MA Thesis – June 2011

Supervisor : Dr. Nivja H. de Jong
Second reader: Prof. dr. René Kager

Native Speakers' Perceptions of Fluency and Accent in L2 Speech

Anne-France PINGET

3427943
Research Master
Linguistics 2009-2011

Utrecht University

Dankwoord

Ik zou alle mensen willen bedanken die op de ene of andere manier ertoe hebben bijgedragen dat deze MA Thesis tot stand is gekomen.

Eerst en vooral mijn begeleidster dr. Nivja de Jong voor haar beschikbaarheid en waardevolle bijsturing zowel tijdens de opzet en de uitvoering van het onderzoek als bij het schrijven. Onze afspraken waren altijd leuk en heel constructief;

Prof. dr. René Kager, mijn tweede lezer, voor zijn inzichtgevende commentaren en suggesties op mijn eerste versie.

Hans Rutger Bosker voor zijn hulp en competentie. Ik heb erg van onze samenwerking genoten;

Dr. Hugo Quené voor zijn raadgeving bij het opzetten van het experiment en de statistische analyse van de resultaten;

Theo Veenker en dr. Iris Mulders voor hun hulp bij het technisch opzetten van het experiment;

Ma famille et mes amis proches pour leur soutien et ces bons moments de temps libre passés ensemble;

Ten slotte alle mensen die als proefpersoon hebben meegedaan.

Anne-France Pinget
Juni 2011

Abstract

The goal of this study is threefold. It is aimed at exploring (i) the relationship between objective properties of speech and perceived fluency, (ii) the relationship between segmental characteristics of speech and perceived accent, and (iii) the relationship between fluency and accent.

We collected 90 speech samples from Turkish and English L2 learners of Dutch. Objective measures of fluency and accent were made for each sample. Forty untrained native speakers of Dutch rated the samples for fluency and accentedness.

The results showed that the temporal measures of fluency were good predictors of fluency ratings, and that their predictive power depends on the type of measures used (i.e. traditional measures per time units, measures per information units, measures that take the L1 into consideration). Furthermore, the segmental measure of accent could predict a small part of accent ratings. Finally, perceived fluency and accent appeared to be weakly correlated, but objective measures of fluency and accent did not add additional explanatory power to the models of perceived accent and perceived fluency respectively.

Content

Part I: Introduction

Part I: Introduction

1.1. Defining the scope

The speech patterns of second language (L2) learners differ from those of native speakers in complex ways. Fluency and accent are often thought to be central aspects of a L2 speaker's speech. Moreover, fluency and accent are likely to be the primary measures of an individual's L2 ability assessed by ordinary native interlocutors on the street, regardless of the speaker's actual proficiency (Derwing Rossiter, Munro & Thomson, 2004). Speaking a language fluently and with a native-like accent is thus frequently the ultimate goal to be attained in mastering a second language.

It seems obvious that L2 speakers are less fluent overall than natives. Despite the fact that the terms "fluent" and "fluency" are regularly used to describe someone's speech production in the L2, there seems to be no consensus concerning what is understood by this concept (Chambers, 1997). Moreover, it is not only the definition of fluency that has been a matter of debate, but also its measurement (Kormos & Dénes, 2004).

Speakers who acquire their L2 later in life are almost certain to exhibit some degree of foreign accent (Scovel, 1988; Patkowski, 1990). Only native speakers of the language can establish how strong this accent is. For both fluency and accent, it is important to know how native speakers perceive a speaker as being fluent or not, and as having a strong accent or not. Thus, detailed studies of which variables underlie native speakers' perception of fluency and accent, and of the process how they balance the multiple features they are attending to (Iwashita, Brown, McNamara & O'Hagan 2008:44) are required in order to gain a good understanding of L2 fluency and accent.

Having a better understanding of these concepts is not only interesting from a theoretical point of view, but has valuable direct implications in several domains. In *language testing*, the candidates' fluency and accentedness is frequently judged. A better delimitation of the features that contribute to non-fluent and accented speech would help human raters to provide a more objective and reliable assessment of L2 speech production based on fixed criteria. Establishing reliable measures of fluency and accent is also important for researchers in *applied linguistics*. In the last few years, we have also witnessed the appearance of numerous software programs for teaching and testing language proficiency. These automatic fluency/accent assessing software programs could also benefit from a clear-cut definition of the criteria distinguishing between a fluent and a non-fluent speaker, and between a speaker with strong foreign accent and a speaker with almost

native-like pronunciation. This knowledge is also useful in *language pedagogy*, since it can help learners to enhance their fluency and diminish their accent.

1.2. Aims of the study

The aim of this study is thus to advance our understanding of the concepts of fluency and accent as characteristics of L2 speech. The study has three main goals. First, we want to investigate the relationship between objective properties of speech and the perception of fluency by native speakers. Secondly, we explore the relationship between segmental characteristics of speech and foreign accent as perceived by native judges. Finally, we investigate the relationship between fluency and accent. We want to gain understanding of how these two aspects of oral production are related. Our expectation is that the perception of fluency is negatively influenced by a strong accent. Conversely, non-fluent speech could possibly be perceived as accented by natives.

A *dual approach* (Cucchiarini, Strik & Boves, 2002) will be adopted: the native perception of fluency and accent in spontaneous L2 speech will be collected in a rating experiment. Subsequently, these ratings will be compared to a number of objective measures of accent and fluency calculated from the speech fragments.

1.3. Outline

In *chapter 2*, we define the two central concepts (i.e. fluency and accent) of this study, review some evidence of previous studies and define our four research questions. *Chapter 3* aims to describe the experimental setting that allows us to answer the questions. In *chapter 4*, we analyze the results obtained in the rating experiment and formulate answers to our research questions. In *chapter 5*, we discuss the findings, the limitations and the implications of this study.

Part II: Background and Research Questions

Part II. Background and research questions

This chapter aims to define the two central concepts under investigation in this study, fluency and foreign accent, and review some previous research on these topics. In the second step, we will have a look at the relationship that may exist between these concepts. The chapter ends with the concrete formulation of the research questions.

2.1. Defining fluency

Pinning down a single definition of the concept of fluency is a difficult task. Studies on L2 fluency often start with the basic distinction made by Lennon (1990, 2000) between two senses of fluency. In the so-called *broad sense*, fluency refers to the global oral proficiency (high command of the L2, overall language performance). Laypeople as well tend to consider fluency as the overall performance, "the communicative acceptability of the speech act" (Sajavaara, 1987: 62). In its *narrow sense*, fluency is considered as one component or one element of oral proficiency, as opposed to other components (e.g. accuracy, appropriacy, etc.). Fluency is an "automatic procedural skill" that encompasses the notions of 'smoothness' and 'fast delivery of speech'. Lennon (2000:26) adopted the working definition of fluency as "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under temporal constraints of on-line processing". This working definition of fluency in a narrow sense has the advantage of being applicable to both native and non-native speakers.

The current study will focus on spoken fluency in L2 speakers and take Lennon's definition of fluency in its narrow sense as a starting point. In other words, we clearly distinguish fluency from proficiency. Indeed, a good linguistic or communicative competence is not always realized in fluent speech, and vice versa, someone's speech can be grammatically/lexically/phonologically 'correct' but not perceived as fluent. As Lennon (1990: 391) explained:

> Fluency differs from the other elements of oral proficiency in one important respect. Whereas such elements as idiomaticness, appropriateness, lexical range, and syntactic complexity can all be assigned to linguistic knowledge, fluency is purely a performance phenomenon; there is (presumably) no fluency 'store.' Rather, fluency is an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently.

Despite the several attempts to pinpoint very precise observable features that characterize a speaker as fluent or not (speech rate, number and duration of pauses, hesitations, etc.), it seems that – in the end – the most tangible fluency is "in the ear of the listener" (Lennon, 1990: 143;

Freed, 1995; Guillot, 1999). L2 performance is fluent if the interlocutor experiences it as fluent. Crucially, fluency should thus be considered from the standpoint of the listener.

In order to pinpoint what is it to be fluent; one can also approach the question by asking what it is to be *non*-fluent. A persistent observation in prior studies (Riggenbach, 1991; Freed, 1995) is that "highly fluent speakers share many features of fluency, while non-fluent speakers are dysfluent in idiosyncratic ways" (Freed, 1995: 255). To put it differently: in order for there to be fluency, it appears that many different conditions have to be met. A fluent speaker needs, for instance, to produce many syllables per minute, he may produce pauses but not within constituents and not of an unnaturally long duration, he should not produce too many false starts or repetitions, etc. In contrast, non-fluency can arise from a single deficiency in any of these different areas (Freed, 1995).

As summarized by Freed (1995), the general picture that emerges from the literature on fluency is that it is a complex phenomenon, which encompasses a multitude of linguistic, psycholinguistic and sociolinguistic features. Reflecting the fact that fluency encompasses so many aspects, studies on L2 fluency have adopted various approaches. Some researchers have investigated the development of fluency longitudinally, comparing the fluency of the same subject group at different points in time (Freed, 1995, 2000; Lennon, 1990; Towell, Hawkings & Bazergui, 1996; Towell, 2002). Others have compared groups of fluent and less fluent speakers and tried to find out how they differ from each other; in what aspect (e.g. Riggenbach, 1991). Another approach consisted in comparing temporal variables of fluency with fluency scores attributed by different types of raters (Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009).

2.1.1. A Threefold distinction

Recently, Segalowitz (2010: 48) proposed to distinguish between three facets of fluency: *cognitive fluency*, *utterance fluency* and *perceived fluency*. (1) *Cognitive fluency* can be defined as the fluency that characterizes a speaker and has to do with the speaker's ability to efficiently plan and execute his speech by integrating the cognitive mechanisms underlying performance. (2) *Utterance fluency* is the fluency that can be measured in a speech sample and has to do with the actual properties of an utterance. One can define utterance fluency objectively by measuring (temporal) aspects of the speech sample such as speech rate, pausing, and false starts. (3) *Perceived fluency* is the judgment that listeners make about the fluency of a speaker on the basis of impressions drawn from their speech sample. The perceived fluency corresponds to what Lennon (1990) and others described as being the only tangible fluency: the fluency "in the ear of the listener".

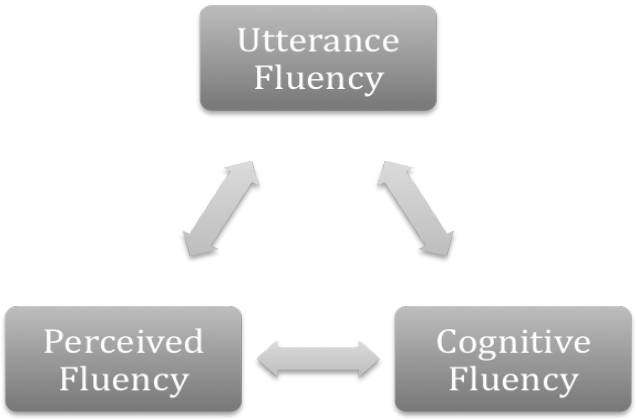


Figure 1: The three facets of fluency as proposed by Segalowitz (2010).

This threefold distinction schematized in Figure 1 is very helpful and relevant when investigating fluency, since each of the three facets correspond to one instance of semiotic models of communication, especially the one proposed by Roman Jakobson (1960). The three key instances in semiotic models of communication are the *sender of the message*, *the message* itself and *the receiver*. These instances correspond with cognitive fluency, utterance fluency and perceived fluency, respectively.

Most studies conducted to date have aimed to investigate the relationship between utterance fluency and perceived fluency. The relationships between utterance fluency and cognitive fluency, as well as between cognitive fluency and perceived fluency have not been studied extensively as yet, except in Segalowitz & Freed (2004) and De Jong, Steinel, Florijn, Schoonen & Hulstijn (*accepted*). In this current paper, we focus again on the relationship between utterance fluency and perceived fluency but with a different approach; namely by taking the L1 fluency into account, as explained in Section 2.1.5. In sections 2.1.2, 2.1.3. and 2.1.4., utterance fluency, perceived fluency and cognitive fluency are each considered in detail.

2.1.2. Utterance fluency: Temporal measures and dysfluency markers

In this section, we focus on the second facet of fluency proposed by Segalowitz (2010); namely *utterance fluency*. As mentioned before, utterance fluency refers to the actual properties of an utterance, the objectively measurable aspects of the speech sample. Tavakoli & Skehan (2005) provided a helpful distinction between three aspects of utterance fluency. These aspects are (1) *speed fluency,* which has to do with the speed at which speech is delivered and can be measured by calculating speech rate such as number of syllables per second (density per time unit); (2) *breakdown fluency,* which has to do with pausing and can be measured by counting the number and length of pauses; and (3) *repair fluency,* which has to do with how often speakers use false starts, make corrections, or produce repetitions. It is an open question whether the different measures of utterance fluency that one traditionally groups into these three aspects are actually related in

practice. In the present study, the objective measures of fluency will exclusively be related to one of these three aspects. In this way, we will be able to assess the role of each aspect separately in perceived fluency.

2.1.3. Perceived fluency: Listener's ratings

Like in previous research (e.g, Cucchiarini et al., 2002; Derwing et al, 2004; Rossiter, 2009), this study aims to investigate whether ratings of native listeners correlate with measures of utterance fluency, and find out which measures are the best predictors of perceived fluency (see research question 1). The innovation of the current study is that we calculate measures of utterance fluency by taking L1 fluency into account, as will be explained in section 2.1.5 (see research question 2).

When investigating perceived fluency, a crucial question has to be asked: on the basis of which aspects do listeners make their judgment about the fluency of a speaker? Lennon (2000: 25) argued that "temporal variables are merely the tip of the iceberg as indicators of fluency". Perceived fluency is probably much more than what can be recorded or transcribed and may include non-verbal aspects such as gesture, self-confidence, etc. Most studies also report influences from a series of non-temporal factors including topic of conversation, situation (e.g. Derwing et al., 2004), accent, grammatical accuracy (e.g. Rossiter, 2009; Kormos & Dénes, 2004) and lexical richness and accuracy (e.g. Lennon, 2000). In the current study, we try to explain the part of variance in listeners' ratings that can be predicted with temporal measures of fluency and not with these various non-temporal factors.

In studies on perceived fluency (e.g. Lennon, 1990; Riggenbach, 1991; Wennerstrom, 2000), it is very common to use specifically trained assessors for evaluating the speech samples. These trained raters are often speech therapists, second language teachers or phoneticians. In contrast, Derwing et al. (2004) and Rossiter (2009) used raters with no linguistic background. Rossiter (2009) explicitly compared the ratings of different groups of judges. She asked experts (L2 teachers and linguistics students), non-experts and advanced non-native speakers to judge twenty-four English L2 learners with various L1 backgrounds. She found strong correlations between objectively-measured fluency variables and subjective ratings irrespective of the raters, trained or untrained, native or non-native. She concluded that "the judges, despite their differing backgrounds, appeared to be paying attention to the same features of oral production when they made their ratings" (2009: 407). The current study will use untrained raters as well.

It is important to note that raters, either trained or untrained, will rate a speech sample according to their own definition of fluency if they do not receive instructions from the experimenters on what precisely to rate. In order to circumvent the potential problem that raters have divergent understandings of fluency, several experimenters decided to instruct their raters

beforehand. Derwing et al. (2004) and Rossiter (2009) specifically instructed their subjects to pay attention to a series of factors (filled and unfilled pauses, false starts, self-repetitions, etc.) when rating the samples. As explained by de Jong et al. (*accepted*), a study with such a procedure and which, in the end, aims to relate listeners' perception to objective measures of fluency runs the risk – from a methodological point of view – of being circular. It is, indeed, more than likely that subjective ratings will correlate with pausing if the experimenter instructs the raters to pay attention to pauses or with speech rate if the raters were instructed to consider the speed of speech. If no instructions are given prior to the rating experiment, raters will use their own definition of fluency to judge the speaking samples. The experimenter then has no control over what the raters are doing precisely, and it is highly probable that the raters will use a definition of fluency in its broad sense as the "global L2 proficiency" instead of considering the pure temporal phenomenon.

In order to solve this problem, it is necessary to know whether it really matters if we instruct the raters or not. Do the raters effectively take the instructions given by the experimenter into account or do they stick to their own, original understanding of fluency? Secondly, it is still unclear if listeners rate fluency holistically, thus basing their rating on their overall impressions or if rating fluency is the sum of the analyses of a definite number of criteria (for instance three or four different temporal factors that the rater unconsciously evaluates before providing his final judgment on fluency). As explained by Derwing et al. (2004), a key question is indeed whether a rater can focus on a certain aspect of fluency while ignoring all the other aspects and variables. These two remaining problems are addressed in the study of Bosker (*in prep*), which has been conducted in parallel to this study.

In conclusion, we see that many prior studies have aimed to correlate objective measures of fluency with perceived fluency. These studies differed widely in their approach and their methodology (e.g. type of speech, type of raters, instructions, temporal vs. non-temporal factors, definition of the measures, etc.). Consequently, the results they provide are not directly comparable to each other. In the present study, our goal is to adopt a consistent methodology: we limit the analysis to purely temporal factors and we provide clear instruction to the untrained raters judging spontaneous speech.

2.1.4. Cognitive fluency: L2 specific problems

Cognitive fluency, which has to do with the speaker's ability to efficiently plan and execute his speech by integrating the cognitive mechanisms underlying performance (Segalowitz 2010) is not the main focus of our study. However, cognitive fluency still deserves some attention, since it is closely related to both utterance and perceived fluency.

Current conceptions of cognitive systems underlying L2 speech production are strongly influenced by the model of Levelt (1989). The model postulates that a speaker has (i) a general knowledge component; (ii) a conceptualizer in which messages are generated; (iii) a formulator in which grammatical and phonological encoding takes place once words have been retrieved from the mental lexicon; and (iv) an articulator that produces overt speech. This model was originally designed for L1 speech production, which was thought to be highly automatized and rapid. More recently, the model has been adapted for bilingual speech production and for L2 production by de Bot (1992) and by Kormos. (2006)

Several proposals have been made to explain where L2 dysfluencies originate in the model and why full automaticity in the L2 cognitive processes is difficult to reach. Segalowitz (2004) claimed that the L2 fluency/dysfluency has its origin in the formulator, in which lexical access, phonological short-term memory and control of attention influence the output of the articulator. Towell et al. (1996) also argued that the site of L2 fluency problems is the formulator, because that is where declarative knowledge is converted into procedural knowledge. O'Brien, Segalowitz, Freed & Collentine (2007) found evidence for the role of phonological memory (understood as a part of the formulator) in L2 fluency. They assessed gains in phonological memory capacities and in oral fluency of L2 learners of Spanish over time. The results indicated that phonological memory was related to the development of L2 oral fluency, and thus that this part of the formulator has a strong connection with L2 fluency.

Derwing, Munro, Thomson & Rossiter (2009) evoked the possibility that L2 oral fluency is also affected in the articulator and the conceptualizer. It is actually possible that different types of dysfluencies result from 'problems' or 'delays' at different stages in the processing system.

As mentioned before, our study is not directly aimed at exploring cognitive fluency. However, better insights into the temporal properties of the L2 utterance, which contribute to the perception of fluency, and into fluency characteristics that are specifically related to the use of the L2 would necessarily enhance our understanding of cognitive fluency.

2.1.5. Fluency in the L2 is related to fluency in the L1

At first glance, it seems obvious that less fluent L2 speech would be characterized roughly by slower delivery of speech, more or longer pauses and more repair strategies. However, many studies have revealed that even native speech is far from being always smooth and continuous; it also exhibits many hesitations and repairs (Lennon, 1990; Riggenbach, 1991). Moreover, it is logical that a speaker who is not so fluent in his L1 (for instance, because he speaks slowly or makes a lot of pauses) cannot be expected to be very fluent in his L2. These facts point at the

importance of taking the L1 fluency characteristics of a speaker into consideration when assessing his L2 fluency.

2.1.5.1. Fluency as an individual characteristic of speech

With the exception of some recent work (e.g. Derwing et al., 2009; De Jong, Schoonen & Hulstijn, 2009), studies on L2 fluency have rarely considered the L1 fluency of their subjects. It is clear, however, that even L1 speakers vary greatly in fluency according to many factors such as personal characteristics, topic, situation, among others.

Recently, De Jong et al. (2009) have demonstrated that the fluency of a speaker in his/her L2 is related to his/her L1 fluency. The aim of De Jong et al. (2009) was to show the need to consider L1 data when studying L2 fluency. They investigated the oral fluency in the L2 Dutch of Turkish and English native speakers by eliciting speech samples in both their L1 and L2. They analyzed different fluency measures (speech rate, pauses, etc.) and reported highly significant L1-L2 correlations for these measures. These high correlations provide strong evidence that a large part of fluency-related phenomena are characteristic of the way individuals speak in general and not just typical for their speech production in the L2. As demonstrated by this study, it is important to obtain fluency data in the L1 to use as baseline measure when investigating the nature of L2 fluency.

As Segalowitz (2010) noted, most researchers have not yet done this, which could potentially have had the effect of individual speech differences unrelated to the use of the L2 providing unwanted sources of noise that may have masked specific L2 fluency phenomena.

2.1.5.2. Segalowitz's proposal

There are several ways one could think of how we could incorporate L1 fluency measures into our analysis of L2 fluency. De Jong et al. (2009) pointed to the need to distinguish between L1-L2 correlations and L1-L2 difference scores. *L1-L2 correlations* (L1 x L2) reflect individual speech properties in the overall fluency; it reveals something about what is common to both languages and thus, not specific to the L2. *L1-L2 difference scores* (L2 results minus L1 results), on the other hand, reflect differences in fluency in the L2 as compared to the L1. So, the difference scores indicate something about how much more difficult the L2 is for the speaker, compared to the L1 as baseline.

Besides correlations and difference scores, Segalowitz (2010) proposed to adopt a new way of calculating utterance fluency measures that also take L1 fluency measures into consideration. Segalowitz' proposal specifically involves the calculation of *residuals* obtained when correlating L1 and L2 fluency with each other. Segalowitz argued that these residuals would be a

more pure measure of L2 fluency than the bare L2 measures, because the role of L1 fluency would have been partialed out in the measures.

Concretely, the purpose of residuals is to isolate the dysfluencies that are specifically related to the use of an L2. First, the assumption is made that the most fluent speech an individual can produce is his L1 speech. We use this L1 as a baseline to partial out the source of variation that is not specifically related to the dysfluencies in L2, but that characterizes a person's general performance in the given test condition. As shown in Figure 2, we correlate – for each objective measure – the L1 result with the L2 result. Then, each L2 result is partialed out to calculate a *residualized score*.

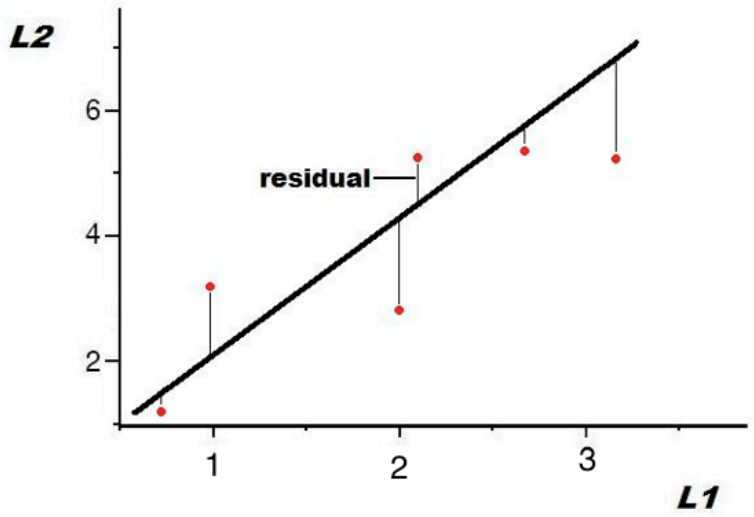


Figure 2: Illustration of the calculation of residuals on the basis of the fluency measures in the L1 and in the L2. Examples: Number of pauses per minute in L1 (x-axe) and in L2 (y-axe).

The residual expresses the difference between the actual observed value and the value that is predicted by the model (given by the regression line). Thus, a speaker with a positive residual (above the regression line) is a speaker who – in our example – produces more pauses in his L2 than what is expected to be the normal L1 to L2 proportion. A speaker with a negative residual (below the regression line) produces fewer pauses in his L2 than what a speaker of his type normally does. The residualized scores allow us to normalize for the L1 in L2 fluency measures.

The goal of this study is to determine whether this new type of objective measures of L2 fluency proposed by Segalowitz (2010) could be significantly better than bare L2 measures (as used in previous studies) at explaining variance in the perceived fluency ratings (see research question 2). In other words, we want to test whether the residualized scores are better predictors of perceived fluency than traditional L2 measures. If this is the case, it would mean that listeners are able to identify a speaker's type; for instance, is it a speaker who produces a lot of "uuh"'s in both his L1 and L2 or is it a speaker who speaks very fast? We want to investigate whether

listeners are capable of detecting this profile (which has nothing to do with the use of the L2, but which is an individual characteristic of the speaker in both his L1 and L2) and whether they take this fact into consideration when providing a judgment on his L2 fluency.

In conclusion, we have – in this section – defined the threefold concept of fluency (i.e. utterance fluency, perceived fluency and cognitive fluency), reviewed previous studies and illustrated on the need to take L1 fluency into account in our analysis.

2.2. Defining Foreign accent

The term *accent* is used to denote "[a] particular way of pronouncing a language, seen as typical of an individual, a geographic region, or a social group. Every speaker of a language necessarily speaks it with some accent or other" (Trask, 1996:4). As pointed out by Richards et al. (1985), this could refer to the region or country, the social class the speaker belongs to and whether or not the speaker is a native speaker.

2.2.1. Accent, foreign accent and intelligibility

In this study, the sociolinguistic dimension present in the above-mentioned definition is not our main focus. We use the term accent especially in the meaning of *foreign accent* (FA) in order to refer to the pronunciation of a language by a non-native speaker that shows deviations from native norms. These deviations characterizing the speaker as a non-native may occur at the phonetic, phonemic or prosodic levels (Gallardo del Puerto, Gomez Lacabex & Garcia Lecumberri, 2007).

Previous studies have shown that both segmental and supra-segmental factors are important for communication effectiveness and efficiency. *Segmental errors*, both at phonemic and allophonic level, can hinder communication, for instance by slowing down word recognition (Smith, 2005; Munro & Derwing, 2008). At the same time, intonation, syllabic structure, lexical stress and rhythm (thus *supra-segmental features*) also help the listener to segment the speech stream and recognize the words more quickly (Cutler, 1984; Cutler & Butterfield, 1992).

There has been a range of studies of native-speaker ratings of foreign accent and intelligibility. What is important is that these studies have clearly showed that foreign accent and intelligibility are two separate concepts. Munro & Derwing (1995), for instance, found that strongly accented speech cannot be equated with lack of intelligibility. Their study examined the correlations between accentedness, comprehensibility and intelligibility in the speech of L2 learners. Eighteen native speakers of English listened to excerpts of English speech produced by 10 Mandarin speakers, transcribed them (intelligibility) and rated them for degree of foreign accent and comprehensibility. Although the utterances tended to be highly intelligible and highly rated for comprehensibility, the accent judgment scores ranged widely on a scale. The findings suggest that although strength of foreign accent is correlated with perceived comprehensibility and intelligibility, a strong foreign accent does not necessarily reduce the comprehensibility or intelligibility of L2 speech. Scheuer (2005: 116) draws the same conclusion that "foreign accent and unintelligibility are not synonymous".

2.2.2. Specific types of accent errors

Abercrombrie (1956) argues that L2 learners – when imitating the accent of native speakers of the target language – concentrate on a limited number of pronunciation problems, which has consequences for intelligibility. In the same way, it has been shown in the above-mentioned studies that, when rating accent of non-native speakers, listeners focused mostly on a certain type of error. The relative importance of the specific pronunciation problems of foreign learners has been termed the "hierarchy of errors" (Johansson 1973, 1975).

Van den Doel (2006) aimed to establish such a hierarchy of errors for Dutch L2 speakers of English and to investigate which factors play a role when native speakers consider these errors. Van den Doel (2006) grouped the errors that L2 speakers may produce into five categories presented in Figure 3: *phonemic, realisational, distributional, stress* and *suprasegmental*.

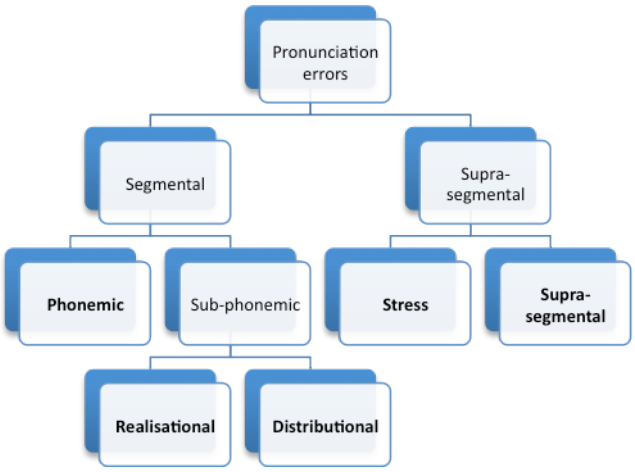


Figure 3: The types of pronunciation errors (based on Johansson, 1978 and Van den Doel, 2006).

The first three categories are used with reference to *segmental pronunciation errors* in non-native speech. The term "*phonemic errors*" refers to the L2 realisations of a particular sound perceived by native speakers as different phonemes (e.g. the realization of the Dutch /w/ as [ʊ] which is perceived as [v] by natives). Johansson (1978) also distinguishes these "phonemic errors" from "sub-phonemic errors". The sub-phonemic errors are – in contrast to the phonemic ones – not perceived as causing the substitution of one phoneme by another. They can be divided into "realisational errors" and "distributional errors". *Realisational errors*, such as the use of the English alveolar approximant [ɹ] instead of the Dutch alveolar trill [r], involve pronunciations which are likely to be perceived by natives as unusual, stigmatised or deviant allophonic realisations of a particular phoneme (Collins & Mees, 2003: 296). *Distributional errors*, such as adopting a rhotic pronunciation while the target L1 is non-rhotic, have to do with the distribution of phonemes in the target language. The border between realisation errors and distributional errors may be vague in some cases.

The two last categories, *stress* and *supra-segmental*, cover errors that are generally all classified in the literature as supra-segmental. *Stress errors* refer to the misplacement of primary stress. These errors are described by Collins & Mees (2003) as the "most significant" errors of Dutch learners, namely the most salient in the ear of Dutch natives. Finally, *supra-segmental errors* are errors connected with supra-segmental phenomena such as intonation, contraction and weakening.

The study of Van den Doel (2006) which was based on the presented categorization of errors, showed a range of interesting results. First, it turned out that errors involving stress and vowel reduction are considered by natives to be among the most important. Second, there is a general tendency for phonemic errors to be considered as more important than those of a realisational or distributional nature. Thirdly, there is no evidence that support the assumption that errors involving consonants are more important than those of a vocalic nature, as it has been suggested for instance in Johansson (1978) and Munro & Derwing (1995). Vowel errors were also ranked very highly in the hierarchy of errors. Finally, the errors classified as very important/salient are often either stereotyped as foreign pronunciations or stigmatized as realisations associated with L1 regional or social varieties. This indicates that foreign accents are not only judged on the basis of intelligibility, but are also evaluated against L1 standards for acceptability.

2.2.3. Factors affecting perceived foreign accent

Most studies on foreign accent focus on the link between *perceived accent* (i.e. foreign accent as perceived by native speakers) and – what could be named – "*cognitive accent*" (i.e. speaker-internal factors that affect accent). The age of L2 acquisition, for instance, has widely been shown to be a powerful predictor of global foreign-accent ratings (Flege, Munro, & MacKay, 1995; Oyama, 1982; Tahta, Wood, & Lowenthal, 1981). A series of comparable studies have described the correlational effects of other cognitive factors such as length of residence in the L2 experience, L2 environment, motivation, quantity and quality of L2 input, relative use of the L1 and L2, and social evaluation of L2 learners on accent measures or ratings (Flege, 1988; Ryan, Carranza, & Moffie, 1977; Oyama, 1982; Flege et al., 1995; Flege, Frieda, & Nozawa, 1997; Moyer, 1999; Trofimovich & Baker, 2006).

Studies on the effect of utterance-internal factors ("*utterance accent*") are more scarce. These studies tend to relate the number of segmental errors in the L2 utterance (i.e. phone substitutions, deletions, or insertions) and its overall prosodic accuracy to the perception of global foreign accent and its comprehensibility (Brennan & Brennan, 1981; Ingram & Pittam, 1987; Anderson-Hsieh, Johnson & Koehler, 1992; Magen, 1998; Munro & Derwing, 1999).

Two studies examined native listeners' responses to longitudinal data collected from foreign speakers at various points in their acquisition of the second language. Major (1987) conducted a study on Brazilian Portuguese speakers acquiring English. The results showed a significant inverse correlation between degree of accuracy in VOT and perceived degree of foreign accent. Another study on Vietnamese children acquiring Australian English (Ingram & Pittam, 1987) directly focused on comparing the effects of consonant and vowel production on perceived accent. The study showed that improvements in vowel quality, more than improvements in consonant production, affected native listeners' judgments of accent change.

Two more recent studies that have explicitly been aimed at assessing the contribution of specific phonetic and phonological factors to the perception of global foreign accent have revealed interesting results. Magen (1998) investigated different segmental and supra-segmental factors in the speech of native Spanish speakers with a heavy accent in English. These factors affected either the syllable structure (e.g. schwa epenthesis); the vowel quality (e.g. vowel reduction, tense-laxness); the consonants (e.g. final deletion, manner, fricative voicing, stop voicing); or the phrasal stress patterns. The goal was to examine the relative weight of these different types of errors on accent ratings. Results showed that listeners were sensitive to syllable structure, final deletion, consonant manner, and phrasal stress. Listeners were, however, not sensitive to voicing differences, which contrasts with the findings of Major (1987) in which accuracy in VOT did correlate with perceived accent.

Anderson-Hsieh et al. (1992) investigated the relationship between raters' judgments of nonnative pronunciation and actual deviance in segmentals, prosody and syllable structure. Speech samples of speakers with 11 different L1 backgrounds were rated on pronunciation. The correlation between these accent ratings and the deviance found in each area of pronunciation showed that (i) errors in all areas have a significant influence on the ratings and (ii) supra-segmental variables proved to have the strongest influence.

In conclusion, all these studies have suggested a relationship between accent error scores and accent ratings. Even though the influence of some factors on accent ratings was not clear (e.g. errors in voicing), most segmental and supra-segmental factors were closely related to accent perceived by native listeners. From the above-mentioned studies, it was shown that errors in vowels correlated especially well with accent ratings. Important to mention is that the calculation of so-called "supra-segmental" factors in these studies was actually often based on segmental measures. For instance, the accuracy of stress was, in fact, computed on the basis of vowel reduction patterns.

In the present study, we focus on *phonemic errors* and their relationship with perceived accent. If we were to analyse supra-segmental errors as well, we would run the risk of considering factors that, by definition, already correlate. Indeed, some supra-segmental errors, such as stress

and rhythm, are highly linked to measures of fluency. For instance, considering lexical stress (which is a supra-segmental aspect of speech) implies that one look at, among other things, the duration of vowels, and vowel duration necessarily has consequence on speed fluency. Since ultimately we aim to incorporate objective measures of accent as predictors for fluency ratings, and the vice versa (incorporate objective measures of fluency as predictor for accent ratings), such an overlap between measures of fluency and measures of accent is not desired. Furthermore, sub-phonemic errors are more difficult to detect than phonemic errors and are considered less salient/important than phonemic errors by raters, as has been shown by Van den Doel (2006). Moreover, an analysis of sub-phonemic errors would require acoustic measures of the phoneme quality, which goes beyond the scope of this study. For instance, the difference between an aspirated realization [t^h] of the Dutch phoneme /t/ as often produced by English speakers and an unaspirated realization [t] may be so subtle that is not detectable within our analysis. Our accent analysis will thus be restricted to phonemic errors.

2.3. Relationship between fluency and accent

In sections 2.1. and 2.2., we discussed the two key notions of these studies: fluency and accent. We reviewed previous studies that investigated the relationship between objective measures of fluency or accent and ratings given by native listeners.

In the present section, we focus on the relationship that may exist between these two aspects. Theoretically, one could first ask whether one is actually able to distinguish fluency and accent. Are we capable of considering these phenomena separately when rating a speech sample? Bond, Stockmal & Markus (2008: 7) claimed – based on indirect evidence – that this is the case: "Although phonological accuracy and fluency appear to be related measures in non-native speech, they are separable properties of speech". There is indeed evidence for the fact that

> Listeners are able to make reasonably consistent and accurate fluency judgments without knowledge of the phonology of a language. Apparently, naïve listeners have expectations about normal fluent speech which they can use as perceptual anchors in judging utterances even when listening to a language which they do not know. (Bond et al. 2008: 7)

As a consequence, it makes sense to consider accentedness and fluency as separate phenomena. However, it is likely that fluency and foreign accent are related in L2 speech. Investigating to what extent this is the case is one important purpose of our study (see research question 4). In section 2.3.1, the potential effects of fluency on foreign accent ratings will be described, while section 2.3.2 focuses on the effects of accent on fluency ratings.

2.3.1. Effects of fluency on accent ratings

Munro and colleagues have conducted several studies that investigate whether fluency has an effect on accentedness and comprehensibility judgments. These studies systematically point to the effect of speech rate in the ratings of foreign accent.

Munro & Derwing (1998) investigated whether fluency has an effect on the perception of accent. They tested whether accented speech speeded with a digital speech compressor-expander by ten percent sounds less accented than speech produced at a normal rate. They found that fast stimuli were rated as less accented than stimuli presented at normal and slowed rates.

Based on their previous studies and on evidence from Anderson-Hsieh & Koehler (1988), Munro & Derwing (2001) started from the assumption that there is indeed a relationship between speech rate and perceived accent. Their aim was to show that this relationship is curvilinear rather than linear. They explained that, as long as the speed of delivery remains manageable from a processing standpoint, the listeners should benefit from the acceleration in speech rate. However, when the same speech is presented at a particularly fast rate, the listeners may be at a

disadvantage, since very fast speech places extra demands on the listener. The speech may therefore be rated as more accented than slightly accelerated speech. Very slow speech may be difficult to process, because listeners are required to keep information in short-term memory for a longer period of time. Furthermore, listeners are also more inclined to notice phonological errors and assign poorer ratings to the speech. In their experiments, they used speech compression-expansion software to increase and decrease speaking rate. They found that both slow and fast stimuli were rated as more accented than stimuli presented at normal rates. A linear regression revealed that the speaking rate could account for 15% of the variance of accent ratings. They claimed that the data pointed to the fact that there is indeed a relationship between speech rate and perceived accent, and that this relationship is curvilinear.

Although these studies provide clear evidence for the role of speech rate in accent ratings, we cast doubt on the methodology used by Munro & Derwing (1998, 2001). In their studies, Munro & Derwing systematically asked their raters to score the stimuli both on accentedness and intelligibility simultaneously. Such a procedure necessarily implies that the ratings on one scale (e.g. intelligibility) will influence the scores on the other scale (e.g. accentedness), since the very same listeners rate both aspects at the same time. Besides, it is obvious that speeded or slowed speech will be rated as less intelligible than speech produced at a normal rate. Therefore, it is very likely that the scores on accentedness were largely influenced by the intelligibility scores, which were logically lower for speeded and slowed speech. Moreover, Munro & Derwing (1998, 2001) based their findings on a manipulation of speech rate only. Speech rate is one way of measuring fluency, but this measure does not encompass all aspects of fluency. Speech rate does not, for instance, take repair strategies into account (i.e. corrections, repetitions, etc.). Thus, it seems clear that more research is required in order to explore the role of fluency in the way listeners rate accentedness.

2.3.2. Effects of accent on fluency ratings

There is a large body of evidence that shows that natives evaluate speakers with non-native accents negatively on a range of different aspects (Eisenstein, 1983; Munro & Derwing, 1995; Leather, 1999; Major, Fitzmaurice, Bunta & Balasubramanian, 2005; Scheuer, 2005). One of these aspects is fluency: accentedness is claimed to be a factor that can potentially influence fluency ratings (negatively). According to Freed (1995), accentedness seems to be one of the most important factors by which raters claimed to be influenced when reporting on their experiences during fluency rating tasks. When asked to describe the criteria on which they based their fluency evaluations, the judges of this experiment indicated that they were also influenced by a variety of non-temporal factors, for instance richness of vocabulary, accuracy of grammar, clarity of voice,

ease, confidence in speech and accent. As Freed (1995: 136) put it, "half of the raters selected 'accent' as an important speech quality which contributed to the evaluation of the subjects as being fluent/non-fluent". Rossiter (2009) made a similar observation concerning the influence of pronunciation.

Several studies have attempted to test this influence of accentedness on perceived fluency experimentally. However, the reported findings vary widely from study to study. In his study of the role of pitch and phrasal segmentation, Wennerstrom (2000), for instance, showed that prosody affects listeners' perception of L2 fluency. Derwing & Rossiter (2003) also found that prosodic accuracy contributes to the overall impression of fluency. The underlying assumption is that inaccurate prosodic patterns are characteristic of accented speech. However, one could wonder whether the analysis of prosody and prosodic accuracy do not directly interfere with pausing (being a component of fluency), since the prosody will necessarily be modified in a speech sample containing a large number of (too long) pauses. Therefore, it is not surprising to find a relationship between prosody and fluency in these studies.

The findings of Derwing et al. (2004) were not so clear-cut either. They examined the relationships between perceptions of fluency, comprehensibility and accentedness. They collected speech samples from 20 beginner Mandarin learners of English. Twenty-eight untrained judges rated fluency, comprehensibility and accent. They found a strong relationship between fluency and comprehensibility, whereas the correlation between fluency and accentedness was somewhat lower. They concluded that their findings show a relatively weak relationship between accentedness and fluency. However, the results of this study should be treated carefully, since, again, the very same group of speakers had to rate the three different aspects (fluency, comprehensibility and accent) in the same speech samples. It is, theoretically speaking, not so surprising to get correlations between aspects that one has to consider at the same time: the rating of one aspect can very much influence the rating of the other. Furthermore, the samples used in their study were drawn from low-proficiency speakers. It is possible that raters have judged fluency and comprehensibility more strictly than accent, since a good accent can be thought to become a requirement only when the L2 speaker reaches a higher proficiency level.

In conclusion, it appears that only a few studies have explored the relationship between fluency and accent. The results of some of these studies have to be treated with caution for the reasons mentioned. Thus, overall we have very few valuable insights into the factors of fluency and accent that may influence listeners' judgments on accent and fluency respectively.

2.4. The present study: Research questions

The present study focuses on two aspects of utterances produced by L2 speakers, *fluency* and *accent*. Both fluency and accent may be assessed objectively or subjectively. The resulting four subcomponents presented in Figure 4 will form the core elements of this study. As mentioned before, objective measures of fluency refer to the specific aspects of utterance fluency that can concretely be measured in speech. Subjective fluency is fluency as rated by natives. In the same way, objective measures of accent refer to the objectively measurable characteristics of foreign accent in the L2 speech and subjective accent refers to the perception of accent by native raters.

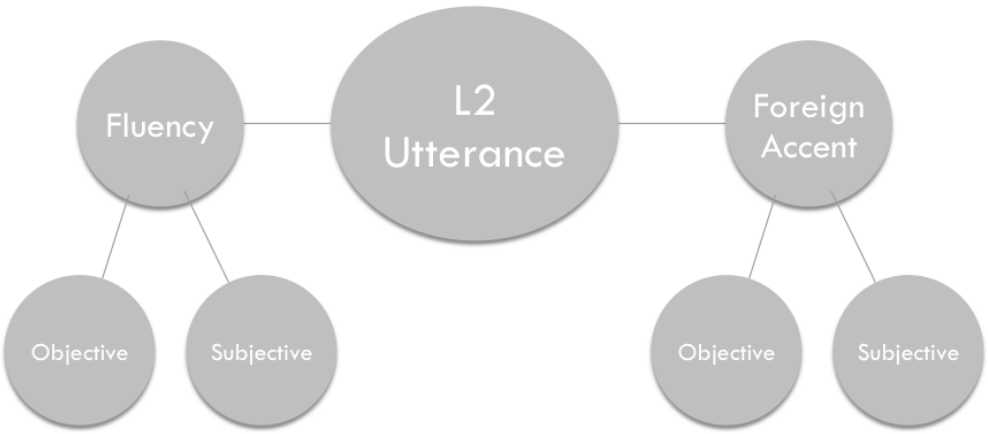


Figure 4: Four components of L2 utterance investigated in this study.

In the previous sections, we have already touched upon the topics we will investigate in the present study. In this section, our four main research questions are presented and hypotheses are formulated on the basis of previous studies.

Research questions 1 and 2 – On fluency

In line with previous research, we want to determine **which objective measures of fluency can predict perceived L2 fluency (RQ1).** Several studies have investigated this relationship before and have found that pausing phenomena (i.e. breakdown fluency) and speech rate are primary factors influencing fluency ratings. Therefore, we logically expect that measures of speed and breakdown fluency will predict fluency ratings. Since we use objective measures of fluency that are specifically related to one single aspect of fluency, we aim to make a distinction between the part of perceived fluency that can be explained by speed fluency aspects and by breakdown fluency aspects separately. With respect to repair fluency, the literature seems to suggest that there is a weak relationship between repair fluency and perceived fluency. For instance, Cucchiarini et

al. (2002) did not find any relationship between fluency ratings and the factor number of dysfluencies (which covered among others repetitions and corrections). Therefore, we do not expect that measures of repair fluency will predict fluency ratings well.

Furthermore, we test a new type of utterance fluency measures proposed by Segalowitz (2010): the residualized scores (i.e. L2 measures from which L1 measures have been partialed out). We want to find out **to what extent these residualized scores correlate with perceived fluency and whether they are able to predict perceived fluency better than the traditional measures of utterance fluency (RQ2).**

Research question 3 – On accent

The third research question concerns accent. Our goal is to investigate **whether a segmental measure of accent can predict perceived accent (RQ3).** By correlating a segmental error measure with perceived accent ratings, we want to determine how much of the accent rating variance can be explained by segmental characteristics of the L2 speech. On the basis of previous studies reviewed in section 2.2.2, we may expect our segmental measure to explain a non-negligible part of the variance in accent ratings given by native listeners.

Research question 4 – On the relationship between fluency and accent

Our fourth research question concerns both fluency and accent, and may be divided into three sub-questions. The first general question we ask is to **what extent fluency and accent ratings are related to each other (RQ4a)**. Previous studies indicate that we may reasonably expect a correlation between accent and fluency: the higher the accent ratings (strong accent), the lower the fluency ratings (not so fluent). Furthermore, we ask **whether an objective measure of accent can predict fluency ratings (RQ4b)**. In the model which tests RQ1, we add an objective measure of accent and check whether this factor adds some explanation of perceived fluency. In this way, we will find out whether accentedness is an interfering factor that plays a role when a native speaker rates L2 fluency. Finally, we ask **whether objective measures of fluency can predict accent ratings (RQ4c).** In the same way, we add objective measures of fluency and check whether these factors add some explanation of perceived accent. Thus, we investigate whether fluency is an interfering factor that plays a role in the perception of foreign accent. The literature described in Section 2.3. indicates that we may expect objective measures of fluency and accent to play a role in perceived accent and fluency respectively.

Part III: Methodology

Part III. Methodology

In order to answer the research questions, we developed an experimental design that allowed us to collect both objective measures of fluency and accent, and measures of perceived fluency and perceived accent on the same data. We selected a range of speech recordings of second language speakers (with L1 Turkish and L1 English). The selection of stimuli is described in section 3.1. In section 3.2., we describe the measures of fluency and accent that we use as *objective measures*. In section 3.3., the rating experiment is described that allows us to collect data of perceived fluency and perceived accent that will constitute our *subjective measures*.

3.1. Stimuli

3.1.1. Speakers

Speech recordings from native and non-native speakers of Dutch were obtained from the "Unravelling second language proficiency" project from the University of Amsterdam[1] (described in de Jong et al., *accepted*). Both the annotations of the recordings and several test scores of every speaker were made available for this study.

Since we know that the proficiency level in the L2 and the impressions natives draw from a speech sample when rating fluency are related to each other (Freed, 1995; Rossiter, 2009), we selected fifteen L2 speakers of both groups of the corpus (L1 English and L1 Turkish) and we matched them for proficiency. Beglar & Hunt (1999) have shown that vocabulary tests are highly representative indicators of overall proficiency. Therefore, we matched our subjects on the basis of a Dutch vocabulary test of 116 lexical items. The test was a 'discrete point' and 'indirect' vocabulary test (Hulstijn, 2010) the speakers took within the same research project. A one-way ANOVA on the vocabulary scores confirmed that there was no statistically significant difference between the vocabulary scores of the two non-native groups (English (mean(SD)= 67.5(15.7)), Turkish (mean(SD)=64.1(18)). Therefore, we assume that the Turkish non-native speakers of Dutch and the English non-native speakers of Dutch are matched for language proficiency. Furthermore, no differences in sex (19f/11m) were found (men: mean(SD)=66.8(17.4); women (mean(SD)=65.2(16.7)) and no interaction between *sex* and *language* (all F's <1).

In addition to these 30 non-native speakers with an intermediate level of Dutch proficiency, we selected 8 native speakers of Dutch who took the same vocabulary test and

[1] Sponsored by the Netherlands Organisation for Scientific Research, grant number 254-70-030. Principal investigators: Nivja de Jong, Margarita Steinel, Arjen Florijn, Rob Schoonen, and Jan Hulstijn, Amsterdam Center for Language and Communication, Faculty of Humanities, University of Amsterdam.

performed the same tasks as the non-native speakers. The recordings of native speakers functioned as a reference. Introducing native fluency and accent into our experiment indeed forced the listeners to use the whole rating scale that we propose in the rating experiment between "highly fluent" and "highly dysfluent" and between "no accent" and "very strong accent" instead of limiting themselves to a smaller part of the scale that is reachable for L2 speakers. The fluency and accent results of these speakers will however not be included in our final analysis. The native speakers were selected based on their score proximity to the native speakers' (n=54) average score on the vocabulary test (mean(SD)=106 on 116(5.32), range=24).

3.1.2. Speaking tasks

All 38 speakers (30 L2 speakers and 8 natives) performed eight different computer-administered speaking tasks in the frame of the "Unravelling second language proficiency" project. The non-native speakers performed these tasks both in their mother tongue and in Dutch (L2). The eight tasks performed in the mother tongue were different, but highly similar to the ones performed in the L2. These tasks had been designed to cover the following three dimensions in a 2 x 2 x 2 fashion: *complexity* (simple, complex), *formality* (informal, formal) and *discourse type* (descriptive, argumentative). The task instructions specifically mentioned that participants should try to imagine that they were addressing people in each task and they were instructed to "role play" accordingly. Participants had 30 seconds preparation time and 120 seconds speaking time per task. As a warm up, participants carried out a practice task.

This design has the advantage of providing us with spontaneous speech in the form of conversational monologue. Some risk of uncontrolled variability is always associated with spontaneous speech. Nevertheless, the fact that the subjects received precise instruction on the situation in which their speech act should take place, we could control to a certain extent for a range of factors such as topic, situation, vocabulary, etc.

All tasks are described in Appendix I. For the analysis of the L1 of the speakers, all eight tasks were selected. This allowed us to gain extensive insight into the profile of a speaker in his L1. For the rating experiment and thus the analysis of the L2, we selected three tasks from the eight (namely Task 2, 4 and 8 presented in Table 1). Our criteria for selection were the different task characteristics and the quality of the sound recordings.

	Task characteristics	Description
Task 2	simple, formal, descriptive	The participant, who witnessed a road accident some time ago, is in a courtroom, describing to the judge what happened
Task 4	simple, formal, argumentative	The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.
Task 8	complex, formal, argumentative	The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues which one of three alternative plans for building a car park he/she prefers.

Table 1. Descriptions of the three tasks selected from the corpus performed by the speakers in their L2.

The recordings of the speaking tasks were of varying sound quality, noise level and total duration (around 2 minutes). Studies on experimental social psychology (e.g. Ambady, Bernieri & Richeson, 2000) have shown that a thin-slice (i.e. a brief excerpt of expressive behavior sampled from a behavioral stream (Bhat, Hasegawa-Johnson & Sproat, 2010)) contains enough information to make rapid and impressionistic judgments about certain behavioral characteristics and that these judgments are reasonably accurate.

Therefore, thin slices of approximately 20 seconds were extracted from approximately the middle of each original recording of 2 minutes. In previous studies, it has been demonstrated that 20 seconds is an appropriate length of speech samples for both fluency and accent evaluation by judges. Derwing et al. (2004) used 30-second samples, while Derwing et al. (2009) took 20-second samples. They determined that this amount of time is sufficient for raters to make reliable assessments. Using longer speech samples would only lengthen the duration of the experiment, which increases the demands on raters, forces them to hold more in memory and increases the likelihood of primacy and recency effects.

Each 20-second thin slice started at an AS-unit boundary, i.e. "a single speaker's utterance consisting of an independent clause, or a sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn & Wigglesworth, 2001) and ended at a silent pause in the speech. The thin slices were subsequently resampled to a sampling frequency of 44100 Hz and scaled to an intensity of 70 dB. As a result, we obtained 114 thin slices (38 speakers x 3 tasks) for experimental use.

3.2. Objective measures of fluency and accent

In this section, we explain how the speech material was transcribed and annotated (3.2.1). Furthermore, we describe the factors we selected as objective measures of fluency (3.2.2.) and of accent (3.2.3).

3.2.1. Transcription and annotation of the speech material

All speech recordings were transcribed and annotated by two research assistants who worked in tight collaboration with each other.[2] Each speech recording was paired with its transcription with the software CLAN. The transcription was split up into so-called "*AS-units*". Foster et al. (2001) have shown that the AS-unit (Analysis of Speech Unit) is the most optimal way of dividing transcribed data into analyzable units for many reasons. As defined by Foster et al. (2001), an AS-unit is "a single speaker's utterance consisting of an independent clause, or a sub-clausal unit, together with any subordinate clause(s) associated with either"[3]. All silent pauses in the recordings were detected by hand. Furthermore, the recordings were annotated with filled pauses such as "uh", "uhm", "er", "mm", etc.), corrections (false starts, reformulations and self-corrections), repetitions (repetitions of exact words, syllables or phrases), number of syllables, lengthening of sounds and lip smacking.

3.2.2. Objective measures of fluency

Using the annotations as described in the previous section, we calculated several objective acoustic measures of fluency. These objective measures are temporal measures of speech and a series of dysfluency markers that have emerged in prior studies (e.g. Grosjean, 1980; Lennon, 1990; Riggenbach, 1991; Freed, 1995; Towell et al., 1996) as most salient in characterizing fluency in non-native speakers and appeared to be potentially related to perceived fluency.

---

[2] Sponsored by Pearson (Pearson Language Tests (PLT) research program). Principal investigator: Nivja de Jong, University Utrecht.

[3] *An independent clause* consists minimally of a clause including a finite verb (like in (1)). An *independent sub-clausal unit* consists of *either* one or more phrases, which can be elaborated into a full clause by means of recovery of ellipted elements from the context (like in (2)) *or* a minor utterance that could be defined as 'irregular sentences' (like in (3)). *A subordinate clause* consists of a finite or a non-finite verb element (like in (4)).

(1) That's right
(2) A: How long you stay here? – B: Three months (1 AS-unit)
(3) Yes
(4) I participate in an organization in France which is called department of agricultural extension

Before presenting each measure, we give some general precisions that concern all temporal measures of fluency. First of all, silences of 0.25 seconds or longer in the speech are considered as pauses. Towell et al. (1996) pointed out that there has been an ongoing debate among researchers about the cut-off point of pause length. If this point is too low, the pause may signal the stop phase of a plosive or may be classified as micro-pauses (Riggenbach 1991) which are not regarded as hesitation phenomena. If the cut-off point is too high, some amount of time may be omitted from the analysis. Therefore, Towell et al. (1996) argued that pauses above 0.25 seconds (they actually use 0.28 seconds for practical reasons explained in Towell et al. 1996: 91) are the most reliable pause exclusion criterion. The micro-pauses of less than 250 milliseconds were therefore left out of consideration.

Secondly, two duration times are used to calculate measures of fluency; one time taking pausing into account, the other not. *Duration 1* is the duration of speech excluding silences (pauses of 0.25 sec or longer), and *Duration 2* is the total duration of speech including silences.

As mentioned before, Tavakoli and Skehan (2005) argued that utterance fluency is a construct that encompasses three specific aspects: speed fluency, breakdown fluency and repair fluency. When proposing temporal measures, we try to keep these aspects separate. In our research, none of the measures interfere with each other. We propose one measure for speed fluency, namely the mean syllable duration (MSD). For breakdown fluency, we selected four measures: the mean number of pauses (between AS-units), the number of silent pauses per minute, the number of filled pauses per minute and the mean length of silent pauses (MLP). Finally, two measures for repair fluency were selected: number of corrections per minute and the number of repetitions per minute. In previous studies on fluency, one traditionally calculates two more measures: speech rate and phonation time ratio. We decided not to select these measures, since they do not measure one specific aspect of utterance fluency, but encompass several aspects. Speech rate is a measure calculated as the number of syllables per total time (including pauses). With speech rate, breakdown and speed fluency are taken together into one measure that encompasses information about both the speed of speech delivery and pausing patterns. Speech rate is thus a measure in which many aspects of fluency are confounded. The phonation/time ratio gives information about both the number of pauses and their length. The more silent pauses the speaker produces and the longer they are, the lower the phonation/time ratio. As we ultimately aim to correlate the different aspects of utterance fluency with perceived fluency, measures in which several aspects already overlap are not desirable.

We could theoretically have computed the four measures that are related to the number of one of the dysfluencies[4] by dividing them either by the duration 1 (speech time), the duration 2

[4] (1) the number of silent pauses, (2) the number of filled pauses, (3) the number of corrections and (4) the number of repetitions

(total time) or by the total number of syllables. When using duration 1 or the number of syllables, we only consider the amount of speech the speaker actually produces. When using duration 2, we take the total amount of time into consideration that was at the speaker's disposal to produce speech. We tested the three different ways of computing the relevant measures and checked the correlations between them. It turned out that all correlations were very strong (all *r*'s > .90). This means that the way of calculating the different measures does not really matter and will probably have no consequences on the rest of the study. In view of this, we chose to use *duration 2* for the calculation, since this was by far the most commonly used one in previous studies and thus allows us to compare our results to previous research.

The seven objective measures of fluency and the way they are calculated are presented in Table 2. *MSD* (*Mean Syllable Duration*) is the inverse of the traditionally calculated articulation rate. *MLP* is the *mean length of silent pause*s. In the analysis, the logarithm of *MLP* will be used instead of the raw measure. Taking the logarithm of time units has the advantage of transforming the measure into a normally distributed equivalent. As far as the measures of repair fluency are concerned, we distinguished between (1) *the number of repetitions per minute* (repetitions of exact words, syllables or phrases) and (2) *the number of corrections per minute* (false starts, reformulations and self-corrections).

ASPECT	NO	ACOUSTIC MEASURE	CALCULATION	ABBREVIATION
SPEED	1	Mean Syllable Duration	dur1/number of syllables	MSD
BREAKDOWN	2	Number of pauses between AS-units	number of all silent pauses between AS-units/number of AS-units boundaries	Number of P (/b/ AS)
	3	Number of silent pauses per minute	number of silent pauses /dur2	Silent P/min
	4	Number of filled pauses per minute	number of filled pauses /dur2	Filled P/min
	5	Mean length of silent pauses	Logarithm of the total length of silent pauses (dur2-dur1) /number of silent pauses	MLP
REPAIR	6	Number of corrections per minute	number of corrections/dur2	Cor/min
	7	Number of repetitions per minute	number of repetitions/ dur2	Rep/min

Table 2: List of selected acoustic objective measures of fluency.
dur1 = duration of speech fragment excluding silences of >250ms;
dur2 = duration of speech fragment including silences.

In contrast to previous studies, we have consistently tried to uniform the seven measures, so that they all become measures of dysfluency (instead of measures of both fluency *and*

dysfluency, as in most previous studies). Concretely, it means that the mean syllable duration (MSD) is calculated inversely. Some researchers have already used similar operations for speech rate (Crystal & House, 1990; Quené, 2008; De Jong et al., *accepted*). Our goal is to systematize these attempts, so that all measures are in line with each other. The main advantage of such a practice is that now – for all measures – the higher the score is, the less fluent the speech is.

As previously explained, only a small sample of each speech recording was selected (20-second thin slice of the 2-minute recording). In order to be sure that our selected thin slices are representative of the original recordings, correlations were calculated between the objective speech measurements of these thin slices (of approximately 20 seconds) and those from the original recordings (of approximately 2 minutes). Strong, statistically significant correlations were found for the majority of the objective measures (r = .70-.90), as given in Table 3.

Measures in L2 thin slices (20 sec)	**Pearson's r correlation with same measures in L2 (whole task) (2 min)**
MSD	.87
Number of P(/b/ AS)	.63
Silent P/min	.77
Filled P/min	.88
MLP	.83
Cor/min	.60
Rep/min	.71

Table 3: Pearson's r correlations between the acoustic measures of the thin slices (20 sec) and the measures of the whole recordings (2 min) (all measures have a significance of lower than .05).

For two measures (number of pauses between AS-units and number of correlations per minute), the Pearson'r were around .60, which is a lower but still moderate correlation.

In conclusion, we argue that these correlations are high enough for our 20-second thin slices to be representative for the whole 2-minute tasks that the non-native speakers performed.

3.2.3. Objective measure of accent

As an objective measure of accent, we calculate a *phonemic error rate*. As explained in Chapter 2, our study will be limited to *phonemic segmental errors*, one of the types of accent errors. In what follows, we explain which phonemes we selected and how we calculated the *phonemic error rate*.

3.2.3.1. Selection of phonemes

Dutch is considered a stress-timed language, and its syllables can have different durations. The Dutch phonology system consists of 11 vowels, 6 diphthongs and 23 consonants (Gussenhoven,

1999). Previous studies on frequent L2 errors in Dutch (e.g. Aan de Wiel, M., van den Brink, G. & S. Struijk van Bergen, 1991) point to a common problem with vowels (both monophthongs and diphthongs) rather than consonants. This may partly be due to the relatively high number of vocalic phonemes in Dutch (Lindblöm, 1984) as can be seen in Figures 5 and 6. Moreover, the difficulties with vowels may be due to the fact that learning to articulate new vowels intrinsically requires more effort than learning to articulate consonants (Flege, 1988).

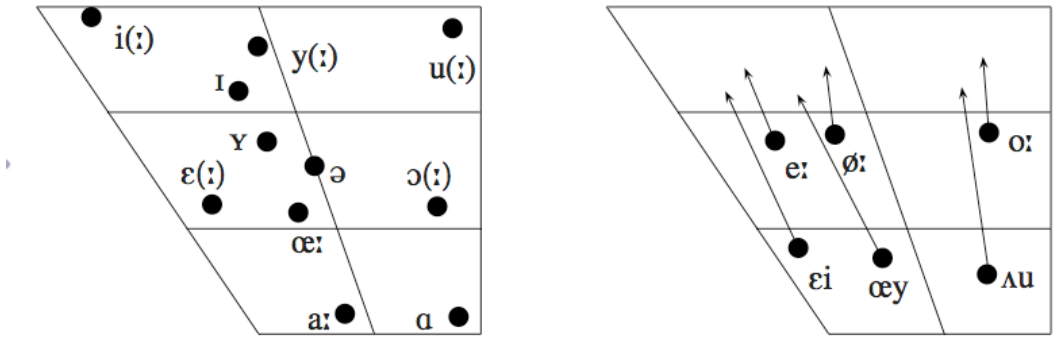


Figure 5. The Dutch vowel system (based on Gussenhoven 1999: 76).

	Bilabial	Labio-dental	Alveolar	Post-alveolar	Palatal	Velar	Uvular	Glottal
Plosive	p b		t d	(c)		k		(ʔ)
Nasal	m		n	(ɲ)				
Fricative		f v	s z	(s) (ʒ)			χ	ɦ
Tap			ɾ					
Approximant		ʋ			j			
Lateral approximant			l					

Figure 6. The Dutch consonant system (based on Gussenhoven 1999: 74).

On the basis of the study of Neri, Cucchiarini & Strik (2006), we made a selection of 12 phonemes (8 monophthongs and diphthongs and 4 consonants) that appeared to be problematic for English and Turkish speakers. Neri et al. (2006)'s study aimed to obtain systematic information on segmental pronunciation errors made by learners of Dutch with different mother tongues. They conducted a corpus study on both spontaneous and read speech, and used five criteria to select the segmental pronunciation errors (perceptual salience, frequency, commonality across speakers of various L1s, persistence, potentiality of hampering the communication). This analysis resulted in a list of Dutch phonemes for each specific L1 that are often pronounced incorrectly. They

discovered that the frequency pattern for mispronounced vowels is very similar across the various L1 groups, while the pattern for mispronounced consonants is more heterogeneous. Based on both the list for spontaneous speech and the list for read speech, we selected the phonemes that were relevant for English and Turkish speakers. The selected vowels and consonants and their most frequent false realizations are displayed in Table 4 and 5.

Target	/ə/	/ɑ/	/œy/	/ʏ/	/a/	/øː/	/ɛi/	/y/
Realized as	Deleted [ɪ] [ɛ] [e] [e.]	[a] [a.]	[ʌu] [ɔu] [ɔi]	[u] [y]	[ɑ] [ɑ.]	[y] [o.] [u]	[ei] [ɑi] [ai]	[u]

Table 4: Overview of the selected vowels and their most common false realizations (on the basis of Neri et al., 2006).

Target	/t/	/x/	/w/	/h/
Realized as	Deleted [d]	Deleted [g] [h] [k]	[u] [f]	Deleted [x]

Table 5: Overview of the selected consonants and their most common false realizations (on the basis of Neri et al., 2006).

We analyzed two possible false realizations: (i) the substitution of the target phoneme by a incorrect realization and (ii) the deletion of the target phoneme. In contrast to Neri et al. (2006), we do not consider the cases where a phoneme is inserted.

3.2.3.2. The phonemic error rate

Our goal is to calculate a *phonemic error rate.* This rate is the proportion of incorrect realizations of the selected phonemes divided by the total number of relevant phonemes produced in the speech of the non-native speakers.

First, all words used in the thin slices were extracted into a database and matched with their CELEX[5] phonetic transcription. Then, each occurrence of one of the 12 selected phonemes was marked. If the same phoneme occurred more than once in the same word, only the first occurrence was taken into account. In one single thin slice, no more than ten occurrences of the same phoneme were analyzed. From the eleventh occurrences of the same phoneme (this was the case with the highly frequent phonemes /ə/ and /t/), we stopped analyzing the occurrences of

[5] The CELEX is a lexical database that includes a list of Dutch syllables, their phonetic transcription and their frequencies created by Dutch Centre for Lexical Information.

that specific phoneme. In total, we analyzed 3512 phonemes (thus on average approximately 39 per thin slice).

For each occurrence of a phoneme, the investigator established whether the realization matched the Dutch phoneme (=1) or whether the target phoneme had been substituted by another (=0). The phoneme was considered as incorrect if the produced sound could be categorized as a different phoneme (not allophone) than the target phoneme. All words transferred from other languages and not existing in Dutch (e.g. "ampel", "so", "and", etc.), non-existing words in Dutch and brand names (e.g. Suzuki Swift) were excluded from consideration. For some words, it appeared that the transcription was reconstructive in nature: the annotators had transcribed words/parts of words that were not entirely pronounced by the speaker. In such cases, the missing sounds were thus left out of consideration, since we cannot judge a sound that is not actually pronounced.

During the analysis, we realized that the 12 phonemes selected by Neri et al. (2006) were indeed often problematic for the non-natives, but it was also clear that phonemes other than the ones we selected posed problems. The /ʌu/ as in 'auto', the final /k/ as in 'eigenlijk' and the /ɪ/ as in 'kinderen' are examples of non-selected phonemes that were often mispronounced and that should be taken into consideration in further research.

On the basis of the analysis of all selected phonemes, we calculate the *phonemic error rate* in the following way (see example in (1)). For each thin slice, we divided the number of false realizations of each selected phoneme by the total number of analyzed occurrences of that phoneme in the thin slice (1a). The score for all 12 phonemes was summed up and then divided by the number of phonemes that effectively presented occurrences in the thin slice (1b).

(1) In thin slice A:

a. Number of false realizations of each phoneme:

0 false /ə/ on 10 => 0/10		2 false /t/ on 9	=> 2/9
1 false /ɑ/ on 2	=>1/2	0 false /w/ on 0	=> 0/0
4 false /œʸ/ on 5	=> 4/5	1 false /y/ on 5	=> 1/5
5 false /ʏ/ on 7	=> 5/7	2 false /x/ on 4	=> 2/4
2 false /a/ on 3	=> 2/3		
0 false /øː/ on 0	=> 0/0		
0 false /ɛi/ on 1	=> 0/1		
1 false /y/ on 1	=> 1/1		

b. [(0/10) + (1/2) + (4/5) + (5/7) + (2/3) + (0/1) + (1/1) + (2/9) + (1/5) + (2/4)]/ 10 = .46

=> *phonemic error rate* = .46

As a result, we obtained the *phonemic error rate* for each thin slice. This measure is made up of information about the realization of all twelve phonemes together. Our aim is thus not to consider the correlation between perceived accent and each specific phonemic realization.

3.3. Perception experiment

3.3.1. Design

On the basis of the speech stimuli described in section 3.1., we designed a rating experiment. This experiment aimed to elicit subjective ratings of fluency and foreign accent. We opted for a between-group design where each group rated a different aspect (either fluency or accent). In contrast to Munro and Derwing (1998, 2001), we did not want the same listeners to rate two different aspects in the same thin slice.

3.3.2. Participants

The experiment took place between January 31st 2011 and February 11th 2011. One hundred normal-hearing native Dutch speakers from the UiL-OTS participant pool participated on a voluntary basis and were paid € 7,5 for their contribution. Sixty participants were included for the benefit of a different project (carried out by Hans Rutger Bosker, UiL-OTS, Utrecht). Thus, 40 participants (36f/4m; age 18-34; mean(SD)=20.98(3.08)) participated in the present study. All participants came from the Randstad[6] and considered themselves as having no marked accent in Standard Dutch. Only two participants reported that they could sometimes be perceived as having an Amsterdam accent. Participants were randomly assigned to one of two groups. Since it has been shown that judgments from non-expert native speaker raters are comparable to those obtained from expert raters (Derwing et al., 2004; Rossiter, 2009), our participants were linguistically untrained native speakers with no experience in phonetics, speech therapy or rating of second language proficiency.

3.3.3. Procedure

The experiment consisted of four parts: (i) the instructions, (ii) a practice phase, (iii) a test phase and (iv) a questionnaire. Participants were in sound-attenuating booths. First, written instructions were presented on the screen (the instructions are displayed in Appendix II). Group 1 (who rated *overall fluency*) received instructions to base their judgments on the specific use of silent and filled pauses, the speech rate and the use of hesitations and/or corrections, and not to rate fluency in the broad sense of language proficiency ("He is fluent in French"). Group 2 (who rated *accent*) received instructions to base their judgments on the pronunciation of specific sounds, word stress and intonation patterns. Their ratings should represent how much the pronunciation of the speakers deviates from the norms of Standard Dutch. Thanks to these clear instructions, we

[6] The Randstad is a central urbanized zone in the Netherlands which comprises the major cities as Amsterdam, Rotterdam, The Hague and Utrecht.

prevented listeners from rating the stimuli on global L2 proficiency and forced them to adopt the specific definition of fluency/accent. It was made clear to the participants of both groups that they would hear both native and non-native speakers. Following the instructions but prior to the actual rating experiment, six practice items were given, so that participants could familiarize themselves with the task. They were allowed to ask questions if they did not understand the instructions. No other instructions than the written instructions as presented in Appendix II were supplied to the participants by the experimenters.

In the test phase, the above-described 114 items (90 thin slices of non-native speakers and 24 thin slices of native speakers) were presented to participants using the FEP experiment software (version 2.4.19, Veenker, 2006). Participants listened to the stimuli over headphones at a comfortable volume. They rated the thin slices presented in one of six different pseudo-randomized orders using interval scaling with semantic differentials. Thus, for each item, they were asked to give their ratings by clicking on one of nine stars on a scale with labels at the very left end and the very right end of the scale. Each item appeared in a new window with the question that recalled the instructions given at the beginning of a session at the top of the screen. This design is graphically presented in Figure 7.

GROUP 1: Overall fluency

What is your judgment on the fluency?										
not fluent at all	*	*	*	*	*	*	*	*	*	very fluent

GROUP 2: Accent

What is your judgment on the accent?										
no accent	*	*	*	*	*	*	*	*	*	very strong accent

Figure 7. Schematic representations of the scales presented to participants.

Halfway through the experiment participants were given the opportunity to pause briefly.

Finally, the participants had to fill in a questionnaire through the CLEO server of the University Utrecht (Creating Language Experiments Online). This questionnaire (presented in Appendix III) concerned participants' background, their attitudes towards and degree of familiarity with the speaker's foreign accent. A small debriefing was designed to collect participants' first impressions of what they had been doing during the test phase.

One entire session of the experiment lasted approximately 45 minutes (114 items x 20s = 38min).

Part IV: Results

Part IV: Results

In this chapter, we present and discuss the results of the study and answer our four research questions that unravel the relations between objective and subjective measures of L2 fluency and accent. The chapter starts with section 4.1. in which we report on a series of preliminary analyses that allow us to get a first insight into the data and verify whether the necessary conditions are met. In Section 4.2, the results for each research question are presented.

4.1. Preliminary analyses

Before formulating answers to our research questions, we first performed several preliminary analyses that allow us to gain insights in the results of the experiment and to check whether important conditions are met before reporting on the results. In the first section, we determine the interrater reliability. A high interrater reliability indicates a high degree of agreement between our raters. In the second section, we compute estimates of the rating scores for each thin slice. This operation allowed us to average over all raters and to correct for the possible influences of random factors. Furthermore, we checked whether there are effects of the L1 in both the objective and subjective measures of fluency and accent. We also verified whether the results are normally distributed. In the last section, we computed residuals and checked for multicollinearity.

4.1.1. Interrater reliability

The Cronbach's alpha coefficient was used to measure the degree of interrater reliability for both groups of raters. The Cronbach's α for the 20 raters of the perceived fluency group was .97 and the Cronbach's α for the 20 raters of the perceived accent group was .98. These results indicate very high levels of agreement among the raters of each group.

4.1.2. Estimates

From the rating experiment, we obtained in total 4560 observations: 2280 observations for fluency (114 items * 20 raters) and 2280 observations for accent (114 items * 20 raters). We calculated estimates of these results for each thin slice by computing a *linear mixed model*. One of the many advantages of this statistical method (extensively described in Quené & Van den Berg 2004, 2008) is that it allows us to include multiple random factors in addition to traditional fixed factors. These models are in fact called *mixed*, since they have this particularity of performing analyses with both fixed and random effects. This operation is aiming at calculating adjusted mean ratings to account for the possible influences of random factors.

It is important to mention that the fluency rating scores have been recoded: the rating scale has been reversed to become a scale of dysfluency (1=very fluent and 9=not fluent at all). This conversion made fluency ratings comparable with accent ratings (where a score "9" already meant a highly marked accent) on the one hand, and with the objective measures of both accent and fluency (for which the higher the score, the less fluent/native-like the speech) on the other hand.

In what follows, different models of estimates with different random factors are carried out and, then compared in order to determine which model explains the most variance and corrects the best for random factors.

4.1.2.1. Estimates of fluency ratings

We performed a first basic model (Model 1) with *fluency ratings* as the dependent variable, *items* (the 114 thin slices) as fixed factors and *participants* (20 listeners) as random factors.

In Model 2, we conserved fluency ratings as the dependent variable, items as fixed factors and participants as random factors, but we added *order of items* as a third random factor. Since the items were presented to the listeners in a specific order, we presumed that the fluency ratings evolved in the course of the experiment. An Analysis of Variance was performed in order to compare the two models. The ANOVA revealed that the most complex model, namely Model 2 is significantly better than Model 1 ($\chi^2(1)=5.0582$, $p=.025$) since it accounts for a potential learning effect (listeners got a better insight into the scale and how they had to rate the speakers) and/or fatigue effect (listeners were bored by the experiment or did not pay as much attention to the last items as to the first ones). The model showed that listeners became stricter throughout the experiment.

In Model 3, we added a fourth random factor, namely *order by participants*. In this model, not only did we control for participants and learning/fatigue effects, but also for how these learning/fatigue effects may differ by participant. It is likely that subject A showed a clear learning effect by adapting his ratings throughout the experiment, but that subject B remained quite consistent in his ratings from the beginning till the end of the experiment. The ANOVA showed that the most complex model, namely Model 3, is significantly better than Model 2 ($\chi^2(1) = 10.768$, $p= .001$). Finally, Model 3 also proved to be better than Model 1 ($\chi^2(2)=15.826$, $p< .001$).

Hence, the most complex model, Model 3 with four random factors, was considered the best model. For the further analysis, we used the estimates for each thin slice calculated by the third model.

4.1.2.2. Estimates of accent ratings

In the second step, we modeled the results of accent ratings in the same way as for fluency ratings. We performed a first basic model (Model 1) with *accent ratings* as the dependent variable, *items* as fixed factors and *participants* as random factors. In Model 2, we added *order* as a third random factor. The ANOVA revealed that the most complex model, namely Model 2 is significantly better than Model 1 ($\chi^2(1) = 5.0582$, $p < .025$). In Model 3, we added the fourth random factor, namely *order by participants* in order to correct for differences in leaning/fatigue between participants. The ANOVA showed that the most complex model, namely Model 3 is significantly better than Model 2 ($\chi^2(1)=6.9243$, $p=.009$). Furthermore, Model 3 proved to be better than Model 1 ($\chi^2(2)=49.06$, $p<.001$).

Consequently, the most complex model, Model 3 with four random factors is considered the best model. The estimates for each thin slice obtained with this third model were used for further analyses.

4.1.3. Descriptive statistics

From this point onwards, we conducted all analyses on the basis of the estimates calculated in the previous section (model 3). These estimates per thin slices were better than the average, as we corrected for internal subject variance, for potential learning and fatigue effects and for the individual variance of these learning and fatigue effects.

In the present section, the distribution of fluency and accent scores is described. Furthermore, we compared the fluency and accent subjective and objective measures across the two language groups.

4.1.3.1. Mean results and distribution

The mean of fluency ratings was 5.067 (SD=1.330, range=6.450) and the mean of accent ratings was 6.031 (SD=1.305, range=5.588). We formally tested the normality by performing a Kolmogorov-Smirnov test, comparing the empirical distribution of the ratings to a comparable normal distribution with the mean and standard deviation. For both groups, we could reasonably assume – from the one-sample Kolmogorov-Smirnov test – that the data are normal (fluency ratings: $D= .986$, $p< .001$; accent ratings: $D= .991$, $p< .001$).

4.1.3.2. Objective and subjective fluency between groups

We conducted several analyses of variance in order to determine whether there were differences in objective measures of fluency between our two groups of speakers (the L1 English speakers and the L1 Turkish). For most objective measures[7], there was no effect of language.

For two objective measures, ANOVA's revealed a difference between L1 English and L1 Turkish speakers: mean syllable duration (L1 English: mean(SD)=266(48) vs. L1 Turkish: mean(SD)=294(55)), ($F(1;84)=7.130$, $p=.009$) and mean length of silent pauses (log) (L1 English: mean(SD)=6.75(.32) vs. L1 Turkish: mean(SD)=7.01(.39)), ($F(1;84)= 11.355$, $p= .001$). In both cases, the L1 Turkish speakers are systematically less fluent than the L1 English: their mean syllable duration and their mean length of pauses are significantly longer than those of L1 English speakers.

The ANOVA performed on the fluency ratings between L1 groups (L1 Dutch, L1 English and L1 Turkish) revealed an effect of language group ($F(2 ;105)=29.968$, $p<.001$). Fluency ratings for the three different L1 groups are displayed in Table 6.

	Mean	**SD**	**Range**
L1 Dutch	3.300	1.235	4.469
L1 English	4.531	1.214	6.147
L1 Turkish	5.602	1.233	5.451

Table 6: Fluency ratings (mean, standard deviation and range)
for the three different L1 groups of native and non-native speakers (n=114).

As expected, the native Dutch speakers scored low on fluency (on average approximately 3, meaning quite fluent). There is, however, considerable overlap between fluency ratings of the native and ratings of the two non-native groups. Interestingly, the Dutch natives are not rated as maximally fluent. In fact, only a few items got a 1 (highly fluent) on the scale. This confirms the fact that not all native speakers may be considered very fluent and that the scale does not only apply to non-natives. The Tukey HSD post-hoc test revealed that the natives were rated as more fluent than both L1 English and L1 Turkish speakers. Furthermore, the L1 English speakers were rated as more fluent than the L1 Turkish speakers.

[7] These measures, for which no difference could be shown, are the mean number of pauses between AS-units, the number of silent pauses per minute, the number of filled pauses per minute, the number of corrections per minute and the number of repetitions per minute.

4.1.3.3. Objective and subjective accent between groups

In this section, we investigate the effects of language group on the objective and subjective measures of accent.

As far as the objective measures of accent are concerned, the ANOVA revealed no difference in phonemic error rate between L1 English (mean(SD)= .139(.091), range= .370) and L1 Turkish speakers (mean(SD)=.132(.124), range=.530) (F<1). We concluded that the phonemic error rate was similar across the two groups.

For the accent ratings, the analysis of variance revealed an effect of language group (F(2;105)= 107.277, p<.001). The accent ratings for the three different L1 groups are displayed in Table 7.

	Mean	**SD**	**Range**
L1 Dutch	1.934	.488	1.764
L1 English	5.945	1.333	5.146
L1 Turkish	6.117	1.286	5.588

Table 7: Accent ratings (mean, standard deviation and range)
for the three different L1 groups of native and non-native speakers (n=114).

As expected, the native Dutch speakers scored low on accent (almost no accent, on average 1.934 on the 9-point scale). Raters, however, did not systematically give a score 1 (no accent) to the native speakers. This is a bit surprising, since all native speakers came from the Randstad and spoke Standard Dutch. We expected that they would be categorized as accentless speakers. In contrast, English and Turkish speakers had a comparably rather high score on accent (around 6 on the 9-point scale). The Tukey HSD post-hoc test revealed that the Dutch native group significantly differed from the L1 English and L1 Turkish group, but there was no difference in accent ratings between Turkish and English speakers. In conclusion, we have seen that the participants of group 2 were very good in recognizing Dutch native speakers and rated them significantly differently from the two non-native groups. Not all natives, however, were rated as purely accentless.

In conclusion, we can say that differences in fluency appeared between our two language groups, both in the objective measures and in the ratings. The L1 Turkish group was overall less fluent than the L1 English group. However, no difference in accent was found between these two groups. L1 Turkish and L1 English speakers got similar phonemic error scores and similar accent ratings.

For further analysis of the results, we ignored the results of the eight native speakers included in our experiment, because (i) we focus on fluency and accent in L2 speech and (ii) the native speakers were originally added to the design of the experiment merely as reference items.

4.1.4. Calculating residuals

To test whether L2 fluency residual scores are better predictors of perceived fluency than original L2 measures, we first calculated these residuals. For this calculation, we performed a linear regression for each objective measure of fluency with the L2 measure as the dependent measure and the related L1 measure as a predictor. Each model determined how well a L1 fluency measure (i.e. mean length of pauses) might predict the same fluency parameter in the L2. In all models, the L1 measure was a significant predictor of the L2 measure. The models and the graphs of each linear regression are presented in Appendix IV and V. Residuals were calculated by subtracting the value predicted by the regression line from the L2 measures. In this way, we obtained the new type of measure proposed by Segalowitz (2010) that we will correlate with fluency ratings.

Furthermore, we computed the correlations for each objective measure of fluency between L1 and L2. High correlations indicate that L1 and L2 fluency were closely related and that residuals were rather small (the range is small), whereas low correlations mean that L1 and L2 fluency were quite different from each other and that residuals were rather big (the range is big). For the L1, we took all eight tasks that the speakers performed either in English or Turkish, whereas for the L2 we only considered the ninety 20-second thin slices that we used in the experimental design. The obtained correlations for each measure are presented in Table 8.

Measures in L2	**Pearson's r Correlation with same measures in L1**
MSD	.18
Number of P(/b/ AS)	.28*
Silent P/min	.37*
Filled P/min	.38*
MLP	.44*
Cor/min	.19
Rep/min	.25*

Table 8: Pearson's r correlations between the thin slices (20 sec) and all tasks of the same speakers in his L1 (sig .05 = *).

Overall, the correlations between L1 and L2 are low (between .18 and .38). Only for MLP, is the correlation moderate (.44). In comparison with de Jong et al. (2009) who analyzed the whole dataset from which our data are a subset, our correlations appear to be lower.

If we compare the three aspects of fluency, it can be seen that the highest correlations were found for the measures of breakdown fluency. We expected these stronger correlations between pausing patterns in L1 and in L2, in line with de Jong et al. (*accepted*), who suggested that pausing patterns are strongly related to personal characteristics, and thus possibly directly

transferred from the L1. Such transfer was less visible for the measures of speed and repair fluency.

The correlations of the L1 English speakers and of the L1 Turkish speakers were compared to each other. No significant differences (all Z's < 1) could be found between the correlations of these two groups.

4.1.5. Multicollinearity

In this section, we investigated whether the different objective measures of fluency cluster together. If some measures appear to highly correlate with each other, we run the risk of a multicollinearity problem. This problem appears when highly correlating factors are added together as predictors in a model, and is undesirable in our case since we want to include all objective measures of fluency in the model that predict fluency ratings.

Since utterance fluency is commonly divided into three aspects (speed, breakdown and repair), we expected to find higher correlations between measures within the same aspect than between measures of different aspects. We computed the Pearson's r for all measures and presented the results in Table 9.

	MSD	**Number of P (/b/ AS)**	**Silent P/min**	**Filled P/min**	**MLP**	**Cor/min**	**Rep/min**
MSD	1						
Number of P (/b/ AS)	-.06	1					
Silent P/min	.24	.29	1				
Filled P/min	.21	-.24	-.23	1			
MLP	.10	.03	**-.42**	-.23	1		
Cor/min	.02	.12	.17	-.02	-.16	1	
Rep/min	.25	-.16	.02	.20	-.01	-.03	1

Table 9: Pearson's r correlations between the objective measures of fluency in the L2.

In general, measures were not strongly intercorrelated: most correlations were weak or very weak (less than .25). Only one correlation (between MLP and number of silent pauses) indicated in bold in Table 9 was moderate. The more silent pauses in the speech, the shorter the pauses. This is an interesting correlation that has – as far as we know – not been shown before. It seems that there must be some kind of balance between the length of the silent pauses and their frequency. In other words, a speaker who tends to pause more often will have shorter pauses, whereas someone who holds long pauses needs fewer pauses.

Filled pauses correlated negatively with most other measures. For instance, the more filled pauses in speech, the lower the number of silent pauses. This could indicate that speakers use either mostly silent or mostly filled pauses, and that the use of one type of pauses is related to the use of the other type.

As far as correlations within aspects are concerned, we need to distinguish between breakdown fluency and repair fluency. Within breakdown fluency (in light grey), measures seemed to cluster together to some extent. Most correlations (either negative or positive) reached approximately .25. Within repair fluency however (in dark grey), the two variables (the number of repetitions per minute and the number of corrections per minute) did not correlate.

In general, we can conclude that the relationships between the different measures were low. The different measures did effectively measure different aspects of fluency and therefore the risk of multicollinearity in further analysis of results is limited. De Jong et al. (*accepted*) asked how the different measures of utterance fluency that one traditionally groups into three aspects actually relate in practice. In view of our data, it appeared that the measure of breakdown fluency showed moderate correlations with each other, while correlations within repair fluency did not load together consistently in contrast to what Tavakoli & Skehan (2005) found.

4.2. Analyses

To summarize, we have conducted a range of preliminary analyses and checked several conditions that had to be met before performing the analyses that will allow us to answer our research questions. Firstly, the interrater reliability proved to be very high, indicating a high degree of agreement between our raters. Secondly, estimates of the rating scores were computed for each thin slice in order to average across raters and to correct for possible influences of random factors. The distribution of these estimates and the effect of language group were checked. Thirdly, we computed residuals and checked for multicollinearity.

4.2.1. RQ1 – Which objective measures are the best predictors of L2 fluency ratings?

The first research question concerned the relationship between objective measures of fluency and scores attributed by raters. The questions were whether the ratings of untrained raters (perceived fluency) correlate with temporal objective measures and which specific objective measures can predict perceived fluency.

To assess the predictive power of the objective measures of fluency, we used a multiple linear regression analysis. The goal was to calculate the variance within fluency ratings that could be explained by objective measures of fluency. This amount of variance of fluency ratings is expressed by the *adjusted R squared* (R^2). In regression, the traditional R^2 coefficient of determination is a statistical measure that gives information about the goodness of fit of a model, thus of how well the regression line approximates the real data points. The adjusted R^2 is a modification of the traditional R^2 that adjusts for the number of explanatory terms (predictors) in a model. Unlike R^2, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance.

We first present the models in which objective measures are introduced separately and, in a second step, we report on the models in which measures are grouped by aspects.

4.2.1.1. Objective measures of fluency as individual predictors

First, we included all the objective measures of L2 fluency at the same time as predictors of subjective fluency in a multiple linear regression. The obtained model is reported in Table 10.

Effects	Estimate	SD	T value	P value	Sig.
(Intercept)	-14.86	1.864	-7.970	< .001	***
MSD	.012	.002	7.357	< .001	***
Number of P	.364	.332	1.096	.276	
Silent P/min	.082	.019	4.395	< .001	***
Filled P/min	.017	.011	1.530	.130	
MLP	2.084	.261	7.981	< .001	***
Cor/min	.059	.025	2.332	.022	*
Rep/min	.065	.026	2.502	.014	*

Table 10: Model of fluency ratings with 7 objective measures as predictors[8].

In Table 10, we see that all objective measures except the number of silent pauses between AS-units and the number of filled pauses per minute are significant predictors in the model. The adjusted R squared of the model was .749 ($F(7;80)=38.110$, $p < .001$) which means that 75% of the variance in fluency ratings can be explained on the basis of these seven objective measures of fluency.

In a second step, we computed seven other models with only one objective measure of fluency each. The adjusted R^2 of these models and thus the explanatory force of each measure of fluency is presented in Figure 8.

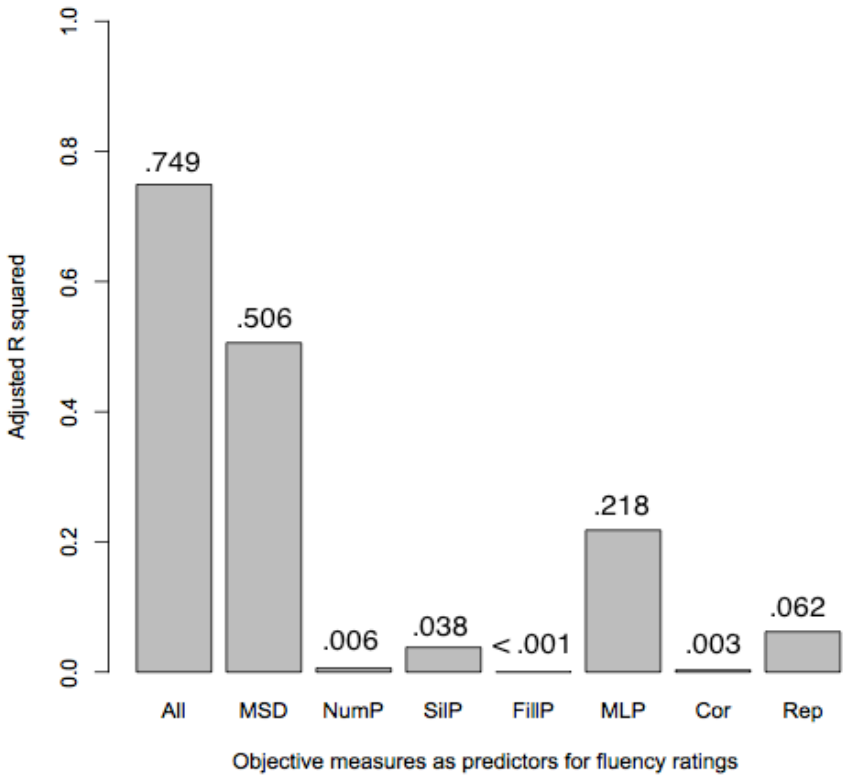


Figure 8: Goodness of fit of the models of subjective fluency with one objective measure of fluency as predictor (expressed in adjusted R^2).

[8] Significance rates: * = .05, ** = .001, *** < .001

It appears that MSD is a good predictor of fluency ratings since it alone accounts for 51% of the variance. The importance of MSD as a predictor has already been tested by Cucchiarini et al. (2002) in spontaneous speech. Cucchiarini et al. (2002) was actually the first study in which pure speed fluency (i.e. MSD or articulation rate) was studied separately from speech rate (i.e. a measure that encompasses aspects of both speed and breakdown fluency). The other studies (e.g. Riggenbach, 1989; Lennon, 1990; Riggenbach, 1991; Freed 1995, 2000; Towell et al., 1996; Kormos & Dénes, 2004) largely concentrated on the analysis of speech rate, which encompasses – as explained previously – aspects of speed as well as pausing. Cucchiarini et al. (2002) found a very weak correlation between articulation rate and fluency ratings, which clearly contrasts with our study.

In general, our measures of breakdown fluency did not seem to be very good explanatory factors. Except for MLP, which explains 22% of the variance, the other breakdown fluency measures had a very small explanatory power. Previous studies have not provided clear-cut results as far as breakdown fluency is concerned. Some researchers have found that the frequency of silent and filled pauses distinguishes fluent and non-fluent speakers (Freed 1995, 2000; Lennon, 1990; Riggenbach, 1989, 1991) and others have got results where the number of pauses and the fluency scores did not correlate (e.g. Rekart & Dunkel,1992; van Gelderen, 1994; Kormos & Dénes, 2004; Derwing et al., 2004). Cucchiarini et al. (2002) found that, for perceived fluency, the frequency of pauses is more relevant than the length and concluded that "less fluent speakers, in general, do not make longer pauses than more fluent speakers, but they do pause more often". In our results however, we found no support for this claim, since the number of silent pauses (SilP) does not explain a large amount of variance of fluency ratings as compared to the length of silent pauses (MLP). Rather, on the basis of our results, we could claim the opposite: less fluent speakers do not in general pause more often than more fluent speakers, but they make longer pauses.

With respect to repair fluency, the number of corrections and repetitions per minute did not appear to explain fluency scores well. This finding is in line with (i) other studies in which perceived fluency was correlated with objective measures (e.g. Cucchiarini et al., 2002; Kormos & Dénes, 2004; Rossiter, 2009 who did not find any strong relationship between fluency ratings and the number of dysfluencies) and (ii) psycholinguistic research. Corley, MacGregor & Donaldson (2007) for instance failed to demonstrate that listeners are affected by repetitions in the ease with which words are integrated into discourse. In contrast to speech containing filled pauses, speech containing repetitions did not show any N400 attenuation effect or a memory effect.

The number of pauses between AS-units and the number of corrections per minute did not appear to correlate well with fluency ratings. We cannot exclude the possibility that this is due to the low correlations found for these measures in section 3.2.2. The correlations for these two measures between the 20-second thin slices and the 2-minute original tasks were lower than for

the other measures. This means that the number of corrections and pauses between AS-units produced in 20 seconds is possibly too low to be representative for longer speech extracts and that we should take the results of these two measures in the linear regression with caution.

Another analysis aimed at testing the importance of each fixed factor in a model is the stepwise linear regression. A stepwise regression starts with the full model (with all objective measures), takes away one factor in each step and checks the effect of this removal on the goodness of fit of the model. If the removal of one factor in the model results in a higher indicator of goodness of fit (AIC[9]), the model has significantly lost some explanatory power. We performed a stepwise regression on our model. It appeared that the only factor whose removal did not significantly affect the goodness of fit was the number of pauses between AS-units. All other factors, when removed from the model caused a significant lowering of the goodness of fit. As a result, we might say that the number of pauses between AS-units is a superfluous measure in the model of perceived fluency.

From the stepwise regression and Figure 8, it is clear that MLP and MSD were the most important measures in the model, since they were the two factors with the biggest explanatory power. In view of this, we can propose an explanation for a fact described in 4.1.3.2. In that section, we showed that there were differences in subjective fluency ratings between L1 English and L1 Turkish speakers. Furthermore, we found that there were also differences in objective measures of fluency between the language groups, namely in MLP and MSD. Since these two measures explain the biggest part of the variance in fluency ratings, it is very likely that raters perceived differences in MLP and MSD and that these differences influenced their judgments.

4.2.1.2. Aspects and combination of aspects of fluency as predictors

Besides the analyses with each measure of fluency included separately, we designed models with measures grouped by aspect and with two aspects combined. The aim of these analyses is to determine which aspect(s) of utterance fluency is the best predictor of fluency ratings.

These models and their explanatory force are presented in Figure 9. The first model is the "full" model with all measures included. The following three models included a combination of two aspects. The last three models included all measures of one single specific aspect of utterance fluency each time.

[9] Akaike Information Criterion (AIC): The lower the AIC, the better the model.

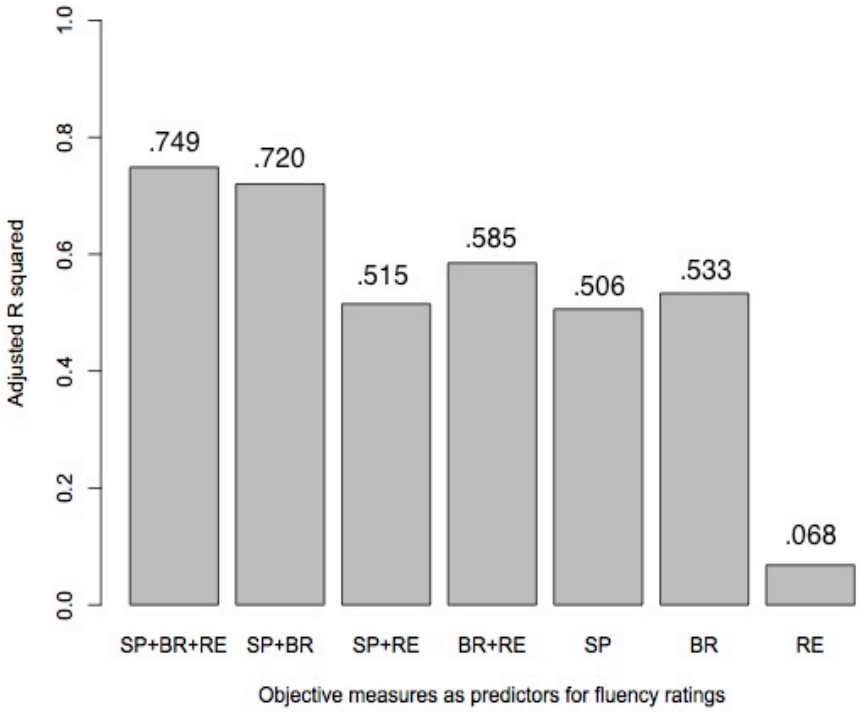


Figure 9: Goodness of fit of the models of subjective fluency with aspect/combination of aspect of fluency as predictor (expressed in adjusted R^2).
SP=speed fluency (MSD)
BR=breakdown fluency (mean number of pauses between AS-units, number of silent pauses per minute, number of filled pauses per minute, MLP)
RE=repair fluency (number of repetitions and corrections per minute).

As shown in the previous section, the measure of speed fluency (MSD) is a good predictor of fluency ratings since it accounted for 51% of the variance, while all measures together can maximally account for 75% of the fluency ratings. The four measures of breakdown fluency together accounted for 53% of the variance, which is about as much as speed fluency. This is very surprising, since the individual measures of breakdown were not strong predictors. When these measures are grouped, however, a bigger part of the variance can be explained. We will come back to this surprising fact and try to formulate an explanation in the next section. The two measures of repair fluency together only accounted for 7% of the variance. As previously explained, the fact that repair fluency measures are weak predictors of fluency ratings is in line with previous studies.

When we grouped different aspects of fluency together, we gained insights into the overlap in explained variance that might exist between these aspects. An ANOVA on the models revealed that the whole model (SP+BR+RE) systematically achieved a significantly higher goodness of fit than all combinations of two aspects (SP+BR, SP+RE and BR+RE). This means that all three aspects (speed, breakdown and repair) are important in the model: no aspect can be omitted.

Furthermore, we tested whether the combinations of two aspects could explain more than the measures of one single aspect. The ANOVA revealed that the combination of speed and breakdown fluency (SP+BR) explained a significantly larger amount of variance (72%) than individual speed of breakdown fluency measures, and was thus the best combination of aspects for predicting fluency ratings. Moreover, the analysis revealed no difference between SP and SP+RE, which means that the addition of repair fluency measures to the speed fluency measure did not significantly affect the goodness of fit. Thus, there was overlap in the variance explained by speed and repair fluency measures. On the other hand, the addition of repair fluency measures to breakdown fluency measures did result in a better model (BR+RE > BR). The overlap between breakdown and repair fluency measures was thus smaller.

4.2.1.3. Alternative frequency measures of dysfluencies

The present section aims to find an explanation for the surprising fact about breakdown fluency described in the previous section. To summarize, we have seen in section 4.2.1.1. that breakdown fluency measures – when added separately into the model – had a very poor explanatory power. Except for MLP that could explain 22% of the variance, the other measures only explained a very small amount (almost negligible) of the variance within fluency ratings. However, in section 4.2.1.2., we grouped these measures in aspects and added all measures of breakdown fluency together in a model predicting perceived fluency. Suddenly, it appeared that all breakdown fluency measures together could account for 53% of the variance, which is much more than the sum of the explained variance of each measure separately.

Our explanation for this fact has to be found in the correlation clusters that exist between length of pauses, number of pauses and perceived fluency. First, we have seen that the more pauses one produces, the less fluent one is perceived. Also, the longer the pauses, the less fluent the speaker is perceived. However, we have found a moderate *negative* correlation between length of pauses (MLP) and number of silent pauses (r=-.42). Thus, the more pauses one produces, the shorter these pauses are. When we added the measures one by one to the model, we ignored this important correlation. As soon as we added information about both the length and the number of pauses, our model became much better than a model with information about merely length or number. Indeed, a model with MLP and the number of silent pauses per minute already predicted 43% of the variance in perceived fluency. This interesting finding points to the fact that listeners are not really capable of considering the length of pauses or the number of pauses separately, but that they included both pieces of information at once in their rating. Listeners did actually not separate these aspects as we did in the objective measures, but rather considered the percentage of time spent pausing (which encompasses information about both the length and the number of pauses).

It is actually not so surprising that the two variables: length of pauses (MLP) and number of pauses correlate negatively in our thin slices. Indeed, both measures are related to time (i.e. the total speech duration: duration 2). Logically, because we have a constant total duration of speech material (e.g. twenty seconds in our thin slice or two minutes for the whole task), the duration and the number of pauses speakers produce are necessarily related to each other. If one produces very long pauses between speech units, the number of these pauses will be small, whereas if one makes many pauses, the duration of these pauses has to be smaller. As explained in this section, this negative correlation poses a problem in our linear regression. Therefore, it appears that we would better use measures of breakdown fluency that are related to the quantity of information given by the speaker instead of related to physical time. As mentioned in section 3.2.2, breakdown fluency measures may also be calculated on the basis of the number of produced syllables (quantity of information) instead of on the basis of duration 2 (i.e. the total speech duration), obtaining the number of silent pauses *per* syllable and the number of filled pauses *per syllable*. We originally did not choose this calculation, because we wanted to stay in line with previous research in which duration 2 has always been used in the computation. Also, we showed that the correlations between the two types of measures were very high. In this section, however, we want to redo the computation with these alternative types of measures per information unit in order to find out whether it influences the results of the regression.

Firstly, the correlation between MSD and the number of pauses per *syllable* was lower (-.10) than the correlation between MSD and the number of pauses per minute (-.42). The small multicollinearity problem present when we added both measures as predictors is thus resolved with the alternative measures. Table 11 presented a comparison of the adjusted R^2 between the traditional type of measure (per time units) as reported in 4.2.1.1. and the alternative type of measures we proposed (per information units).

		Measures /minute	Measures /syllable
Individual measures	**Num of silent pauses**	.038	**.369**
	Num of filled pauses	<.001	**.106**
	Num of corrections	.003	**.128**
	Num of repetitions	.062	**.154**
All	**All measures**	.749	**.777**
Aspect	**Speed**	**.506**	**.506**
	Breakdown	.533	**.696**
	REpair	.068	**.259**
Combination of Aspects	**SP+BR**	.720	**.728**
	SP+RE	.515	**.566**
	BR+RE	.585	**.654**

Table 11: Comparison of goodness of fit of the models of subjective fluency with objective measures of fluency as predictors (either traditional measures per time units or alternative measures per information unit) (expressed in adjusted R^2). Bold=model with highest adjusted R^2.

From Table 11 it is clear that all models computed with the alternative type of objective measures of fluency (*per syllable*) are better than the models with the measures *per minute*. Big differences can be found for breakdown fluency measures: number of silent pauses (4% vs. 37% of the variance in fluency ratings explained) and the number of filled pauses (0% vs. 11%). Consequently, a model with the number of pauses per syllable and MLP as predictors could explain 62% of the fluency ratings instead of 43% with MLP and the number of pauses per minute. The same argument applies to repair fluency measures as well. They appear to better correlate with fluency ratings when they are calculated as a ratio to the number of syllables: the number of corrections (0% vs. 13%) and the number of repetitions (6% vs. 15%). From the new models with the alternative type of measures, it is now clear that breakdown fluency measures are good predictors of fluency ratings, even better than speed fluency. Pausing patterns thus clearly contribute to the perception of a speaker as fluent or non-fluent.

In conclusion, we have highlighted here two important facts that may explain why the results for breakdown fluency as a predictor for perceived fluency were not clear-cut in previous research. *First*, it appeared that the length of pauses on the one hand, and about the number of pauses on the other hand, are – when they are taken apart – weak predictors of fluency ratings. However, when both types of information are added together in the model, they can explain much more variance. We concluded that listeners considered pausing information as a whole:

both the duration and the number of pauses influenced their ratings. *Secondly*, we found that objective measures could better be calculated as a ratio to the number of syllables. These measures are theoretically better than measures per time unit, since they are not by definition already negatively correlated (because we have a fixed total duration of the speech samples). These measures also turned out to be better predictors of the variance of fluency ratings than the measures per time unit. It seems thus that listeners based their judgments on the number of dysfluencies (e.g. pauses, corrections or repetitions) related to the quantity of information (i.e linguistic units a speaker produces; e.g. syllables, words, etc.) and not related to time units. In the following analyses, we therefore used the measures per syllables instead of per minute.

4.2.2. RQ2 – Segalowitz' proposal

Segalowitz (2010) has proposed a new type of utterance fluency measure: the residualized scores (see explanation in sections 2.1.5.2 and 4.1.4). We want to find out to what extent these residualized scores correlate with perceived fluency and whether they are able to predict perceived fluency significantly better than the traditional measures of L2 utterance fluency.

We performed the same regressions as in the previous sections with the residualized scores instead of the L2 measures. The models were compared on the basis of the adjusted R squared. The results are presented in Table 12.

		L2	Residuals	L1 * L2
Individual measures	**MSD**	**.506**	.495	.506
	Number of P	.006	**.012**	0
	Silent P/syl	.369	**.373**	.376
	Filled P/syl	**.106**	.061	.084
	MLP	**.218**	.106	.235
	Cor/syl	**.128**	.084	.152
	Rep/syl	.154	**.161**	.148
All	**SP+BR+RE**	**.777**	.660	.820
Aspects	**Speed**	**.506**	.495	.506
	BReakdown	**.696**	.549	.740
	REpair	**.259**	.223	.267

Table 12: Goodness of fit of the models of subjective fluency with measures of fluency as predictors (either L2 measures, residuals, or the interaction between L1 and L2 measures) (expressed in adjusted R2). Bold=highest goodness of fit for the model.

For most of the models, it appeared that the traditional L2 measures are better predictors than the residuals. The adjusted R^2 was higher for L2 fluency measures, with the exception of three models. These models are models with one single measure as a predictor (the number of pauses between AS-units, the number of silent pauses per syllable and the number of repetitions per syllable), and are marked in bold in Table 13. The amount of variance explained by the models with residuals in these three cases is as high as the variance explained by the models with traditional L2 measures. Unfortunately, we have no way to test whether the adjusted R squared are significantly different. No ANOVA could be computed, since the models have the same number of predictors.

For some measures, it appears thus that the residuals are as good as L2 measures at predicting perceived fluency; but for most of the models, L2 measures still appeared to be better predictors than residuals.

Furthermore, a third type of regression was conducted: regressions with L1 and L2 measures as predictors of fluency ratings. The idea behind this computation is that L1 fluency measures in interaction with L2 measures could influence perceived fluency. In other words, within this third type we investigated whether the L1 can influence the effect of L2 on fluency ratings. The low articulation rate of a speaker in his L1 could, for instance, encourage the listeners not to be too strict when ratings his L2 fluency. The idea of using both L1 and L2 information (in the models with residuals or with L1*L2) rests on the assumption that listeners have some insights in a speaker's L1 fluency characteristics, even though they do not hear the speaker in his L1. In other words, they have some impression about the type of speaker (e.g. someone who speaks very low or someone who produces a lot of filled pauses). As can be seen in Table 13, the models with L1*L2 measures as predictors achieved a goodness of fit that is in most cases slightly superior to the models with traditional L2 measures. However, ANOVA's performed on the two types of models revealed in all cases that there was no difference (all F's <1). As a result, we may say that measures of L1 fluency and their interaction with L2 measures did not add any explanation of the variance.

To conclude, we have performed models with three different types of variables: traditional L2 measures, residuals (a type proposed by Segalowitz, 2010) and the combination of L1 and L2 measures (plus their interaction). Segalowitz claimed that residuals are better objective measures of fluency because they partial out the role of the L1 in L2 speech. We have shown that residuals are not better than traditional L2 measures at predicting perceived fluency. The L1 and L2 measures did not increase the goodness of fit of the models as compared to the models with L2 measure alone. Therefore, we conclude that taking L1 into account (as residuals or as separate predictors) does not result in better predictions of fluency ratings.

4.2.3. RQ3 – Can a phonemic measure of accent predict L2 accent ratings?

In the third research question, our goal was to investigate whether the phonemic error rate can predict perceived accent. To investigate this, we used a linear regression analysis with phonemic error rate and language group as fixed factors. There was no effect of language group and no interaction between phonemic error rate and language group. This means that there was no effect of the language group on the way the phonemic error rate affects the perception of listeners. Raters were thus not influenced by the L1 of a speaker when rating accentedness. The obtained model with phonemic error rate as fixed factor is presented in Table 13.

Effects	**Estimate**	**SD**	**T value**	**P value**	**Sig.**
(Intercept)	5.261	.196	26.829	< .001	***
Phonemic Error rate	5.685	1.133	5.018	< .001	***

Table 13: Model of accent ratings with the phonemic error rate as a predictor.

In Table 13, we see that the phonemic error rate as an objective measure of accent had an effect on accent ratings. The model achieved an adjusted R-squared of .214, which means that 21% of the variance in the accent ratings can be explained for by segmental errors in the speech. This is a rather small amount, but this result is in line with previous research (e.g. Anderson-Hsieh et al., 1992; Magen, 1998). Our segmental measure of accent turned out to be a non-negligible predictor of subjective accent, but a large part of the variance still has to be explained, probably by other complementary measures of accent (e.g. supra-segmental or sub-phonemic measures).

4.2.4. RQ4 – What is the relationship between fluency and accent?

While the first three research questions considered fluency and accent separately, we investigated in our fourth research question the relationship between accent and fluency. The first general question that will be asked in 4.2.4.1. is to what extent fluency and accent ratings are related to each other. In the second step, we investigated whether the objective measure of accent can be a good predictor of fluency ratings and whether, conversely, objective measures of fluency might help to predict accent ratings.

4.2.4.1. Correlation between fluency and accent ratings

Firstly, we investigated to what extent accent ratings and fluency ratings are related to each other. Our hypothesis was that a strong accent might possibly influence fluency ratings: the higher the accent ratings (strong accent), the higher the fluency ratings (not so fluent).

We found that accentedness was significantly correlated to perceived fluency, (r=.25, p=.017). The correlation and the regression line are presented in Figure 10. Indeed, the higher the accent ratings, the higher the fluency ratings: the stronger the perceived accent, the more strongly the speaker is perceived as dysfluent. These results were similar to previous studies, in which the two types of ratings were correlated. Wennerstrom (2000), Derwing & Rossiter (2003) and Derwing et al. (2004) also showed that a weak accent contributes to the overall impression of fluency; and vice versa - that a L2 speaker with a relatively marked accent (as perceived by raters) will be judged more severely on fluency. The correlation we obtained is, however, much weaker than what Derwing et al. (2004) found (r=.49.).

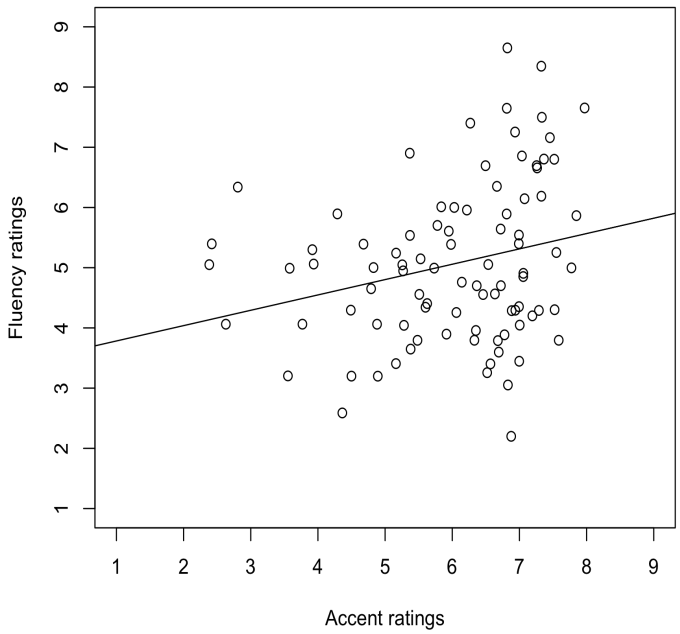


Figure 10: Correlation between accent ratings (x-axe) and fluency ratings of the non-native speakers (y-axe) and regression line (n=90).

4.2.4.2. Does the objective measure of accent add some explanation to fluency ratings?

In the second step, we asked whether an objective measure of accent might be a predictor of fluency ratings. We added the objective measure of accent to the model as a predictor of fluency ratings and checked whether this factor could strengthen the explanatory power of the model. If it appears to be the case, we could conclude that accentedness (in the form of segmental errors) is an interfering factor when a native speaker rates L2 fluency.

To investigate this, we used a linear regression analysis with phonemic error rate and all objective measures of fluency as fixed factors. We also included the factor language group, but there was no effect of this factor, nor any interaction between the objective measures of fluency and language group. This means that there was no effect of the language group on the way these measures affect the perception of listeners. Raters were thus not influenced by the L1 of a speaker in any way

when rating fluency. Therefore, we only reported the model with the other fixed factors. The obtained model is presented in Table 14.

Effects	Estimate	SD	T value	P value	Sig.
(Intercept)	-9.490	1.277	-7.430	< .001	***
MSD	0.006	0.002	2.405	0.018	*
Number of P	0.448	0.319	1.404	0.164	
Silent P/syl	8.521	1.823	4.674	< .001	***
Filled P/syl	2.578	1.552	1.661	0.101	
MLP	1.593	0.196	8.129	< .001	***
Cor/syl	12.584	3.752	3.354	0.001	**
Rep/syl	10.752	3.307	3.251	0.002	**
Phonemic Error rate	-0.167	0.647	-0.258	0.797	

Table 14: Model of fluency ratings with 7 objective measures and the phonemic error rate as predictors.

Like the model presented in section 4.2.1, the number of pauses between AS-units and the number of filled pauses per syllable did not significantly affect the fluency ratings. The phonemic error rate, our objective measure of accent, had no effect either. The model achieved an adjusted R-squared of .775. The model with all objective measures of fluency (with measures per syllable) had an adjusted R-squared of .777. An ANOVA between these two models confirmed that there was no difference ($F<1$). Therefore, we concluded that the phonemic error rate as an objective measure cannot successively predict fluency ratings. Perceived fluency is thus not influenced by the number of segmental errors a speaker may produce.

4.2.4.4. Do objective measures of fluency add some explanation to accent ratings?

Finally, we tested whether objective measures of fluency are predictors of accent ratings. Objective measures of fluency were added to the model presented in Table 15 as predictors of accent ratings. The question was whether these factors could add some explanation of variance, thus whether fluency is an interfering factor when a native speaker rates foreign accent.

We performed linear regressions with phonemic error rate as a fixed factor and one of the measures of fluency separately. Furthermore, a regression with accent error score and all objective measures of fluency as fixed factors was performed. Language group was also added as a fixed factor in all models, but showed no significant effect in any model. Moreover, there was no interaction between the objective measures and language group. This means that there was no effect of the language group on the way these measures affect the perception of listeners. Raters were thus not influenced by the L1 of a speaker in any way when rating accent. The goodness of fit of the models are reported in Table 17 and expressed in adjusted R squared.

Fixed factors		Adusted R^2
Accent	**Measures of fluency**	
Phonemic error rate +	**-**	.21
	MSD	.20
	Number of P (/b/ AS)	.19
	Silent P/syl	.21
	Filled P/syl	.20
	MLP	.21
	Cor/syl	.21
	Rep/syl	.22
	All	.22

Table 15: Goodness of fit of the models of subjective accent with phonemic error rate and measures of fluency as predictors (expressed in adjusted R2).

The basic model presented in Table 13 with the phonemic error rate as predictor could account for 21% of the accent ratings. All the other models with one or more objective measure of fluency as additional predictors did not reach a higher adjusted R-squared. ANOVA's between the basic model and the other models presented in Table 17 did not reveal any difference (all *F*'s<1). Even though previous studies (Munro & Derwing, 1998; Munro & Derwing, 2001) seem to indicate that objective measures of fluency are factors that play a role in perceived accent, we found no support for this finding in our analysis. No objective measure of fluency (added separately or combined) was a significant predictor of accent ratings. This fact casts doubts on the finding of Munro & Derwing (2001) who claimed that speech rate alone could already account for 15% of the accent ratings.

Part V: Discussion and Conclusion

Part V: Discussion and conclusion

In this chapter, we summarize and discuss the findings of the study. Furthermore, we highlight a number of concrete implications of this study and acknowledge the study's limitations.

In this study, a dual approach was adopted (Cucchiarini et al., 2002): fluency and accent ratings were assigned by native speakers to spontaneous speech produced by non-natives, and were compared with a number of objective measures calculated in the same fragments. Our experimental design tried to eliminate, through instruction a broad range of side effects from non-verbal and non-temporal factors, which in previous studies were shown to potentially influence perceived fluency in an uncontrolled way. By matching our selected speakers for proficiency level and having them perform the same tasks, we could collect spontaneous speech with a fair amount of control of grammatical accuracy and lexical richness.

The results of the experiment showed that it is possible to obtain reliable ratings of fluency and accent with untrained native speakers: reliability was very high in both rater groups (Cronbach's α of .97 and .98). This fact may be surprising when compared to the much lower degrees of interrater reliability obtained in previous studies (e.g. Riggenbach, 1991; Freed, 1995). The high reliability coefficients in our study may be a consequence of the instructions. Indeed, the subjects were instructed not to take the proficiency of a speaker into account, but to concentrate on the definition of fluency/accent we provided.

The *first* research question we addressed was to what extent fluency ratings (perceived fluency) are related to temporal measures of speech. The data analyzed provided a range of interesting results. Firstly, the findings of this study indicate that there is a set of variables that are good predictors of fluency scores given by untrained native speakers. These objective measures of fluency are *the articulation rate, the number of silent pauses, the number of filled pauses, the mean length of silent pauses, the number of repetitions* and *the number of corrections*. One of the measures, however, did not significantly help predict fluency ratings (*the number of silent pauses between AS-units*). In comparison with other studies, the measures we used were strictly complementary: all measures assessed one specific aspect of fluency. On the basis of our analysis we argued that the objective measures of fluency that give information about the frequency of a specific dysfluency (e.g. silent or filled pauses, corrections or repetitions) could better be calculated on the basis of information units (i.e. how much speech the speaker effectively produces = the number of syllable) instead of time units. This new type of measures has two main advantages: (i) these objective measures will not correlate by definition with each other (in contrast to measures related to time) so we avoid the problem of multicollinearity when we add these measures as predictors for fluency ratings, and (ii) these measures predict a bigger part of the variance of fluency ratings than the traditional measures (i.e. those related to time units).

Secondly, it appeared that three measures were particularly important for perceived fluency: *the articulation rate*, *the mean length of silent pauses* and the *number of silent pauses*. This finding is in line with Cucchiarini at al. (2002) who conducted a similar experiment with spontaneous speech. In other studies (e.g. Kormos & Dénes, 2004; Derwing et al., 2004), breakdown fluency measures did not correlate with fluency ratings. Thus, evidence for the role of these measures is quite divergent. We proposed an explanation that could account for this contradictory evidence and which is related to the particular correlation clusters between the duration of pauses, the frequency of pauses and perceived fluency. In our data, the correlation between frequency and duration of pauses (calculated as ratio to time units, as in previous studies) clearly indicated that the more pauses a speaker produces, the shorter these pauses are. In our data, it also appeared that the combination of these two types of information (frequency and duration of pauses) explained a larger part of the variance than the measures taken separately. This indicates that listeners cannot separate these two aspects of breakdown fluency and that their perception of breakdown fluency is holistic.

Overall, we found evidence for the fact that breakdown and repair fluency measures affect the perception of fluency. This contrasts to some extent with previous studies (Rekart & Dunkel, 1992; van Gelderen, 1994; Kormos & Dénes, 2004; Derwing et al., 2004; Rossiter, 2009), in which the number of pauses and the number of repairs did not or only weakly correlate with fluency ratings. Our study, however, indicates that fluency as a concept needs to include information about speed, smoothness and repair. In other words, listeners take all three aspects distinguished by Tavakoli & Skehan (2005) (speed, breakdown and repair fluency) into account when providing a judgment on fluency. Nevertheless, our analysis made clear that speed and breakdown fluency played a bigger role than repair fluency. In conclusion, our findings confirmed that a high proportion of the variation in fluency ratings can be explained by a cluster of objective measures of fluency (78%) and therefore fluency is primarily a temporal phenomenon. Indeed, fluency can successfully be captured by *global* measures (i.e. measures that take the duration and frequency of speed, pausing, repair patterns into consideration). It would be interesting to investigate to what extent *local* measures (i.e. measures that would take the position of these dysfluencies in the speech into account) can add some explanatory power to the model (see explanation in section 5.2.).

In our *second* question, we focus on the role of L1 in L2 fluency. As explained, our experimental design was meant to eliminate a range of side effects in order to isolate the phenomenon of fluency that is purely related to the use of an L2. Following this reasoning, we also eliminated the differences in fluency that were strictly related to personal characteristics by taking L1 fluency measures into account (in the calculation of residual scores). The second research question was to what extent residualized scores as alternative types of utterance fluency measures proposed by Segalowitz (2010) correlate with perceived fluency and whether they could possibly be better predictors than traditional L2 measures. The analyses showed that residuals could predict a non-

negligible part of the variance in fluency ratings. However, the part was smaller than the part explained by traditional L2 measures. Residuals are thus presumably better utterance fluency measures, since they allow us to partial out the role of L1 in L2 speech, but the hypothesis that these residuals would also be better at predicting perceived fluency than pure L2 measures cannot be maintained on the basis of our results.

With respect to the *third* major objective of this study, our results confirmed the findings of previous research. Our goal was to relate accent ratings to a measure of segmental errors. This measure, *the phonemic error rate*, was based on a selection of 12 Dutch phonemes that have been shown to be difficult for English and Turkish learners of Dutch (Neri et al., 2006). This segmental measure could account for 21% of the variance within accent ratings. The number of segmental errors can thus predict a small part of variance in accent ratings, as shown by Anderson-Hsieh et al., 1992; Magen 1998. Nevertheless, it is clear that perceived accent is much more than segmental errors only.

In our *fourth* research question, fluency and accent were investigated in relation to each other. First, the relationship between fluency ratings and accent ratings was investigated. The tested hypothesis was that a strong accent could lead listeners to evaluate the speakers as less fluent. We found a weak but significant correlation between fluency and accent ratings (r=.25), showing that these ratings are weakly related. In the first analyses, our findings offer valuable insights into the objective measures of accent and fluency that can respectively predict accent and fluency ratings. Our temporal measures of fluency could predict 78% of the variance in fluency ratings and our segmental measures of accent could predict 21% of the variance in accent ratings. Since fluency and accent ratings are weakly related, the question that emerges is whether accent could explain some of the remaining 22% of fluency ratings, and similarly, could fluency explain some of the remaining variance in accent ratings? In both cases, the results showed that this was not the case. Neither objective measures of fluency nor the objective measure of accent could add any explanatory power to accent and fluency ratings respectively. Consequently, listeners do not seem to pay attention to accent (at least to segmental accent errors) when they are rating fluency, and they do not seem to be influenced by temporal factors of fluency when rating accent. Fluency and accent are clearly different aspects of L2 speech, and listeners are perfectly able to distinguish between the two. These findings are to some extent in contradiction with previous studies. Munro & Derwing (2001), for instance, showed experimentally that fluency could help to predict perceived accent. The different outcomes between our study and their work can partly be explained by the instructions given to the raters in the experiment. Munro & Derwing (2001) did not give any specific instructions to the raters. They played the samples from a tape in a classroom and asked the participants "to rate comprehensibility and accentedness on two 9-point scales", whereas in our study participants received extensive instructions on the computer screen in which fluency/accent was defined. Therefore, it is possible that listeners

focused intensively on the aspects described in the instructions, so that fluency or accent in our study did not interfere in their ratings. Moreover, Wennerstrom (2000) and Derwing & Rossiter (2003) could predict perceived fluency with measures of accent. In these two studies, however, the measures of accent were measures of prosodic accuracy, whereas our analysis was focused on segmental errors in speech.

5.1. Implications

It is clear that language testing practices could benefit from this type of research for two main reasons. First, instructions to human raters on how to rate fluency could be improved on the basis of the present research. It could include more information concerning the specific aspects of utterance fluency are related to the ease and smoothness of L2 linguistic processing. Second, automatic fluency assessing software could also be developed and improved, since we now have better insights into which factors contribute to the perception of fluency and how well each measure may predict human ratings.

We have seen that our untrained raters (when instructed) are able to provide consistent and reliable judgments on fluency and accent, as the interrater reliability was very high. Thus naïve listeners are all able to provide judgment on L2 speech so that the use of trained listeners such as phoneticians and speech therapists seems superfluous.

An important insight for both language testing and language teaching is that we have shown that fluency and accent are different aspects of L2 speech, untrained raters are perfectly capable of distinguishing between the two aspects. Consequently, they should be assessed separately.

As far as *fluency* is concerned, we have shown that it can indeed be defined in its narrow sense by "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under temporal constraints of on-line processing" (Lennon 2000: 26). Segalowitz (2010) has claimed that residuals are theoretically better measures of specific L2 fluency. We do not deny this, but cast doubt on the idea that they could be good measures for assessing L2 fluency, since they correlate less with perceived fluency than traditional L2 temporal measures. In our data, one of the temporal measures, the *mean length of silent pauses (MLP),* turned out to be quite a good predictor of perceived fluency. However, De Jong et al. (*accepted*) who used exactly the same measure, found that the MLP is only very weakly related to language proficiency. Thus it seems that a strong relation between a measure of utterance fluency and perceived fluency does not imply a strong relation between this same measure and cognitive fluency.

When it comes to *accent*, we have shown that perceived accent is partly based in segmental errors. Since segmental errors in the speech could only explain a limited amount of variance, it is clear that future software aimed at automatically evaluating foreign accent should take more parameters into account than only segmental errors in the speech.

5.2. Limitations of the present study and suggestions for further research

Our study is correlational by nature. We correlated measures of utterance fluency with perceived fluency, a segmental measure of accent with perceived accent, and subjective and objective measures of accent and fluency with each other. Caution should be taken in this type of study with the interpretation of results. Even if accent ratings, for instance, turn out to significantly relate to perceived fluency, our experimental design does not allow us to make firm statements on causality relationships nor on the direction of relationship. It would be too risky to claim that accent has an impact on fluency ratings on the basis of this correlation or, the other way around; high fluency leads to a perception of reduced accent. Consequently, we are not able to make strong claims about what causes variation in the ratings, but our study gives important insights into how aspects are related to each other and how much variability in one aspect may be explained by the variation in another aspects. This study should thus be considered as a starting point for further research that could, in a controlled way, assess the consequence of variation in one specific aspect on the perception of fluency or accent.

In our study, perceived fluency was correlated with objective measures of fluency and accent. Our temporal measures of fluency could predict 77% of the variance in fluency ratings. Segmental accent errors produced by L2 speakers could not account for the remaining variance. Consequently, there is still 23% of the variation that remains unexplained. It is possible that listeners are influenced by local measures of fluency (see explanation in the next paragraph), by other factors of a linguistic nature (e.g. lexical choice, grammatical errors, etc.) or even by non-linguistic factors (e.g. quality of the voice, self confidence, etc.). As already stated by Rossiter (2009), further research on non-temporal factors that affect the perception of fluency is recommended.

A limitation of the present study is that we have only considered the temporal phenomena of fluency (e.g. the pauses) globally and not locally. However - as Lehtonen (1978: 67) pointed out,

> fluency does not always imply an uninterrupted flow of speech which is grammatically perfectly irreproachable. To be fluent in the right way, one has to know how to hesitate, how to be silent, how to self-correct; how to interrupt and how to complete one's expression, and how to do all this fluently, in a way that is expected by the linguistic community and that represents normal, acceptable and relaxed linguistic behavior.

From pausological research for instance, it appears that pauses are difficult to interpret. Indeed, the presence of pauses does not always indicate limited fluency. According to Chafe (1980), pauses may reflect either time required to focus on a new thought (conceptualization) or time required to put this new thought into words (formulation). Thus, the function and placement of pauses is as important as their frequency. In fluent-sounding speech, pauses typically occur at predictable places; there are so-called *juncture pauses* (Hawkins, 1971) and appear at clause boundaries. In contrast, dysfluent-sounding pauses occur within clauses and tend to disturb the impression of smoothly flowing speech (Hawkins, 1971). Despite this fact, the present study focused exclusively on quantitatively measurable features.

In order to investigate fluency phenomena locally, a much more discourse-oriented approach (such as Pawley (2000), for instance) is required. Research on cognitive fluency could highly benefit from the combination of these *global* and *local* insights. Indeed, it would make it possible to define specific types and locations of dysfluencies and investigate their sources within the processing system.

As far as accent is concerned, our segmental measures of accent could predict 21% of the variance in accent ratings. Temporal measures of fluency could not account for the rest of the variance. More measures of accent are definitively required in order to check how much of the variance within accent ratings can be explained by measures purely related to accent. Both sub- and supra-segmental measures should be considered in further investigation. However, the risk with supra-segmental measures that fluency and accent overlap at some point should be taken into account. For instance, considering lexical stress (which is a supra-segmental aspect of speech) requires that one look at, among other things, the duration of vowels, and vowel duration necessarily has consequences on speed fluency. In order to improve our measures of accent, future similar studies on Dutch could strongly benefit from the project called *Transcription Quality Evaluation* (TQE)[10] coordinated by dr. Helmer Strik aimed at developing automatic foreign accent evaluation. TQE is designed to assess foreign accent in Dutch by means of automatic phoneme recognition and could provide us with scalar data of how well a non-native realization fits a target Dutch phoneme.

This study used data from English and Turkish learners with an intermediate proficiency level in Dutch. However, previous research has clearly shown that fluency is not a static aspect which an L2 speaker displays or not. Fluency is dynamic and evolves with the proficiency level (Towell et al., 1996; O'Brien et al., 2007). It would be interesting to investigate whether the relationships we highlighted in this study change over time. We found, for instance, a negative correlation between the frequency and the duration of pauses. The question is whether a comparable correlation could be found in low proficiency speakers and in speakers who have achieved a near-native level.

In our experimental design, we included two different language groups. Even though we matched our two language groups for proficiency level, the analysis revealed differences in objective measures of fluency between L1 English and L1 Turkish. We did not extensively discuss these differences, but the possibility exists that they are actually due to cross-linguistic differences in the L1 (i.e. L1 Turkish speakers may produce longer silent pauses by nature, even in their mother tongue). Riazantseva (2001) investigate this possibility by comparing three measures of breakdown fluency in the speech of Russian L2 speakers of English and native English speakers. For one of the measures (pausing duration), a cross-linguistic difference was found: the pause durations in L1 Russian were on average longer than in L1 English. Riazantseva concluded that (i) pausing duration is a language-specific feature and (ii) initial transfer of these language-specific pausing patterns can be overcome with increased proficiency in the L2. Derwing et al. (2009) did not replicate this finding in their study

[10] http://lands.let.ru.nl/~strik/research/TQE.html

comparing Mandarin and Slavic speakers. However, it may be difficult to directly compare the results of these two studies because firstly, in the study by Derwing et al. (2009) the participants had lower proficiency level and secondly, in Derwing et al. (2009) only pauses longer than 400 ms were considered, which is much higher than Riazantseva´s (2001) cut-off of 100 ms. In this type of study, finding out whether there actually are cross-linguistic differences would definitely be required. In our own data, a deeper analysis of the L1 characteristics is required in further research.

In conclusion, we have shown how objective measures of fluency and accent relate to the perception by human listeners of fluency and accent in the speech of non-native speakers. In this way, this study provides new insights into our knowledge of the L2 phenomena of fluency and accent, and in the way these concepts are related to each other.

References

AAN DE WIEL, M., VAN DEN BRINK, G. & S. STRUIJK VAN BERGEN (1991). Diagnostiek van uitspraakproblemen van tweedetaalverwervers van het Nederlands. In: *Levende Talen* (466), p. 507–511.

ABERCROMBIE, D. (1956). Teaching pronunciation. In: ABERCROMBIE, D. (ed.*). Problems and principles: studies in the teaching of English as a second language*. London: Longmans, p. 28–40. Reprinted in Brown (1991), p. 89–95.

AMBADY, N., BERNIERI, F. & RICHESON, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral Stream. In: *Advances in Experimental Social Psychology* (32), p. 201-272.

ANDERSON-HSIEH, J. & K. KOEHLER (1988). The effect of foreign accent and speaking rate on native speaker comprehension. In: *Language Learning* (38), p. 561–613.

ANDERSON-HSIEH, J., R. JOHNSON & K. KOEHLER (1992). The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. In: *Language Learning* (42), p. 529–555.

BHAT, S., HASEGAWA-JOHNSON, M. & R. SPROAT (2010). Automatic Fluency Assessment by Signal-Level Measurement of Spontaneous Speech. *2010 INTERSPEECH Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Makuhari, Japan.

BEGLAR, D. & HUNT, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. In: *Language Testing* 16 (2), p. 131-162.

BOND Z., STOCKMAL V. & D. MARKUS (2008). A note on native and non-native accentedness judgments, *Ohio University Working Papers in Applied Linguistics, www.ohiou.edu/linguistics/workingpapers/2008/ouwpal_2008.html.*

BOSKER, H. R., QUENÉ, H., SANDERS, T., PINGET, A. & N. DE JONG (in prep). *What do oral-fluency raters evaluate? The effect of instructions on fluency ratings.*

DE BOT, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. In: *Applied Linguistics* (13), p. 1–24.

BRENNAN, E. M., & BRENNAN, J. S. (1981). Measurements of accent and attitude toward Mexican-American speech. In: *Journal of Psycholinguistic Research* (10), p. 487-501.

CHAFE, W. (1980). Some reasons for hesitating. In: DECHERT, H. & M. RAUPACH (eds.), *Temporal Variables in Speech*, The Hague: Mouton, p. 169-180.

CHAMBERS, F. (1997). What do we mean by fluency? In: *System* (25), p. 535–544.

COLLINS, B. & I. M. MEES (2003). *The phonetics of English and Dutch*, Leiden: Brill.

CORLEY, M., MACGREGOR, L. J., & D. I. DONALDSON (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. In: *Cognition* (105), p. 658–668.

CRYSTAL, T. & HOUSE, A. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. In: *The Journal of the Acoustical Society of America* (88), p. 101-112.

CUCCHIARINI, C., STRIK, H., & BOVES, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. In: *Journal of the Acoustical Society of America* (111), p. 2862–2873.

CUTLER, A. (1984). Stress and accent in language production and understanding. In: GIBBON, D. & H. RICHTER (eds.), *Intonation, Accent and Rhythm. Studies in Discourse Phonology*, 77–90. Berlin/New York: de Gruyter.

CUTLER, A. & S. BUTTERFIELD (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. In: *Journal of Memory and Language* (31), p. 218-236.

DERWING, T. & ROSSITER, M. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. In: *Applied Language Learning* (13), p. 1-18.

DERWING, T., ROSSITER, M., MUNRO, M. & THOMSON, R. (2004). L2 fluency: Judgments on different tasks. In: *Language Learning* 54, p. 655-679.

DERWING, T., MUNRO, M., THOMSON, R. & ROSSITER, M. (2009). The relationship between L1 fluency and L2 fluency development. In: *Studies in Second Language Acquisition* 31(4), p. 533-557.

VAN DEN DOEL, R. (2006). *How Friendly are the Natives? An Evaluation of Native-speaker Judgements of Foreign-accented British and American English*. LOT Doctoral dissertation, University Utrecht.

EISENSTEIN, M. (1983). Native reactions to non-native speech: a review of empirical research. In: *Studies in Second Language Acquisition* (5), p. 160–176.

FLEGE, J. (1988). Using visual information to train foreign-language vowel production. In: *Language Learning* (38), p. 365–407.

FLEGE, J., M. MUNRO & R. MACKAY (1995). Factors affecting strength of perceived foreign accent in a second language. In: *Journal of the Acoustical Society of America* (97), p. 3125–3134.

FLEGE, J., E. FRIEDA & T. NOZAWA (1997). Amount of native-language (L1) use affects the pronunciation of an L2. In: *Journal of Phonetics* (25), p.169-186.

FOSTER, P., TONKYN & G. WIGGLESWORTH (2001). Measuring spoken language: a unit for all reasons. In: *Applied Linguistics* 21 (3), p. 354-375.

FREED, B. (1995). What makes us think that students who study abroad become fluent? In: FREED, B. (Ed.), *Second language acquisition in a study abroad context*. John Benjamins, Amsterdam, p. 123-148.

FREED, B. (2000). Is fluency, like beauty, the eyes, of the beholder? In: RIGGENBACH, H. (Ed.), *Perspectives on fluency*. The University of Michigan Press, Michigan, p. 243-265.

GALLARDO DEL PUERTO, F., GÓMEZ LACABEX, E. & M. GARCÍA LECUMBERRI (2007). *The assessment of foreign accent by native and non-native judges.* Paper presented at the Phonetics Teaching & Learning Conference, UCL, August 24-26, 2007.

VAN GELDEREN, A. (1994). Prediction of global ratings of fluency and delivery innarrative discourse by linguistics and phonetic measures – oral performance of students aged 11–12 years. In: *Language Testing*, (11), p. 291–319.

GROSJEAN, F. (1980). Temporal variables within and between languages. In: H. DECHERT & M. RAUCHPACH (eds.). *Towards a Cross-Linguistic Assessment of Speech Production*. Frankfurt, P. Lang, p. 39-53.

GUILLOT, M. (1999). *Fluency and its teaching, Multilingual Matters*, Clevedon, England.

GUSSENHOVEN, C. (1999). Illustrations of the IPA: Dutch. In: *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press. p. 74-77.

HAWKINS, P. (1971). The syntactic location of hesitation pauses. In: *Language and Speech* (14), p. 277-288.

HULSTIJN, J.H. (2010). Measuring second language proficiency. In: E. BLOM & S. UNSWORTH (Eds.), *Experimental methods in language acquisition research* (EMLAR). Amsterdam: Benjamins, p. 185-199.

INGRAM, J. & PITTAM, J. (1987) Auditory and acoustic correlates of perceived accent change: Vietnamese school children acquiring Australian English. In: *Journal of Phonetics* (15), p. 127-145.

IWASHITA, N., BROWN, A., MCNAMARA, T. & O'HAGAN S. (2008) Assessed Levels of Second Language Speaking Proficiency: How Distinct? In: *Applied Linguistics* 29(1), p. 24-49.

JAKOBSON, R. (1960). Closing statements: Linguistics and Poetics. In: SEBEOK, T. (ed.) *Style in Language*, New-York.

JOHANSSON, S. (1973). The identification and evaluation of errors in foreign languages: a functional approach. In: J. SVARTVIK (ed.) Errata: papers in error analysis. Lund: Gleerup. p. 102–114.

JOHANSSON, S. (1975). *Papers in contrastive linguistics and language testing*. Lund: Gleerup.

JOHANSSON, S. (1978). *Studies in error gravity: native reactions to errors produced by Swedish learners of English*. Gothenburg: Acta Universitatis Gothoburgensis.

DE JONG, N., SCHOONER, R. & J. HULSTIJN (2009). *Fluency in L2 is related to fluency in L1*. Paper presented at the Seventh International Symposium on Bilingualism (ISB7), Utrecht, The Netherlands.

DE JONG, N., STEINEL, M., FLORIJN, A., SCHOONEN, R. & J. HULSTIJN (*accepted*). *Linguistic skills and speaking fluency in a second language*. In: *Language Testing*.

KORMOS, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Erlbaum.

KORMOS, J. & M. DÉNES (2004). Exploring measures and perceptions of fluency in the speech of second language learners. In: *System* (32), p. 145-164

LEATHER, J. (1999). Second language speech research: an introduction. In: J. LEATHER (ed.). *Phonological issues in language learning*. Oxford: Blackwell. p. 1–58.

LEHTONEN, J. (1978). On the problems of measuring fluency. In: M. LEIWO & A. RÄSÄNEN (eds.). *AFinLA: year book 1978*, p. 53-68.

LENNON, P. (1990). Investigating fluency in EFL: A quantitative approach. In: *Language Learning* (40), p. 387-412.

LENNON, P. (2000). The lexical element in spoken second language fluency. In: RIGGENBACH, H. (Ed.), *Perspectives on fluency*. The University of Michigan Press, Michigan, p. 25-42.

LEVELT, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

LINDBLOM, B. (1984). Can the models of evolutionary biology be applied to phonetics problems? ln: M. P. VAN DER BROECKE & A. COHEN (eds), Proceedings of the Tenth International Congress of Phalletic Sciences.Dordrecht: Foris.

MAGEN, I. (1998). The perception of foreign-accented speech. In: *Journal of Phonetic*s (26), p. 381-400.

MAJOR, R. C. (1987) English voiceless stop production by speakers of Brazilian Portuguese. In: *Journal of Phonetics* (15), p. 197-202.

MAJOR, R., S. F. FITZMAURICE, F. BUNTA & C. BALASUBRAMANIAN (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. In: *Language Learning* (55), p. 37-69.

MOYER, A. (1999). Ultimate attainment in L2 phonology: the critical factors of age, motivation and instruction. In: *Studies in Second Language Acquisition* (21), p. 81-108.

MUNRO, M. J. & T. M. DERWING (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. In: *Language Learning* (45), p. 73-97.

MUNRO, M. J. & DERWING, T.M. (1998). The effects of speech rate on the comprehensibility of native and foreign accented speech. In: *Language Learning* (48), p. 159-182.

MUNRO, M. J. & DERWING, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. In: *Language Learning* (49), Supplement 1, p. 285-310.

MUNRO, M. J. & DERWING, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: accentedness and comprehensibility of L2 speech: The role of speaking rate. In: *Studies in Second Languague Acquisition* (23), p. 451-468.

MUNRO, M,J. & DERWING, T.M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. In: *Language Learning* (58), p. 479-502.

NERI, A., C. CUCCHIARINI & H. STRIK (2006). Selecting segmental errors in L2 Dutch for optimal pronunciation training. In: *IRAL - International Review of Applied Linguistics* (44), p. 357-404.

O'BRIEN, I., SEGALOWITZ, N., FREED, B. & COLLENTINE, J. (2007). Phonological memory predicts second language oral fluency gains in adults. In: *Studies in Second Language Acquisition* 29(4), p. 557-582.

OYAMA, S. (1982). The sensitive period and comprehension of speech. In: KRASHEN, S., SCARCELL, R. & LONG, M. (Eds.), *Issues in second language research*. London: Newbury House, p. 39-51

PATKOWSKI, M. (1990). Age and accent in a second language: A reply to James Emil Flege. In: *Applied Linguistic* II, p. 73-89.

PAWLEY, A. (2000). The one clause at a time hypothesis. In: RIGGENBACH, H. (ed.), *Perspectives on Fluency*, Ann Arbor: University of Michigan Press, p. 163-199.

QUENÉ, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In: *Journal of the Acoustical Society of America* 123(2), p. 1104-1113.

QUENÉ, H. & H. VAN DEN BERGH (2004). On Multi-Level Modeling of data from repeated measures designs: A tutorial. In: *Speech Communication*, 43 (1-2), p.103-121.

QUENÉ, H. & H. VAN DEN BERGH (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. In: *Journal of Memory and Language*, (59), p. 413-425.

REKART, D., & DUNKEL, P. (1992). The Utility of Objective (Computer) Measures of the Fluency of English as a Second Language. In: *Applied Language Learning* (3), p. 65-85.

RIAZANTSEVA, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. In: *Studies in Second Language Acquisition* 23(4), p. 497-526.

RIGGENBACH, H. (1989). *Nonnative fluency in dialogue verse monologue speech: A microanalytic approach.* Unpublished doctoral dissertation, University of California, Los Angeles.

RIGGENBACH, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. In: *Discourse Processes* (14), p. 423-441.

ROSSITER, M. (2009). Perceptions of L2 fluency by native and non-native speakers of English. In: *The Canadian Modern Language Review*, 65 (3), p. 395-412.

RYAN, E. B., M.A. CARRANZA & R. W. MOFFIE (1977). Mexican American reactions to accented English. In: J. W. BERRY & W.J. LOANER (eds.). *Applied cross-cultural psychology*. Amsterdam: Swets & Zeitlinger, p. 174–178.

SAJAVAARA, K. (1987). Second Language speech production: Factors affecting fluency. In: DECHERT, H. & M. RAUPACH (eds.). *Psycholinguistic Model of production*, p. 45-65.

SCHEUER, S. (2005). Why native speakers are (still) relevant. In: DZIUBALSKA- KOŁACZYK & PRZEDLACKA (Eds.). *English pronunciation models: A changing scene*. Berne: Peter Lang. p. 111–130.

SCOVEL, T. (1988). *A Time to speak: A Psycholinguistic Inquiry into the Critical Period for Human Speech*, Newbury House/ Harper Row, New York.

SEGALOWITZ, N. (2004). *Real-time cognitive processing efficiency and second language fluency acquisition*. Paper presented at the annual meeting of the American Association of Applied Linguistics, Portland, OR.

SEGALOWITZ, N. (2010). *Cognitive Bases of Second Language Fluency*. Routlegde: London/New York.

SEGALOWITZ, N. & FREED, B. F. (2004). Context, contact and cognition in oral fluency acquisition: Learning Spanish in At Home and Study Abroad contexts. In: *Studies in Second Language Acquisition* (26), p. 173-199.

SMITH, R. (2005). *The role of fine phonetic detail in word segmentation*. Unpublished Doctoral Dissertation, University of Cambridge, UK.

TAHTA, S., WOOD, M. & LOWENTHAL, K. (1981) Foreign accents: factors relating to transfer of accent from the first language to the second language. In: *Language & Speech* (24), p. 265–272

TAVAKOLI, P. & SKEHAN, P. (2005). Strategic planning, task structure and performance testing. In: ELLIS, R. (ed.). *Planning and task performance in a second language Amsterdam*, John Benjamins, p. 239-273.

TOWELL, R. (2002). Relative degrees of fluency: a comparative case study of advanced learners of French. In: *International Review of Applied Linguistics in Language Teaching* 40 (2), p. 117-150.

TOWELL, R., HAWKINS, R. & N. BAZERGUI (1996). The development of fluency in advanced learners of French. In: *Applied Linguistics* (17), p. 84-119.

TRASK, R. (1996). *A dictionary of phonetics and phonology*. London: Routledge.

TROFIMOVICH, P., & BAKER, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. In: *Studies in Second Language Acquisition* (28), p. 1-30.

VEENKER, T.J.G. (2006). *FEP: A tool for designing and running computerized experiments* (version 2.4.19) [UiL-OTS computer program].

WENNERSTROM, A. (2000). The role of intonation in second language fluency. In: RIGGENBACH, H. (Ed.), *Perspectives on fluency*. The University of Michigan Press, Michigan, p. 102-127.

Appendices

Appendix I: Tasks performed by the speakers in their L1 and L2

	Task characteristics	**Description**
Task 1	simple, informal, descriptive	The participant speaks on the phone to a friend, describing the apartment of friends who have recently moved house.
Task 2	**simple, formal, descriptive**	**The participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened**
Task 3	simple, informal, argumentative	The participant advises his/her sister on how to choose between (or combine) child care, further education, and paid work.
Task 4	**simple, formal, argumentative**	**The participant is present at a neighborhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.**
Task 5	complex, formal, descriptive	The participant tells a friend about the development of unemployment among women and men over the last ten years.
Task 6	complex, informal, argumentative	The participant discusses the pros and cons of three means of transportation (public transportation, bicycle, automobile) on how to solve the problem of traffic congestions.
Task 7	complex, formal, descriptive	The participant works at the employment office of a hospital and tells a candidate for a nurse position what the main tasks in the vacant position are.
Task 8	**complex, formal, argumentative**	**The participant, who is the manager of a supermarket, addresses a neighborhood meeting and argues which one of three alternative plans for building a car park he/she prefers.**

Table a. Descriptions of the tasks performed by the speakers in their L2 (Dutch) (the three selected tasks are marked in bold).

	Task characteristics	Description
Task 1	simple, informal, descriptive	The participant speaks with a friend and describes the type of apartment he is looking for.
Task 2	simple, formal, descriptive	The participant who has just seen a crime/accident occurring on the street, describes to the police officer what had happened.
Task 3	simple, informal, argumentative	The participant advises his/her brother on how to choose between quitting his current job and dive in a new domain, or remaining at his current job while studying part-time for his new career.
Task 4	simple, formal, argumentative	The participant is present at a neighborhood meeting in which an official has just proposed to build a new casino at a location near a school. Participant gets up to speak, and suggests another location that would be more acceptable
Task 5	complex, finormal, descriptive	The participant tells a friend about a piece in the newspaper about home sales in rural vs. suburban areas.
Task 6	complex, informal, argumentative	The participant is the principal of a high school and calls a new science teacher to tell him about the courses to be taught.
Task 7	complex, formal, descriptive	After watching a movie about global warming, the participant discusses the problem with a friend and tries to convince him that more solar/wind energy is the best solution.
Task 8	complex, formal, argumentative	The participant, who is the manager of an elderly home, addresses a Board of Directors' meeting and discusses the advantages and disadvantages of building more facilities.

Table b. Descriptions of the tasks performed by the speakers in their L1 (English or Turkish).

Appendix II

Instructions – group of fluency raters

Welkom bij dit luisterexperiment!

In dit experiment kijken we naar hoe vloeiend bepaalde sprekers Nederlands spreken.Met vloeiendheid bedoelen we NIET hoe goed iemand in een taal is (''Hij spreekt vloeiend Frans''), maar eerder hoe soepel het proces van spreken verloopt. Iemand kan dus veel grammaticale fouten in het Nederlands maken maar wel soepel en dus vloeiend spreken.

Jouw taak is om spraakfragmenten te beluisteren en te beoordelen.
Baseer je oordeel telkens op:
- het gebruik van pauzes. In spraak komen twee typen pauzes voor: stille en gevulde pauzes. Stille pauzes zijn stiltes in de spraak (kort of lang) en gevulde pauzes zijn \uhm\'s\ en \uh\'s\ (kort of lang). Zo kunnen er in spraak bijv. geen en/of zeer korte van zulke stille en gevulde pauzes zijn of bijv. juist zeer veel en/of zeer lange van zulke stille en gevulde pauzes zijn.
- het spreektempo: de snelheid van spreken, bijv. zeer snel of zeer langzaam.
- het gebruik van herhalingen en correcties. Een herhaling is als iemand twee keer (gedeeltelijk) hetzelfde zegt, een correctie is als iemand zichzelf verbetert door (gedeeltelijk) opnieuw te beginnen. Zo kunnen er in spraak bijv. geen of juist zeer veel herhalingen en/of correcties zijn.

Geef je oordeel aan op een schaal met links 'volstrekt niet vloeiend' en rechts 'volstrekt vloeiend'.

Er zijn spraakfragmenten van zowel moedertaalsprekers als niet-moedertaalsprekers van het Nederlands. De geluidskwaliteit van de spraakfragmenten kan verschillen, maar laat je hierdoor niet afleiden.
Beluister een fragment helemaal en geef DAARNA pas je oordeel. Als je al eerder klikt, wordt dit niet gerekend als je oordeel maar verschijnt de melding: 'Te vroeg!'. Bovenin het scherm loopt een tellertje af dat aangeeft hoeveel fragmenten er nog volgen. Halverwege krijg je de gelegenheid om even te pauzeren. Blijf wel tijdens deze pauze gewoon in de luistercabine. Het hele experiment duurt ongeveer 40 min. Na afloop vragen we je een korte vragenlijst op de computer in te vullen over het experiment.;

Eerst volgen er zes oefenfragmenten. Hierna krijg je de mogelijkheid om vragen te stellen over het experiment.

(practice phase)

Dus om het kort te herhalen: Wat je moet doen is de fragmenten beoordelen op vloeiendheid.
Baseer je oordeel telkens op:
- het gebruik van pauzes: bijv. geen en/of zeer korte stille en gevulde pauzes of juist zeer veel en/of zeer lange stille en gevulde pauzes.
- het spreektempo: de snelheid van spreken, bijv. zeer snel of zeer langzaam.
- het gebruik van herhalingen en correcties: bijv. geen of zeer veel.

Bedankt en veel succes!

(test phase)

Instructions – group of accent raters

Welkom bij dit luisterexperiment!

In dit experiment kijken we naar het accent bij sprekers van het Nederlands. Een spreker heeft een accent als zijn uitspraak in bepaalde mate afwijkt van de Nederlandse norm, van de standaardtaal.
Jouw taak is om spraakfragmenten te beluisteren en te beoordelen op accent. Baseer je oordeel telkens op de uitspraak van bepaalde klanken, de woordklemtonen en de intonatie van de zin.

Geef je oordeel aan op een schaal met links 'geen accent' en rechts 'zeer sterk accent'.

Er zijn spraakfragmenten van zowel moedertaalsprekers als niet-moedertaalsprekers van het Nederlands. De geluidskwaliteit van de spraakfragmenten kan verschillen, maar laat je hierdoor niet afleiden. Beluister een fragment helemaal en geef DAARNA pas je oordeel. Als je al eerder klikt, wordt dit niet gerekend als je oordeel maar verschijnt de melding: 'Te vroeg!'.Bovenin het scherm loopt een tellertje af dat aangeeft hoeveel fragmenten er nog volgen. Halverwege krijg je de gelegenheid om even te pauzeren. Blijf wel tijdens deze pauze gewoon in de luistercabine. Het hele experiment duurt ongeveer 40 min. Na afloop vragen we je een korte vragenlijst op de computer in te vullen over het experiment.;

Eerst volgen er zes oefenfragmenten: oefen met deze fragmenten de beoordelingstaak. Hierna krijg je de mogelijkheid om nog vragen te stellen over het experiment.

(practice phase)

Dus om het kort te herhalen: Wat je moet doen is de fragmenten beoordelen op accent.
Baseer je oordeel telkens op de uitspraak van bepaalde klanken, de woordklemtonen en de intonatie van de zin.

Bedankt en veel succes!;

(test phase)

Appendix III: Questionnaire

PPN nummer : _____

Hier zijn twee vragen over het experiment waaraan je net deelgenomen hebt:

1. Waarop heb je vooral gelet tijdens het experiment? *(een of meerdere dingen?)*

- ☐ de vloeiendheid
- ☐ het spreektempo
- ☐ de pauzes
- ☐ de herhalingen en correcties
- ☐ de grammaticale fouten
- ☐ de woorden die de sprekers gebruiken
- ☐ het accent
- ☐ andere: _____

2. Welke ANDERE factor(en) heeft/hebben jouw oordeel mogelijk beïnvloed tijdens het experiment?

- ☐ de vloeiendheid
- ☐ het spreektempo
- ☐ de pauzes
- ☐ de herhalingen en correcties
- ☐ de grammaticale fouten
- ☐ de woorden die de sprekers gebruiken
- ☐ het accent
- ☐ andere: _____

Hier volgen nog een paar vragen over jezelf:

1. Hoe zou je je taalvaardigheid (schrijven, lezen, spreken en luisteren) in de onderstaande talen in het algemeen beoordelen? (1=geen kennis; 5=heel vloeiend)

	1	2	3	4	5
Duits	☐	☐	☐	☐	☐
Engels	☐	☐	☐	☐	☐
Frans	☐	☐	☐	☐	☐
Turks	☐	☐	☐	☐	☐
Chinees	☐	☐	☐	☐	☐

2. Hoe vaak spreek je Nederlands met ….? (1=nooit; 5=heel vaak)

	1	2	3	4	5
Chineestaligen	☐	☐	☐	☐	☐
Turkstaligen	☐	☐	☐	☐	☐
Engelstaligen	☐	☐	☐	☐	☐
Franstaligen	☐	☐	☐	☐	☐
Duitstaligen	☐	☐	☐	☐	☐

3. Ik waardeer het als …… moeite doen om Nederlands te spreken (1=helemaal NIET mee eens; 5=helemaal mee eens)

	1	2	3	4	5
Franstaligen	☐	☐	☐	☐	☐
Chineestaligen	☐	☐	☐	☐	☐
Duitstaligen	☐	☐	☐	☐	☐
Engelstaligen	☐	☐	☐	☐	☐
Turkstaligen	☐	☐	☐	☐	☐

4. Ik vind dat alle …… die in Nederland wonen, goed Nederlands zouden moeten spreken (1=helemaal NIET mee eens; 5=helemaal mee eens)

	1	2	3	4	5
Engelstaligen	☐	☐	☐	☐	☐
Franstaligen	☐	☐	☐	☐	☐
Turkstaligen	☐	☐	☐	☐	☐
Chineestaligen	☐	☐	☐	☐	☐
Duitstaligen	☐	☐	☐	☐	☐

Persoonlijke gegevens:

1. Geslacht:
   ☐ Mannelijk
   ☐ Vrouwelijk

2. Leeftijd: _____

3. Geboorteland: _____

4. De postcode (zonder letters) van de plaats waar je het grootste deel van je leven gewoond hebt: _____

5. Wat is je moedertaal? _____

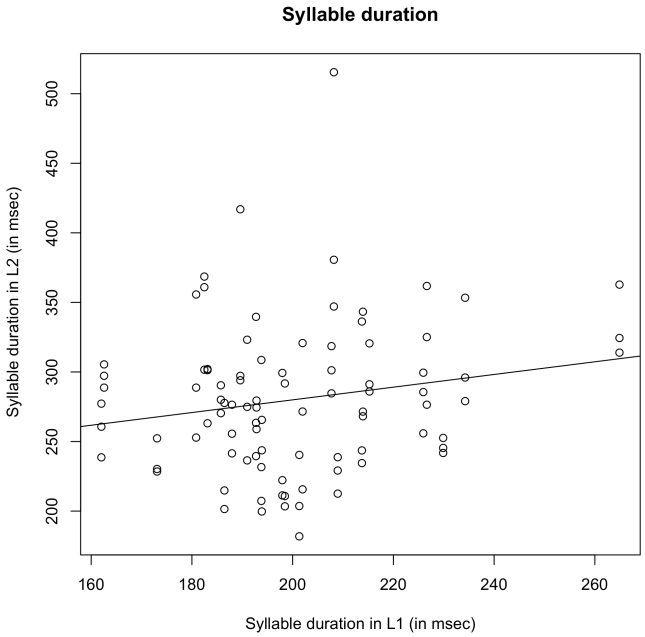
6. Heb je een accent als je Nederlands spreekt?
   ☐ Ja
   ☐ Nee
   Zo ja, welke? _____

7. Heb je gehoorproblemen?
   ☐ Ja
   ☐ Nee
   Zo ja, welke? _____

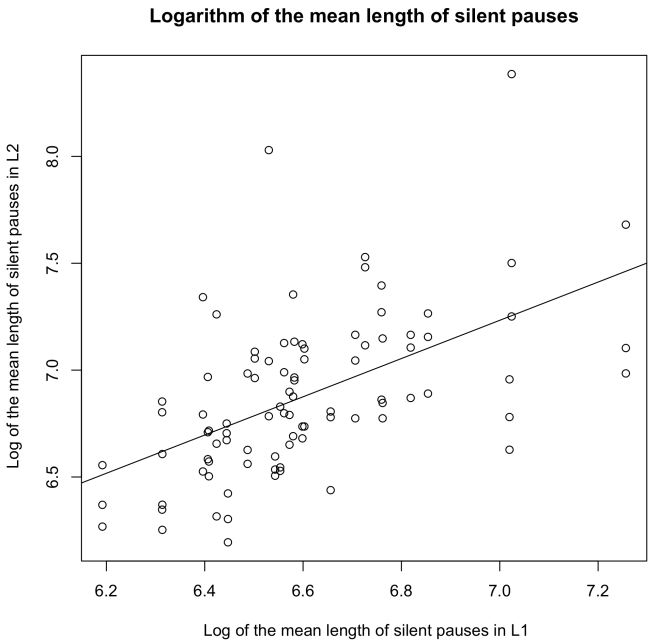
8. Heeft je beroep of je opleiding enige relatie met het vreemdetalenonderwijs?
   ☐ Ja
   ☐ Nee
   Zo ja, welke? _____

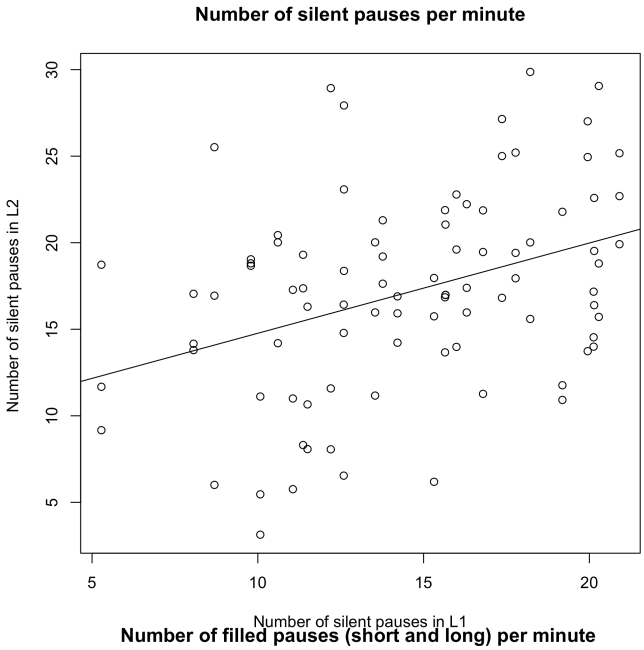
Appendix IV: L1-L2 graphs for objective measures of fluency

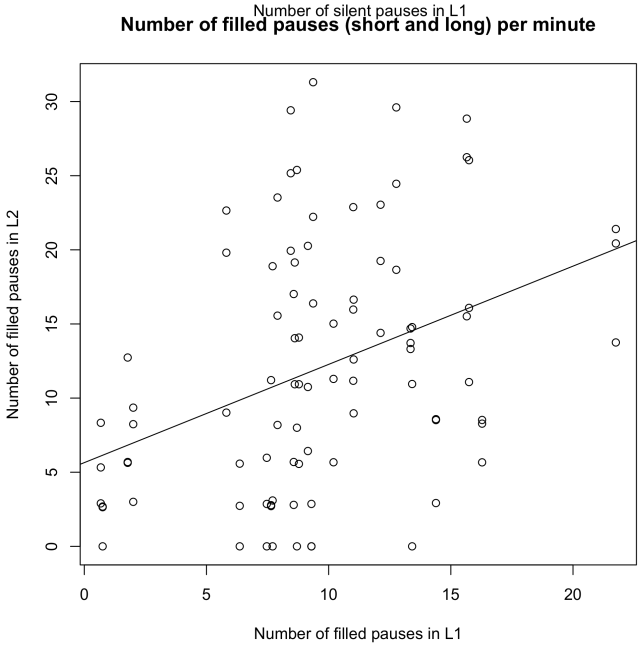
Speed fluency

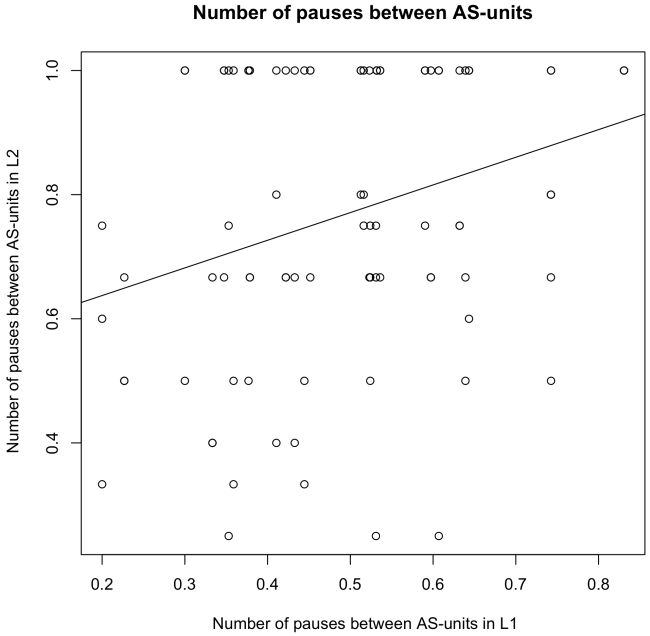


Breakdown fluency

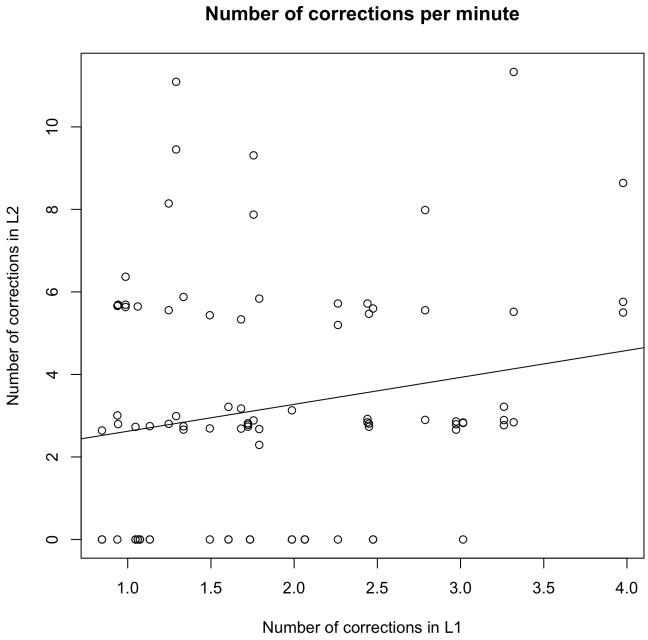


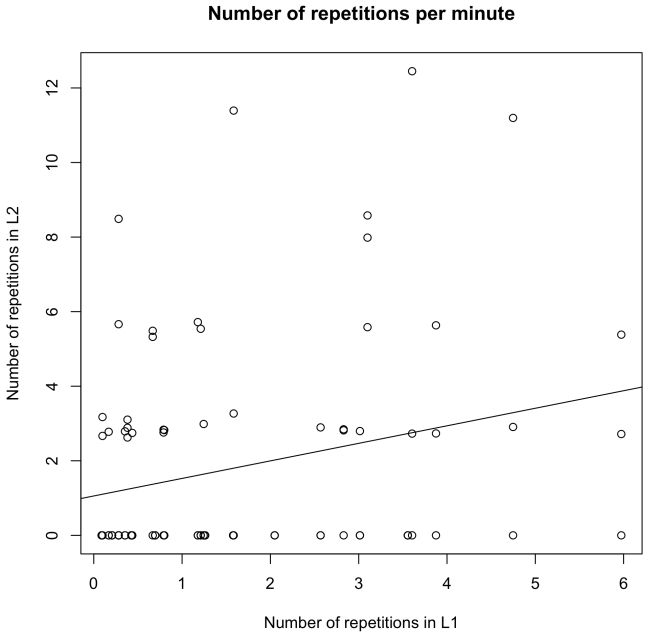

Number of silent pauses per minute
Number of silent pauses in L2
Number of silent pauses in L1
5
10
15
20
25
30


Number of filled pauses (short and long) per minute
Number of filled pauses in L2
Number of filled pauses in L1
0
5
10
15
20
25
30


Number of pauses between AS-units
Number of pauses between AS-units in L2
Number of pauses between AS-units in L1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1.0

Repair fluency





Appendix V: L1-L2 models for objective measures of fluency

MSD

	Estimate	SE	t value	p value	Sig.
(Intercept)	188.729	52.038	3.627	< .001	***
MSD	.456	.258	1.765	.081	
Adjusted R-squared		.023			

Number of P (between AS)

	Estimate	SE	t value	p value	Sig.
(Intercept)	.548	.082	6.714	< .001	***
Number of P	.446	.159	2.802	.006	**
Adjusted R-squared		.072			

Silent P/min

	Estimate	SE	t value	p value	Sig.
(Intercept)	9.553	2.102	4.546	< .001	***
Silent P/min	.522	.140	3.739	< .001	***
Adjusted R-squared		.127			

Filled P/min

	Estimate	SE	t value	p value	Sig.
(Intercept)	5.647	1.858	3.039	.003	**
Filled P/min	.663	.175	3.798	< .001	***
Adjusted R-squared		.131			

MLP

	Estimate	SE	t value	p value	Sig.
(Intercept)	.969	.976	.992	.324	
MLP	.895	.148	6.056	< .001	***
Adjusted R-squared		.286			

Cor/min

	Estimate	SE	t value	p value	Sig.
(Intercept)	1.970	.735	2.680	.009	**
Cor/min	.653	.356	1.833	.070	
Adjusted R-squared		.026			

Rep/min

	Estimate	SE	t value	p value	Sig.
(Intercept)	1.056	.432	2.445	.017	*
Rep/min	.471	.193	2.439	.017	*
Adjusted R-squared		.052			