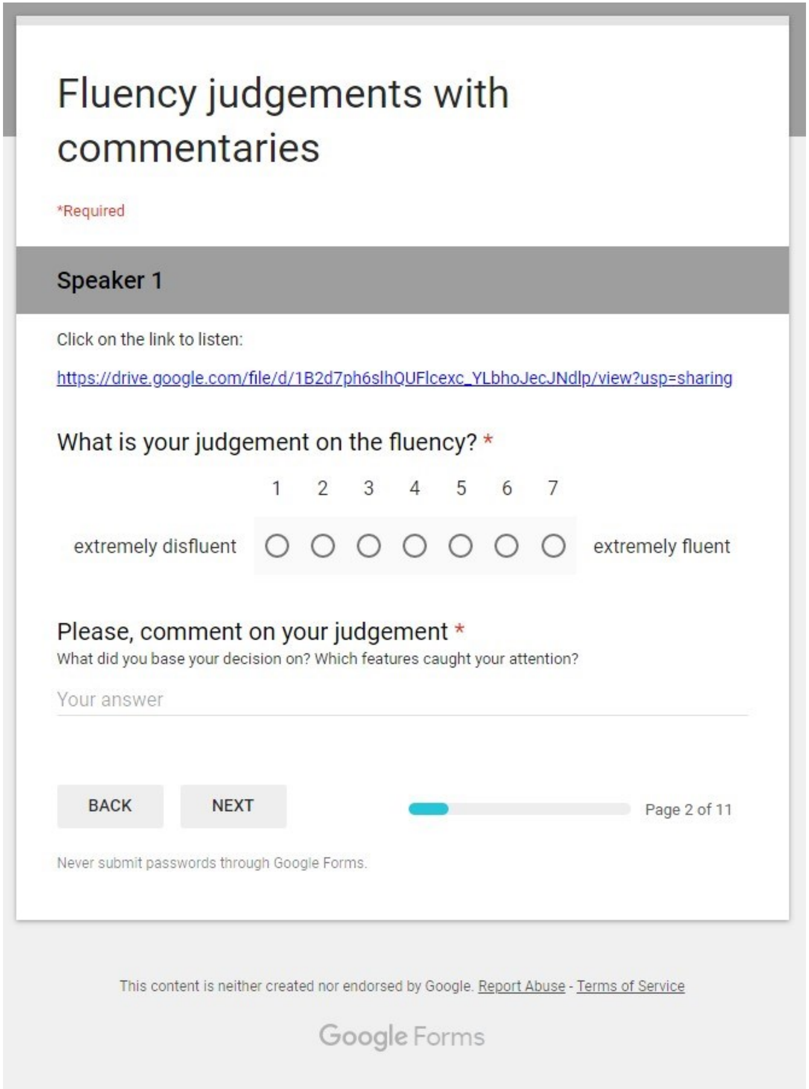
Please, before you start, read the following text carefully.

You will hear 10 recordings (1 minute each), for each recording, you will be asked for judgement of fluency on a scale from 1-7 and a commentary on the process of judging, the features that influenced your decision, etc. By "fluency" we do not mean overall speaking proficiency, fluency is regarded as one component of such overall proficiency, as opposed to complexity and accuracy, you shouldn't base your decision on grammar or number of mistakes, but on temporal features, pauses, repetitions, etc.

Please make sure you have good sound quality, ideally headphones, and calm environment before you begin.

To listen to the recording, you will have to click on the link provided in the question. You can listen as many times as needed, and you can also go back if you feel you misjudged a recording. The whole procedure should take about 30 minutes.

2. An example of a question in the survey



Fluency judgements with commentaries

*Required

Speaker 1

Click on the link to listen:

https://drive.google.com/file/d/1B2d7ph6slhQUFlcexc_YLbhoJecJNdlp/view?usp=sharing

What is your judgement on the fluency? *

	1	2	3	4	5	6	7	
extremely disfluent	○	○	○	○	○	○	○	extremely fluent

Please, comment on your judgement *

What did you base your decision on? Which features caught your attention?

Your answer

BACK NEXT Page 2 of 11

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service

Google Forms

3. Instructions for the main study survey

Fluency judgements plain

Please, before you start, read the following text carefully.

You will hear 25 recordings (90 seconds each), for each recording, you will be asked for judgement of fluency on a scale from 1-7.

By "fluency" we do not mean overall speaking proficiency, fluency is regarded as one component of such overall proficiency, as opposed to complexity and accuracy, you shouldn't base your decision on grammar or number of mistakes, but on temporal features, pauses, repetitions, etc.
Please make sure you have good sound quality, ideally headphones, and calm environment before you begin.

To listen to the recording, you will have to click on the link provided in the question. You can listen as many times as needed, and you can also go back if you feel you misjudged a recording. The whole procedure should take about 45 minutes.

4. The raters' commentaries

Recording 1 - CZ001	Recording 2 - CZ012	Recording 3 - CZ041	Recording 4 - CZ031	Recording 5 - CZ030
commentary	commentary	commentary	commentary	commentary
The flow of the talk was good, there seemed to be no problem formulating ideas in English but the speech was slower than a natural speaker. I could definitely hear an accent but it didn't distract from what was being said. Pauses were made in appropriate areas for thought formulation, not searching for words.	Spoke in at a natural if almost quick pace, with very few pauses in response. The accent was noticeable, but negligible.	This speaker's accent was noticeable. Also, slow speaking with many pauses to think of how to say the words. May "Uhs" in places that aren't natural for native speakers.	There pace was fairly natural but with pauses of Uh and um. Seemed to be thinking in L1 but translating fairly quickly to English	The person seems to be speaking at a normal pace though there are some short pauses for formulation.
This person seemed extremely fluent and if not for a few mispronunciations of certain words that were not a result of a British accent as far as I am aware I understood them perfectly.	The person seemed to be speaking Australian English with a slight accent, the "r"s were much more pronounced than a typical drawl.	This speaker seemed to have difficulty with "th," "d," and "t," sounds. They also spoke more slowly than a native or fluent speaker.	Excellent grammar but trouble with the "th," sound pronunciation - uses "d," instead.	Good grammar use, irregular stresses on syllables and struggle with some letter sounds (th).
The use of "um" and the pauses are very common in American speaking	Usage of "um", rhythm and flow of speech, slight pauses to catch breath all are characteristics of someone who has knowledge of the subject being discussed	Flow of speech is interrupted and not fluid	Solid flow of speech with mild interruptions for "um"	Speech is fluid with few to no mistakes
I think the pauses, stresses, and inflections sounded quite natural. The speech was a little bit slow.	The pacing and accent sounded native, good phrases were used	a lot of pauses and "um/uh" I think non-native english speakers sometimes do this when they are thinking of a word or how to phrase something. the language used was a little bit more simple in this recording and was repeated a few times	Inflections of words in the sentences and speed of speech sounded natural but it didn't seem to have natural flow. The speaker shortened words a lot when they were speaking (it sounded a little bit choppy/staccato)	it seemed to come naturally to the speaker, maybe a little bit slow. some pronunciation and stresses were not so natural sounding, but it was a well paced story which could be a mark of a someone who is more fluent
some unnatural pauses, lack of linking	extremely fluent	too many and unnaturally placed "um" filer	choppy, non-flowing sentences, lack of linking but she sounds nervous	some unnatural word stress, linking and pauses

Recording 6 - CZ012	Recording 7 - CZ001	Recording 8 - CZ030	Recording 9 - CZ031	Recording 10 - CZ041	
commentary	commentary	commentary	commentary	commentary	rater
This seems like a native speaker. No pauses, the speed of conversation is quick and intuitive.	The pace is medium, but thought out. Some phrases have unusual pauses between words where there wouldn't be in a native speaker.	This person has a natural flow but there are few areas where the pausing between words is a little long, but its not distracting.	This speaker seems to struggle a little with the pace, it's a little slow. Uhs and um's in places that seem to indicate internal translating before speaking.	this person speaks slowing and deliberately, trying to find the correct English word to say. Many Ums and pauses.	R1
Subject spoke very quickly with a consistent accent and did not seem to struggle with any major pronunciations.	Subject spoke proficiently but had some trouble with sounds like "w."	Can tell a story with minimal grammar errors, only some minor pronunciation issues.	Slow, hesitant speech but minimal grammar errors and good pronunciation.	Speaker is proficient but speed is slow and there were many mispronunciations.	R2
Speech is fluid, usage of "awesome", no repetition of words, usage of "um"	Speech is fluid, little to no repetition of words	Speech is fluid, word choice is awkward, meter and lilt of speech	Meter and lilt of speech, word choice is awkward	Meter and lilt of speech, pronunciation	R3
the accent sounds native, inflections and pacing were great. Specific/natural words words were used (not just synonyms, which I have noticed that many non native speakers use instead of the most natural word that would come to mind for a native speaker of english)	The speaking was a little bit hesitant but phrasing and word choice were good	sounded calm and comfortable, like she is quite used to speaking english. she used "like" the same way a native would	She sounded a little bit stressed and this was a little bit choppy sounding, I noticed that some phrasing a little bit off too, but her accent was quite good/clear	It was difficult to understand a word or two. She did use some advanced words, but some words/phrases that she used weren't quite the same as a native would use. The speech was also a little bit choppy/staccato. it might be important to note that with this speaker, I feel like I have to focus more on what she is saying; while with some of the others, I didn't need to pay as much careful attention to fully understand.	R4
some unnatural emphasis	the whole recording flows together but with a strange rhythm	overall fluent but the pronunciation threw me off	too many "uh" fillers, unnatural use of the word "well" as a filler	overall not very fluent	R5