Oral fluency in a second language: A research agenda for the next ten years.

Pauline Foster

St. Mary's University

This paper reviews how the construct of oral fluency in a second language (L2) has been defined and researched over the last twenty-five years. The emerging picture is somewhat kaleidoscopic, as domains of cognitive, social, individual and linguistic influences on L2 speech have been opened up for study. L2 fluency research presents a wealth of directions for future exploration, five of which have been laid out here as an achievable, though not comprehensive, agenda for the coming decade. Four of these studies focus on the relationship between variables such as perceived fluency, utterance fluency, idiomaticity, task familiarity, vocabulary size and learner self-reflection, while the fifth focusses on supporting fluency development in the L2 classroom. As an aid to prospective researchers, the five studies are laid out in practical detail, together with suggestions on how these might need to be altered to fit the local context in which the research is undertaken

1. Introduction

Oral fluency is no simple matter to define. Filmore's (1979) attempt comprises not just psycho-motor and cognitive dimensions (speed, smoothness, coherence) but also social, emotional and aesthetic dimensions (appropriateness and creativity), the full set of which would only be found in an idealised and very eloquent first language user. By contrast, in common parlance a speaker of a second language (L2) is deemed 'fluent' simply by being good at it. (Hern 2018 is a recent lay example.) This equivalence of fluency with global L2 proficiency (Lennon, 2000) is too vague a definition for research purposes, and in much of the recent research literature it occupies a restricted space, teamed up with complexity, accuracy and lexis in a quartet of spoken performance indicators known as CALF. In this context, fluency refers to surface smoothness in speaking, born of an ease in mobilising linguistic knowledge under the pressure of real-time processing. This is Lennon's (ibid) 'narrow' definition of fluency, and the one we will encounter most in this paper.

CALF measures capture a profile of L2 ability at one point in time, or across time. As learners become more proficient, they typically display greater syntactic and lexical variety, make fewer grammatical and semantic errors, and increase the smoothness of their delivery (Kormos, 1999; Schnadt & Corley, 2006; van Hest, 1996). But it is not warranted

to expect across-the-board improvement in CALF measures because learners can become faster and smoother in deploying L2 knowledge that remains restricted or inaccurate. Increased fluency with flat-lining accuracy is a typical characteristic of fossilisation (Schmidt, 1983). But L2 fluency is not a stable feature of a learner's performance, being susceptible to a host of influences. These are: cognitive load (Foster & Skehan, 1996; Kormos, 2000; Skehan & Foster 1997); individual differences in speaking style (De Jong, Steinel, Florijn, Schoonen & Huilstijn, 2012; Riazantseva, 2001); and deliberate choices in interactional contexts where speed of expression would be inappropriate. Fluency is thus exceptional in the CALF quartet by its association with both progress and lack of progress in an L2, and its susceptibility to a variety of internal and external influences.

2. Theoretical and empirical perspectives on L2 fluency.

It is useful to situate any exploration of fluency inside a theoretical background. For this Levelt's (1989) model of speech processing has the twin benefits of being accessible and widely accepted. It describes three stages through which thoughts become speech. The *Conceptualiser* is a pre-linguistic stage wherein ideas to be expressed are conceived; they pass to the *Formulator* which encodes them, i.e. dresses them in appropriate lexical and grammatical structures. The *Articulator* translates these into a phonetic plan for the vocal tract and facial muscles to act upon. Importantly from a fluency perspective, the components work in parallel. While a pre-linguistic idea is being hatched, the previous idea is being formulated, and another that has been hatched and formulated is in the process of being articulated. At any one time, all three components are active, rather than standing idle for the others to catch up. This is not a robotic production line, however. All components are subject to quality control by the speaker, who may choose to abandon and re-work an idea or a formulation at any point before, during or after articulation, and this is manifested as an interruption or slowing down in one or more of components of the model, resulting in stretches of serial rather than parallel processing.

In first language (L1) performance, speakers have automatised knowledge of grammatical forms, and rapid retrieval of items from an extensive, cross-referenced, idiomatic lexicon. They have effortless control over their vocal tract to articulate the phonemes of their language(s). Consequently, they do not switch to serial processing because of lack of linguistic knowledge or skill. They might be struggling with the pre-linguistic formation of concepts, or deciding to rephrase an utterance because it does not express their thought

well enough, or they might be correcting slips of the tongue, or inserting pauses where the context demands it. Seamless L1 fluency is not just unusual; it's unnecessary and undesirable, suggestive of glibness, insensitivity, and an overbearing attitude towards an interlocutor (Foster, 2013). Deliberate hesitations, pauses or repetitions are ways of showing helpfulness, respect, caution or polite emphasis. L2 speakers behave with the same human instincts; they can be disfluent because of a change of mind about what to say, or to be accommodating to an interlocutor, or because they are sensitive to the context of the interaction. Nevertheless, and especially at lowers levels of proficiency, a combination of lack of L2 knowledge, a small and less integrated L2 lexicon, articulatory difficulties, and overloaded attentional resources are likely to lie behind a disfluent L2 performance.

Research into L2 fluency has developed a plethora of dependant variables relating to speed, breakdown and repair phenomena (Skehan, 2003). These include length and frequency of filled or unfilled pauses, end- and mid-clause pausing, speech and articulation rates, pace, phonation/time ratio, mean length of run, number of false-starts, reformulations, self-corrections, lexical replacements, and repetition of word-initial phonemes, syllables, whole words and even whole phrases. Kormos (2006) applying Levelt's model to an L2, concluded that as the Formulator draws on declarative L2 knowledge, it necessarily acts more slowly until that knowledge had been automatised. She suggests that incidence of end-clause pausing reflects how easily the Conceptualiser is creating a pre-verbal message, while incidence of mid-clause pausing reflects the ease (or difficulty) with which the Formulator is acting upon it. Speed of articulation is evidence of how far L2 declarative knowledge has been automatised, and incidence of repair measures reflects an L2 speaker's monitoring of speech for error. To this one might add that a smaller and less integrated L2 lexicon would contain fewer lexicalised chunks or sentence stems (Pawley & Syder, 1983) both of which support more fluent formulation and articulation (Tremblay, Derwing, Libben & Westbury, 2011). Additionally, lack of practice with pronouncing phonemes outside their L1 range will involve L2 speakers in slower and more effortful operation of the vocal tract, as will lack of familiarity with typical phonological reductions (e.g. the schwa in English). Hence, an increase in articulation rate, a decrease in and migration of pauses to clause-final position, and a decrease in incidence of repair measures, are all suggestive of L2 speakers' progress in automatizing L2 knowledge and expanding their lexical and syntactic resources. In such a

view, becoming more fluent reflects growing proficiency and the smoother working of Levelt's model.

Segalowitz (2016) adds another perspective to considerations of L2 fluency. He sees it in terms of three domains: cognitive, utterance and perceived. Cognitive fluency is a measure of the speaker's ability to turn ideas into speech; utterance fluency is a measure of the speech itself; and perceived fluency is a listener's assessment of speaker's ease in turning thoughts into speech. Segalowitz notes that utterance fluency is the domain that has received by far the most research attention, being most amenable to quantitative analysis of the dependent variables that capture speed, repair and breakdown. Such analyses lend themselves readily to cross-sectional and longitudinal investigations of utterance fluency, and through comparisons with similar studies, to the possibility of generalisations. They also lend themselves to examination of the impact on fluency of a variety of independent variables such as L1 background, L2 proficiency, intelligence and ethnic group identity. However, with such a wide number of measures of utterance fluency on offer, and with no consensus on which capture it best, cross-study generalisations remain of limited validity (Guz, 2015).

Perceived fluency is an interesting addition to the discussion; speech which may be characterised by measurable disfluencies when subjected to careful transcript analysis may not have appeared to be disfluent to the listener in real time. The judicious use of lexical fillers (*kind of, you know, as a matter of fact, to be honest*) can maintain the impression of comfortable fluency while buying the speaker necessary time to finish an utterance. In some contexts, listeners do not necessarily notice pauses, even ones lasting several seconds (Butcher 1980:334), and more recently Préfontaine (2013) has shown that, in French at least, they tolerate long pauses if they come at syntactic boundaries rather than between them. Rossiter (2009) showed that accuracy in pronunciation, grammar and word selection affected listeners' judgement of L2 fluency. Saito, Ilkan, Magne, Tran and Suzuki (2018) investigated the relationship between perceived and measured fluency, and found that in assessing L2 speech, native speaker judges used speed (articulation rate) as the primary indicator of fluency, and pausing (mid- and end-clause) as a secondary indicator, but paid no significant attention to incidence of repairs (repetitions and self-corrections). As long as speech is delivered at a speed they are used to, and with the pausing largely behaving as oral punctuation marks, listeners are tolerant of repairs and do not regard them as disfluencies.

As much of the research referenced above is of a psychological or psycholinguistic nature, it could appear to have little direct relevance to L2 pedagogy or assessment. Teachers may be used to thinking of fluency in Lennon's (2000) broad sense of global L2 proficiency, and indeed as something which does not need to be taught *per se* but which emerges as a testament to the successful learning of and practice in using L2 grammatical and lexical structures. (Rossiter, Derwing, Minintin, & Thompson, 2010; Tavakoli & Hunter, 2018.) From an oral testing perspective, an L2 performance characterised by disfluencies might reflect an individual's idiosyncratic style as much as inadequacy in L2 knowledge or skill. There is evidence (De Jong, Groenhout, Schoonen. & Hulstijn, 2015; Derwing, Munro, Thompson & Rossiter, 2009; Raupach, 1980) that shows significant correlations between L1 and L2 fluency, with Derwing et al. (p533) concluding that fluency must be conceived of as an underlying trait, and not a simple indication of proficiency. Exploring what such a trait might be founded upon, Zuniga and Simard (2019) looked at incidence of self-repair in 58 participants with English L2 and French L1, carrying out picture-cue narration tasks that were designed to be similar in terms of plot, length and vocabulary. They did one task in French and the other in English, in counterbalanced order to neutralise any practice or language effect. The participants were also cloze-tested for their English proficiency, and their attentional control was measured through a Divided Attention Test which assessed how well individuals perform two tasks simultaneously. Results showed a strong positive correlation between L1 and L2 repair behaviour ($r = .613$), similar to the correlation scores obtained by Derwing et al. (2009). Further hierarchical regression analyses revealed that attentional control and L1 self-repair behaviour explained 40% of the variance with L2 self-repair behaviour, while proficiency level contributed very little. The authors concluded that stable individual traits are more closely linked to L2 self-repair than proficiency: speakers who are prone to self-repair in their L1 will be prone to it also in their L2, and this will not be shifted by increasing L2 proficiency. In other words, self -repair in a learner's L2 performance is not susceptible to classroom intervention.

In summation, the independent variables that affect L2 fluency in speech can be cognitive, social, physical, as well as linguistic. A speaker may be slow with generating a pre-verbal message and alter it before, during and after putting it into words. This can be related to an idiosyncratic speaking style, underlying individual traits, the subject matter, the social context, or the stakes of the situation. A speaker may insert disfluencies for the benefit of the listener, to aid comprehension, soften a blow, to show sympathy or delicacy, or to

focus attention on something important. On the linguistic side, disfluencies can reflect L2 knowledge that is not automatised, or is lacking, requiring laborious paraphrasing with the resources at hand. Finally, insufficiently skilful control of the vocal tract may mean L2 words take longer to articulate, especially if the L2 lexicon stores them largely as single items rather than frequently used chunks. It is a quite challenge for fluency research to pick these variables apart.

3.0 A Research Agenda.

The research task suggestions set out below cannot and do not reflect the current richness of L2 fluency research. They are a selection of ideas distilled from a wide choice of past and current publications, but not a comprehensive one. A discerning reader will at once see gaps. There is, for example, no task here that deals directly with cognitive fluency, nor is there one that tackles the confound of eloquence and fluency which characterises the highest-level descriptor bands in some oral exam formats. Also, there is very little here that addresses social or emotional influences on L2 speaking. But the tasks offered are not meant to be taken as prescriptive guides to the research studies they envisage; they can be used as stimuli for researchers to explore tangent lines of enquiry. Because research studies always need a context, it will be necessary anyway for researchers to adjust the suggested designs towards something that is more practical for their local situation and available resources. This will include tweaking the tasks towards other L2s than English, towards contexts with little access to funding or equipment, and towards recruiting participants of a variety of ages, proficiencies and motivations.

3.1 Task one

What is the relationship between perceived L2 fluency and idiomaticity of use?

Usage-based perspectives of language acquisition (e.g. Ellis, 2002; Ermen & Warren, 2000; Pawley & Syder 1983) suggest that the perception of speaker fluency is to a greater or lesser extent a perception of idiomaticity, i.e. the selection of strings of words that are not idiosyncratically constructed for the nonce, but which arise through colligational and collocational bonds, held in common by a speech community, and intuitively acquired by a learner's frequent interactions with others over a prolonged period of time. Erman and Warren have calculated that as much as 55% of a speaker's output is pre-fabricated to

some degree, and there is evidence (Bybee, 2002; Gregory, Raymond, Bell, Fosler-Lussier & Jurafsky, 1999) that L1 phonetic reductions typically found in the oral production of collocations arise from them being processed as single neuromotor units.

A number of research studies into L2 fluency (Di Silvio, Diao & Donovan, 2016; Marriott, 1995; Mőhle & Raupach, 1983; Siegal, 1995; Towell, Hawkins, & Bazergui, 1996;) report learners who had spent some time living in the target language community were more fluent upon their return than their pre-study abroad selves, or than otherwise equivalent groups who had not gone abroad. These studies looked at the effect of immersed and prolonged exposure to the target language, but were not designed to measure the extent to which the increased fluency of the learners was related to increases in idiomatic usage (e,g. collocations and other multiword sequences) though Towell et al. (1996) suggested these could be the key factor.

It would be very useful to pursue the question of idiomaticity and fluency by recourse to a corpus of spoken L2 performance collected for test purposes. For example, the corpus used by Tavakoli and Uchihara (2019) is made up of recordings of the speaking component of the standardized Test of English for Educational Purposes (TEEP, n.d.) employed by the University of Reading to measure applicants' speaking proficiency before they begin their studies. Other schools or universities across the world are likely to have a similar resource of spoken test recordings collected for the same purpose of assessing student oral ability in the language of instruction, whatever that might be. The Spoken Language Corpus (n.d.) complied by Lancaster University and Trinity College, exists expressly for research purposes, comprising Trinity test performances of English learners at a range of levels from advanced to lower intermediate. What is necessary in any such corpus is that the spoken performances will have been graded for levels of *perceived oral fluency*. The research task envisaged here would take from such a spoken corpus a random sample of such test performances, across all proficiency levels of perceived oral fluency. The size of the sample will be constrained by the resources of time and money available to the researcher, but the larger, the better. The anonymised data can be transcribed and analysed for incidence of formulaic language use, i.e. "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray, 2002:9).

Analysis for incidence of formulaic usage in the transcripts can be done by machine searches that look for incidence of contiguous chunks of words (ngrams). For example, Garner and Crossley (2018) used TAALES, or Tool for the Automatic Analysis of Lexical Sophistication 2.0 (Kyle, Crossley & Berger, 2018), as did Tavakoli and Uchihara (2019). But, while ngram analysis is undoubtedly a fast instrument, and as such is very convenient, it suffers from over-reporting. As Stengers, Boers, Housen and Eyckmans (2011:15) note, "Corpus statistics do not always generate word strings which coincide with people's intuitions about what constitutes a formulaic sequence". Indeed, two- or three-word chunks such as *for the*, *and it*, *in the only* are very frequent in English language texts of all kinds, and would be flagged up by a ngram analysis, but they are clearly not formulaic sequences. An ngram analysis is susceptible to text subject matter. As Tavakoli and Uchihara (2019) found, the subject matter of their text threw up trigrams such as *sites for leisure* and *important for national* which are also not formulaic. Automatic ngram searches can also under-report. Words with an obvious idiomatic relationship are not always contiguous, having other words inserted between them. For one example, an ngram search of *turkeys voting, in all their sad millions, for Christmas* would miss the idiomatic relationship between *turkeys, Christmas* and *voting*, whereas a human rater with a knowledge of nativelike selections in English would spot it at once.

Analysis by intuition is a sharper tool than its machine equivalent, even though labour-intensive and time-consuming (Foster, 2001, Stengers et al, 2011). It yields valuable and powerful data because it draws on intuitive knowledge of unfathomable depth. It is carried out by a group of trained raters, working independently, who read the research transcripts and mark out the language which, in their L1 judgement, had not been constructed word by word, but which is fully or partially fixed in memory. Once such an analysis is complete, it must be checked through interrater reliability such that only the words identified by all of the raters are taken as having been identified as formulaically related. At this point, the degree of formulaicity in performance can be expressed as, for example, the percentage of total words that are in in wholly or partially prefabricated sequences. The relationship of this score to the participants' perceived fluency score can then be computed, and as the participants are from a range of proficiency levels, the developmental path of idiomaticity across levels could open for exploration.

Taking this task a little further, as in Tavakoli and Uchihara (2019), other utterance fluency measures of repair, breakdown and speed can also be computed and then related to both the

participants' test score for perceived fluency and their measure of idiomaticity. Such an analysis would throw idiomaticity into the mix of Saito et al.'s (2018) conclusion that, in the ear of the listener, speed is the primary indicator of fluency, pausing is the secondary indicator, while incidence of repair has no bearing at all. Additionally, practical teaching and testing implications could be explored. To what extent is it feasible to work on developing idiomaticity in classroom learners? To what extent is it valid to give implicit credit in an oral test for a dimension of language which is sensitive to someone's context of learning?

3.2 Research Task Two

What is the relationship between speakers' familiarity with the content of the speaking task, their utterance fluency and their perceived fluency?

As set forth in Levelt's model, speaking is a tripartite process: conceiving ideas to be expressed (content), selecting appropriate grammatical and lexical structures (form), and then articulating these structures (performance). For this to be a smooth process, content, form and performance demands must not exceed the speaker's limited attentional capacity. In terms of L2 task performance, Skehan (2003) proposes a trade-off: when the demands of a speaking task exceed attentional capacity, an L2 speaker will prioritise one CALF dimension of performance over another. A speaking task will be more fluently performed if the learner can use off-line planning time (Ellis 2009) or can think through the content before having to start talking about it, and this effect is seen to be greatest on more cognitively demanding tasks (Foster & Skehan 1996; Mehnert, 1998; Skehan & Foster,1997; Wigglesworth, 1999.) Such off-line planning can also take the form of a task rehearsal or repetition (Lynch & Maclean, 2000). Additionally, speakers can find and exploit on-line planning time, especially in an interactive task when they have the role of listener to a partner's turn. (Ellis, 2009). Another variable that lessens the cognitive load of a task is content familiarity; when speakers know well the subject matter of a task, they need to give less of their attention to conceive ideas about it, and this allows them more attention to form, and potentially greater fluency and accuracy in performance.

Bui and Huang (2018) designed a within-participants study to test the effect of content familiarity on fluency. They recruited 58 participants with the same L1 (Cantonese), of broadly similar experience of studying English L2 (12-15 years of classroom teaching) and comparable proficiency (i.e. B2 on the Common European Framework of Reference, 2001.). The participants were studying either nursing or computer science at a Hong Kong

University, and were asked to do two very similar tasks involving a discussion of a computer virus, and a biological virus. Their performances on both tasks were transcribed and coded for a comprehensive list of nineteen measures of speed, breakdown and repair fluency. Results showed that the participants' performances were indeed significantly influenced by how well they knew the topic; content familiarity was related to increased speed and decreased mid-clause pausing. As the authors themselves note, one shortcoming of their study is that it does not include any measure of perceived fluency, so it is not known if a listener would also judge the fluency of the performances as significantly different, and along the same lines.

It would be useful to do a close replication of this neatly designed study with the addition of a measure of perceived fluency, using trained raters to assess the tasks performances according to the fluency descriptors of International English Language Testing Suite (n.d.) or Cambridge First Certificate in English (n.d.), or another made for the nonce. In this way, an important question could be addressed. In the ears of the listener, does content familiarity impact on qualitative judgements of perceived fluency in the same way that content familiarity impacts on quantitative measures of utterance fluency? Other close replications could follow, changing the tasks to other designs paired for degree of familiarity of content, or changing the level of L2 proficiency of the participants. If a number of these altered variables return interesting results, an approximate replication of Bui and Huang could follow that involves them all. Ultimately, these replications could illuminate what raters are reacting to in assessments of speaker fluency, and how far this is allied to L2 proficiency or the appearance of content expertise. This in turn might pose validity questions for designers of speaking tests and reliability questions for the design of rating bands.

3.3 Research Task Three: What can learners' reflections on episodic disfluencies contribute to an understanding of their L2 speech processing?

Research Task Two involved the perspectives of L2 performance afforded by using the external measures of speakers' utterance fluency and perceived fluency. To these we can add learners' own perceptions of how fluent they are. This is an interesting dimension to consider for at least two reasons: to see how far learners' subjective ratings match up to objective performance measures, and to explore from a pedagogic angle what learners identify as holding them back from being able to express themselves with greater ease. Préfontaine's (2013) study explored the former. She asked 40 mixed-ability adult learners of French to

select from a list of six statements the one that best reflected what they felt they were able to achieve in the language. These statements corresponded to the Common European Framework of Reference (CEFR) band descriptors (2001:28-29). The lowest and least fluent level equated to CEFR band A1, ("*I can manage very short isolated utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.*"), while the highest and most fluent level equated to CEFR band C2, ("*I can express myself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.")* Having assessed themselves globally in this way, the participants carried out three monologic narrative tasks, after each of which they rated their own performance using a sliding scale on eight fluency variables, such as *smoothness, rhythm*, and *reasonable speed*. Three native-speakers of French then listened to the recordings of the task performances and used the same CEFR band descriptors to give a global assessment of each speaker's proficiency, in addition to task-specific assessments on the eight fluency measures. In addition to taking measures of perceived fluency from both the speakers' and judges' perspectives, Préfontaine ran the transcribed data through a Praat analysis (De Jong & Wempe, 2009) to compute objective measures of a range of temporal variables, including speech and articulation rates, phonation-time ratio, pausing frequency and length, and mean length of run. Multiple correlation analyses showed that learners' assessment of their L2 skills across all three tasks were significantly similar to those of the native-speaker judges, and also returned moderate to strong correlations between the self-perceived fluency variables and Praat-generated utterance fluency measures, most significantly for mean length of run and average pause time. Préfontaine concluded that not only did L2 learners perceive their fluency in ways that lined up with objective ratings, they also showed themselves to be alert to how the three tasks made subtly different demands on their speech.

One limitation to Préfontaine's study is that it used narrative monologues rather than interactive tasks to generate data; this was necessary for the use of Praat, but is not necessary if a Praat analysis is dispensed with. A further limitation to Préfontaine's design is that the six global fluency descriptors were taken from the A1 – C2 CEFR bands and as such offer the speaker a rather narrow choice of categories in which to fit themselves. The sliding-scale perception on eight fluency measures allowed more flexibility, but these too took single snapshots of an overall task performance. A more process-oriented view of L2 fluency however might see it as something that varies not just from task to task, but from moment to

moment during a task as a speaker encounters different L2 processing or situational demands. The research task proposed here will further explore the insights learners can provide on their own performance, using interactive tasks to generate the data, and a stimulated recall procedure (Gass & Mackey, 2000) to allow learners to give broader and finer-grained feedback.

In contrast to a monologue where one person speaks through the entire task, conversations are a far more common mode of human interaction, in which interlocutors exchange the roles of speaker and listener in a collaborative, social endeavour. Conversation analysis (e.g. Sacks, Schegloff & Jefferson, 1974) looks at how each speaker behaves in such interactions, taking account of what the other is saying and fashioning a response that is timely (neither overlapping nor allowing a discernible pause to open up) and appropriate (i.e. following the Co-operative Principle of Grice (1975) and the socio-pragmatic norms accompanying speech acts such as requesting, and apologising). Maintaining speed, rhythm, and smoothness in such circumstances is likely to be harder than in a monologic task, but is nevertheless part of the profile of being a fluent speaker. The research envisioned here will, in addition to one monologic narrative, put the participants into interactive tasks of such types as an exchange of information, topical discussion, decision-making and role-playing. The tasks will all need to be piloted under multiple examples to weed out any which are not engaging enough to sustain interest. To exclude the possibly confounding influence of different interlocutors, each of the interactive tasks will need to be performed by the same dyads, and in a counter-balanced design that will obviate any task practice effect. But because the data analysis will be qualitative and intensive, the number of participants can remain relatively small and can include a range of proficiencies.

The tasks will be recorded and (ideally immediately) afterwards the recording will be played back to the participants for the stimulated recall procedure. The interactive data will require two researchers as the dyads have to be given a stimulated recall at the same time. During the recall procedure, the researcher will stop the recording at any point and invite the participant to illuminate the thought processes that he or she was experiencing in listening or speaking, with particular regard to ease or difficulty in either role. This will have to be done using the L1 of the participant, so that there is no language barrier to expressing these thoughts as precisely as possible. The goal will be to tease out where L2 processing stumbled and what caused that to happen: was it lack of vocabulary to formulate a thought, provoking a silent mental search for a paraphrase, or was it hesitation over the most accurate grammatical

formulation while various alternatives were pondered, or was it a stumbling over the pronunciation? Was an apparent dysfluency due to a change of mind over what best to say, or a deliberate slowing down to indicate delicacy in delivering a contrary opinion, or asking for a favour? Was it necessary for the speaker to pause because the interlocutor's turn had been hard to process and needed another pass to decode? Was it because the task was cognitively demanding or socially delicate, and required more thought before speaking? Crucially also, the researcher needs to find out in which parts of the recording the interaction was smoothly unproblematic, and why. Was it down to the employment of an appropriate formulaic phrase, or because the subject matter was well enough known and therefore predictable?

These data should yield a rich episodic picture of what sustains L2 fluency, and what acts to derail it at an individual, task-specific level. This in turn could inform pedagogy if it throws up the conclusion that the underlying cause of much dysfluency is the simple lack of vocabulary, and /or individual variation in how highly a speaker or a listener rates being accurate over being nimble in interaction. It could also illuminate how some learners might wish to be taken as fluent listeners to the point of strategically disguising a stretch of incomprehension with enthusiastic and encouraging back-channel signals. Cross-task comparisons will be also be useful if they reveal that a speaker's fluency in one task is not matched by similar fluency in another, and how this might affect the design of tasks for testing purposes. It might also reveal where L2 learners' fluency is sustained or hampered by a knowledge of L2 socio-pragmatic formulae.

3.4 Research Task Four: What is the relationship between productive vocabulary size and fluency performance measures?

It is well established (e.g. Laufer, 1998; Nation, 2001; Read, 2000; Schmitt, 2014) that a person's vocabulary in any language is of two sorts. Words that a person recognises and understands when encountering them in reading or listening comprise a receptive (passive) vocabulary. A subset of these is known as a productive (active) vocabulary and comprises words that a person calls into use when speaking or writing. Laufer (1998) makes a further useful distinction: a person producing a word in response to a direct cue (such as being asked to fill a gap in a sentence, or provide an antonym) is drawing on *controlled* productive knowledge, while a person producing a word spontaneously is drawing on *free* productive

knowledge. It can be argued that the more a person can draw on free productive knowledge, the more fluently that person is able to speak.

The relationship between dimensions of L2 vocabulary and aspects of L2 speech proficiency has attracted research interest for some time. Koizumi and In'nami (2013:902) present a useful survey of nine studies carried out between 2000 and 2012 that employed a variety of instruments to measure vocabulary in terms of *size* (how many words), *depth* (semantic, grammatical and collocational information attaching to words), and *speed* (reaction time in lexical retrieval). These measures were then compared to aspects of L2 oral performance in interviews, picture descriptions and other speaking tasks. Vocabulary size was the most popular dimension for investigation, with seven of the studies measuring this through tests of receptive knowledge or controlled productive knowledge. Only one study (De Jong, Steinel, Florijn, Schoonen & Huilstijn 2013) investigated the relationship between oral fluency and size of controlled productive vocabulary, although the size measurement involved display of collocational knowledge, so it included an element of vocabulary depth as well. Using transcript data from a range of descriptive and argumentative tasks, the authors compared the 179 participants' vocabulary size/depth score to their utterance speed, breakdown and repair scores, as measured by a Praat analysis. A range of correlations emerged, from strong (for speed), to moderate (for repair), to very weak (for breakdown).

Koizumi and In'nami's survey highlights the relative paucity of studies into productive vocabulary knowledge and L2 speech, and this may in part be due to concerns about the validity of the tests employed. For example, sentence completion tasks may elicit receptive as well as productive knowledge, may only touch the surface of a person's productive knowledge, or, as in De Jong et al. (2013), combine size with depth. To avoid these concerns about sentence completion tests, Uchihara and Saito (2017) investigated the relationship between L2 oral ability and size of controlled productive vocabulary by means of a measure, Lex30, developed by Meara and Fitzpatrick (2000). The 39 L2 English participants were presented with an on-screen list of 30 English words and invited to type up to four other words they could associate with each prompt word. Scoring was done by each participant's list of answers being compared to the JACET list of 8,000 basic words in English (Committee of revising JACET, 2003) and one point was awarded for each word outside the 1,000 most frequent, giving a raw score out of a possible maximum of 120. In addition, the percentage of infrequent words to total words was calculated for each participant. Speech samples were elicited by showing participants a sequence of seven pictures, and asking them to narrate

what was happening. Only data from the final three pictures was used in analysis, made up of a random 10-seconds taken from each picture description, combined into a 30-second sample, and yielding a mean of 45 words (range 35 -61) for each participant. A global rating for perceived comprehensibility and accentedness was given by five novice native-speaker raters who listened to the samples. They rated the performances on a 1,000-point scale from *hard to understand, heavily accented* (0) to *easy to understand, little accent* (1000). An additional rating for perceived speech rate was given by five expert raters, again on a 1000- point cline between *non-target-like* (0) and *target-like* (1000). Pearson correlation analyses revealed a significant, weak-to-moderate relationship between speech rate and the raw Lex30 scores ($r = .342, p = .033$) but no significant relationship was found for the other oral ability scores in comprehensibility or accentedness. The percentage Lex30 scores did not correlate significantly with any of the other variables, though scores for comprehensibility and speech rate were significantly linked ($r = .518\ p < .01$).

Uchihara and Saito's study is interesting in that it shows a significant relationship between size of controlled productive vocabulary and perceived speech rate, which echoes De Jong et al.'s (2013) finding of a strong and significant correlation between size and speed. A strength of Uchihara and Saito's study is that it measures the size of controlled productive vocabulary in a way that avoids the confound with depth in the method chosen by De Jong et al., while potential limitations are that it used very short speech samples (30 seconds yielding an average of 45 words) from one kind of task, and employed a subjective global assessment of only one dimension of perceived fluency. De Jong et al. (2013) in a very much more extensive study employed eight speaking tasks, involving combinations of three design elements: simple or complex in terms of topic, formal or informal in terms of setting, and descriptive or argumentative in terms of discourse. This yielded 16 minutes of transcribed data which could be machine coded (through Praat) for a variety of measures of breakdown, repair and speed fluency. However, both studies collected oral data through lab-based monologues. In Uchihara and Saito, the participants were in a sound-proof booth, tasked with describing a series of pictures, each with three key word prompts. In De Jong et al. the eight tasks were computer –administered, with participants given on-screen instructions. Each task came with a photo picture to set the scene, and one or two visual-verbal cues. The participants were told to imagine they were in the situation described by the picture(s) on the screen, and depending on the task, had to imagine they were addressing an individual listener or a whole room of them. Neither study used spontaneous interactive speech, and for very

practical reasons. Monologues have the research advantage that in that there is no confounding interlocutor voice to distract a human rater, or render a Praat analysis difficult. However, this does not mean that accommodations cannot be made so that interactive data, the most common speaking environment, can be investigated for fluency. Tavakoli (2016) provides an interesting discussion of this.

The research task envisaged here will further explore the relationship between free productive vocabulary knowledge and a variety of utterance fluency measures. As noted above, free productive vocabulary knowledge is what a speaker draws on in spontaneous speech, therefore a large sample of spontaneous speech is required. The tasks used by De Jong et al. (2013) will produce a large sample (16 minutes), and could be easily adapted so that the participant will still be the main voice, but have a supportive interlocutor providing friendly eye contact, back channel behaviour, and perhaps the occasional encouraging prompt to maintain the interaction. The interlocutor would need to be trained to try to keep his or her contributions from overlapping with the participants' words. For a Praat analysis, all the digital recordings would have to be pruned of these interventions so that only the participant data remains.

Ideally, if productive vocabulary size is going to be compared to speech fluency, it should be measured via a spoken test (Uchihara & Clenton, 2016), especially as there is some question as to whether the written version of Lex30 measures vocabulary *recall* rather than *use* (Baba, 2002; Fitzpatrick & Meara, 2004). When Fitzpatrick and Clenton (2012) compared results for the oral and written version of Lex30 on the same participants, they found a similar mean score (15.6 vs 16.6) and a correlation that was borderline weak, though significant *($r = 0.391$, $p < 0.01$)*. Given that Fitzpatrick and Clenton's (2012) assessment is that Lex30 is overall a valid measure for productive vocabulary knowledge, an oral version of Lex 30, if available, should be used here instead of the written version. For a further analysis, the words from each participant's spoken corpus could be lemmatised (Bauer & Nation, 1993) and compared to the JACET 8000 word list. This will give an indication of what number and what percentage of their freely produced words lie outside of the 1000 (or 2000) most frequent in English. Additionally, it would be possible to measure reaction time to the prompt words to give an idea of lexical fluency, i.e. how quickly a word can be retrieved from memory.

It will be possible from these data to have an idea of each participant's controlled productive vocabulary, free productive vocabulary, breakdown fluency, repair fluency, and speed

fluency, and to explore the size and significance of correlations between them. When exploring correlations of this nature, it is crucial to be able to report high inter-rater reliability in the coding. In terms of strengthening the reliability of the vocabulary test, there would be great benefit in increasing the number of items to make an oral Lex45 or even an oral Lex60, were this possible. Additionally, oral data can usefully be collected from native speakers, as this will give a parallel set of freely-produced words prompted by the eight tasks, and also a native speaker base-line for fluency measures. Interpretations of how fluently or with what vocabulary L2 English learners perform speaking tasks are more valid when set against how L1 users perform them.

3.5 Research Task Five: How can L2 fluency development be supported in the language classroom?

A research agenda for the next ten years would not be complete without including a task that starts from a focus on pedagogy, in contrast to other parts of the research agenda where pedagogical implications might be garnered only after the research is complete.

The importance that communicative language teaching (CLT) places on opportunities for classroom learners to speak is not specifically tied to targeting any one aspect of L2 fluency, rather it is assumed that that fluency as a whole goes hand-in-hand with practice, as declarative knowledge is transformed into procedural knowledge, and laborious production gets faster and less taxing. Plentiful role-plays, discussions, and problem-solving tasks are incorporated into CLT lesson plans to provide that practice. But there are classroom language techniques which research has suggested as supportive of fluency development in a more targeted way. For example, the simple act of performing a speaking task more than once has been shown to lead to fewer instances of repair and breakdown disfluencies in the subsequent repetitions (Foster & Hunter, 2016; Lynch & Maclean, 2001). Spending a few minutes preparing for a speaking task, rather than launching straight into it, has also been found to support a more fluent performance (Foster & Skehan, 1996; Skehan & Foster, 1999). Towell et al. (1996) ascribed the increased fluency in their participants' L2 French to their plentiful exposure to formulaic language after a year living in France. With this in mind, Boers, Eyckmans, Kappel, Stengers & Demecheleer, (2006) trained a class group of L2 learners to look for formulaic sequences in their English L2 reading. After a year of classes, they compared the fluency of this group with the fluency of a control group who had received no

such training, and found that the group whose consciousness of formulaic sequences had been raised was more fluent than the control group.

Rossiter et al. (2010) looked at how language textbooks deal with fluency, taking their local ESL provision as the context (Edmonton, Canada). They examined 28 CLT textbooks and teacher resource guides which claimed explicitly to help develop oral fluency, together with 14 CLT textbooks that had a more general pedagogic focus. They sorted the classroom activities in these books into five categories: free production, use of formulaic sequences, rehearsal/repetition, consciousness-raising and fillers. They found that in the textbooks the most common type of activity was free production, with formulaic sequences and rehearsal/ repetition coming in a close second. Somewhat surprisingly, in the textbooks that advertised themselves as promoting fluency, there were no materials at all that involved the use of fillers, or fluency consciousness-raising. By contrast, all five categories of fluency activity were represented in the teacher resource guides, with consciousness-raising and use of fillers appearing in nearly half of them. The authors assume from their survey that students using their (admittedly small) spread of textbooks will most likely not get instruction in the use of fillers, nor will they have their consciousness raised about fluency, unless the teachers themselves design original materials with these things in mind. Even though the teacher guides offer a wider range of fluency-enhancing activities, it is not known to what extent teachers avail themselves of these to supplement the learner textbooks. Rossiter et al.'s conclusion is that, compared to the focus on grammatical and lexical development, ESL learners in Edmonton are rather short-changed on oral fluency; it is a "neglected component in the communicative language classroom" (p.583). The ubiquity of free-production activities in their survey however suggests that the neglect is benign rather than blameworthy; teachers might well suppose that their learners' fluency development is sufficiently taken care of by general speaking practice, and requires no specific intervention.

It might also be the case that L2 teachers tend to conceive of fluency in Lennon's (2000) 'broad' sense of general proficiency arising from increasing grammatical and lexical knowledge. To explore teachers' understanding of fluency and how it informs their professional practice, Tavakoli and Hunter (2018) administered a questionnaire to 84 language teachers in the UK. It included questions on what they understood by the term L2 fluency, how confident they were in their ability to teach it, and what sort of activities they used to promote its development in their classes. Results showed that while 17% felt their understanding of speech fluency was limited, a comfortable 83% majority reported they

understood it to a large extent or to some extent. When it came to what factors contributed to fluency however, the teachers were less sure, with 29% choosing to describe their knowledge as limited or even non-existent. In terms of the teacher's definitions of fluency, the questionnaire garnered 452 examples, an overwhelming majority of which (345 or 76%). could be classified as relating to ease and confidence in speaking, or L2 proficiency in general. Only 61 definitions (13%) related to Lennon's 'narrow' sense of surface smoothness in speaking, with infrequent pauses or hesitations.

In response to questions about promoting fluency in the classroom, only around 10% of the teachers said that they knew 'to a large extent' how to teach fluency, how to help learners improve their fluency, what activities to use to promote fluency, and what learning strategies improved fluency. Consistent majorities of between 80% and 87% felt they had only 'some' or 'limited' knowledge of these things. Yet a question about overall confidence in helping learners improve speech fluency threw up a rather surprising 24% who were largely confident and 45% who were somewhat confident, indicating that the teachers' confidence was not tied to any specifics of their classroom practice.

The questionnaire invited the teachers to give three examples of what activities they used to promote fluency in their classrooms. This could have given a maximum of 252 (84x3) examples, but 57 (23%) of the slots remained unfilled, revealing that many of the teachers were using only one or two types of fluency-oriented activities. The 195 examples that were supplied were sorted into a (slightly adapted) version of Rossiter et al.'s (2010) five categories of fluency development activities, and were found overwhelmingly (69%) to favour communicative 'free production' such as role-plays, debates, group work tasks and conversations. Striking was the tiny number of classroom fluency activities that could be connected to research studies, such as consciousness-raising types (only 4 mentions), planning and repetition types (only 7 mentions), and discourse strategy types (only 8 mentions). This result is understandable when set against the responses to the question about teachers' knowledge of recent fluency research findings; a third of the respondents said they had hardly any. Encouragingly however, 73% said that they thought research could help improve their teaching practice.

The results of this study (and a similar one in Chile reported by Morrison, 2018) reveal that while the teachers were confident what L2 fluency means, and confident in how to develop it in the classroom, their definition is aligned to general L2 proficiency and the classroom

activities they use are, overwhelmingly, of the general speaking practice sort. Tavakoli and Hunter conclude that, in line with Rossiter et al. (2010:345); "fluency, in its focussed and narrow sense, might very well be neglected in the L2 classroom." This is obviously not an ideal state of affairs, and for it to change fluency researchers and L2 classroom practitioners will need to join forces to explore how to support fluency development (in Lennon's narrow sense) through classroom activities that have both empirical and pedagogic validity.

The studies described above are based on small sample sizes taken from a particular locale. Both therefore offer opportunities for close replications set in different locales. It would for example be interesting to examine how fluency is handled in a set of teaching texts from other parts of the world, and also interesting to explore teacher fluency beliefs and fluency practice in other parts of the world. However, the main study envisaged here is not a replication. It is a longitudinal look at the impact on L2 fluency of activities designed to develop it along more narrowly defined lines than afforded by general speaking practice.

Finding the right setting will be crucial. Ideally, this will be an educational establishment where a foreign language is taught to at least two intact class groups, tested as the same level of proficiency, and taking a course that follows the same CLT-type syllabus, spread over several months or even longer. (The minimum number of participants would have to be determined by calculating the necessary statistical power of such a study, and basing the sample size on that.) By making this a longitudinal study we avoid the limitation of cross-sectional designs that may show no effect at all for a one-off treatment because development is not necessarily something that is immediately detectable. By excluding second languages, we can limit the influence of target language exposure outside the classroom. By using at least two class groups at the same proficiency we can assign classes to an experimental or control condition. By using intact classes, we can be sure the learners are with the same classmates and teacher throughout. By following the classes for a whole course, the initial pre-test and eventual post-test will not be too close together to render any outcome dubious. By having at least 20 learners in each group we have enough participants to allow for some quantitative analyses, and for attrition when inevitably some participants drop out. Importantly, by setting the research in an intact classroom, rather than a laboratory setting with paid participants, we preserve as far as possible the ecological validity of the treatment. Equally important is having the classroom teacher(s) as part of the research team, making sure the design and execution of the treatment is pedagogically practical and defensible. The learners will have paid for the teaching and so cannot be exploited by being asked to perform

a classroom task that has excellent research credentials but which would not pass muster with a teacher.

The pre-test task (A) will be recordings of the learners made at the outset of the course, before any teaching has started. This could be of an interaction between each learner and the teacher, on an informal conversational topic, lasting at least five minutes, though longer is better. The recording can be given to a trained rater for a perceived fluency analysis, and should also be closely transcribed to allow for an utterance fluency analysis of a host of variables pertaining to the treatment. The post-test task (B) would then be a parallel interaction made at the end of the course so that fluency development of the control and experimental groups can be captured and compared. Because it is next to impossible to design two tasks that are completely parallel in terms of content, and because we do not wish task content or task order to be confounding variables, the participants will do the tasks in a counterbalanced order, i.e. half will carry out task A as their pre-test, while the other half will carry out task B as their pre-test; and likewise for the post test, tasks A and B will be presented in a counterbalanced way.

The treatment will be a particular type of fluency instruction that will be inserted into the experimental group's classes, and be absent from the control group's classes. For example, Nattinger and DeCarrico (1992) suggest that L2 learners who learn to deploy idiomatic fillers appropriately (such as '*you know'*, '*well*' and '*I mean'* in English) are able to plug what would otherwise be noticeable breaks in the flow of their speech. To find out if learners respond to a classroom intervention that targets this strategy, the experimental group will be given activities designed to encourage their use. This could be expanded to include idiomatic gambits (Wray, 2002) such as '*actually*' or '*let me see'*, or '*what I am getting at is*…'. These also buy processing time for a speaker who is not sure what thought to conceive next, or how best to formulate it.

The precise design of such intervention tasks will be something for the teachers and researchers to collaborate on, and will be dependent on the precise aspect of fluency they choose to investigate, as well as the kind of activity topics that the learners are familiar with. However, Rossiter et al. (2010) is a very good starting point for anyone looking for practical ideas. They lay out activities that develop learner awareness of formulaic sequences and discourse markers through consciousness-raising, planning strategies, data analysis, textual analysis, interactional management and poster presentations.

As the learners would all be following a CLT syllabus, with equal and plentiful 'free production' activities for both experimental and control groups, increases in fluency for both would be expected to show up in the post-test at the end of the course. The comparison of pre-test and post- test performance would provide insight into whether the class time spent on specific fluency-enhancing interventions had had a sufficiently significant impact to warrant a recommendation to teachers that there is a demonstrably effective *and* practical way to develop their learners' fluency.

References.

Baba, K. (2002). Test review: Lex 30. *Language Testing Update*, *32*, 68-71.

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography, 6,* 253-279.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral fluency: Putting a lexical approach to the test. *Language Teaching Research, 10,* 245-261.

Bui, G., & Huang, Z. (2018). L2 fluency as influenced by content familiarity and planning: Performance, measurement, and pedagogy. *Language Teaching Research, 22(1)*, 94–114.

Butcher, A. (1980). Pause and syntactic structure. In H.W. Dechert & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp 85-90). The Hague: Mouton.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Cambridge First Certificate in English, (n.d.). *Assessing Speaking Performance – Level B2.* Retrieved February 20, 2020 from https://www.cambridgeenglish.org/images/168619-assessing-speaking-performance-at-level-b2.pdf

Committee of Revising JACET Basic Words (Ed.). (2003). JACET list of 8000 basic words. Tokyo: Japan Association of College English Teachers.

Common European Framework of Reference, (2001). Language Policy Unit, Strasbourg. Retrieved February 20, 2020 from https://rm.coe.int/16802fc1bf

De Jong, N.H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385 - 390. https://doi.org/10.3758/BRM.41.2.385

De Jong, N.H., Groenhout, R., Schoonen, R. & Hulstijn, J.H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior, *Applied Psycholinguistics, 36(2),* 223-243.

De Jong, N.H., Steinel, M.P., Florijn, A.F., Schoonen, R., & Huilstijn, J.H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*, 5-35.

De Jong, N.H., Steinel, M.P., Florijn, A.F., Schoonen, R., & Huilstijn, J.H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34,* 893-913.

Dechert, H. W. & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler*. The Hague: Mouton.

Derwing, T., Munro, M., Thomson, R., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31(4),* 533-557.

Di Silvio, F., Diao, W., & Donovan, A. (2016). The development of L2 fluency during study abroad: A cross-language study. *The Modern Language Journal*, *100(3),* 610-624.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24(2)*, 143-188.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity and accuracy in L2 oral production. *Applied Linguistics, 30(4),* 474-509.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20(1)* 29-62.

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

Fitzpatrick, T. & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing, 27(4)*, 537 -554.

Fitzpatrick, T. & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *VIAL: Vigo International Journal of Applied Linguistics, 1*, 55- 73.

Foster, P. & Hunter, A. (2016). When it's not what you do but the way that you do it: How research into second language acquisition can help teachers make the most of their classroom materials. In B. Tomlinson (Ed.), *SLA and Materials development for Language Teaching.* (280-292). New York: Routledge.

Foster, P. & Skehan, P. (1996). The influence of planning time on performance in task-based learning. *Studies in Second Language Acquisition*,*18*, 299-234.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain, (Eds.), *Researching pedagogic tasks: Second language teaching, learning and testing.* (75-93). London: Routledge.

Foster, P. (2013). Fluency. In C.A. Chapelle, (Ed.), *The International Encyclopedia of Applied Linguistics,* Oxford, UK: Wiley-Blackwell.

Garner, J. & Crossley, S. (2018). A latent curve model approach to studying L2 N-Gram development. *The Modern Language Journal, 102*, 494-511.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum.

Gregory, M., Raymond, W., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society, 35*, 151-166.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.). *Studies in Syntax and Semantics III: Speech Acts.* (183-98). New York: Academic Press,

Guz, E. (2015). Establishing the fluency gap between native and non-native speech. *Research in Language, 13(3)*, 230-247.

Hern, A. (2018). 'Becoming fluent in another language as an adult might be impossible – but I'm still going to try.' *The Guardian*. Retrieved August 2, 2019, from https://www.theguardian.com/education/shortcuts/2018/may/02/fluent-another-language-adult-impossible-try

International English Language Testing Suite, (n.d.). Retrieved July 29, 2019, from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

Koizumi, R. & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4(5)*, 900-913.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning*, *49(2),* 303-342.

Kormos, J. (2000). The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition, 22(2*), 145-169.

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

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030-1046.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different*? Applied Linguistics*, *19(2),* 255–271.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: University of Michigan Press.

Levelt, W. (1989). *Speaking: from intention to articulation*. MA: MIT Press.

Lynch, T. & Maclean, J. (2000). Exploring the benefits of task repetition and recycling in classroom language learning *Language Teaching Research 4(3),* 221-250.

Marriott, H. (1995). The acquisition of politeness patterns by exchange students in Japan. In B, Freed (Ed.), *Second language acquisition in a study abroad context*. pp197-224 Amsterdam: John Benjamins.

Meara, P. & Fitzpatrick, T. (2000) Lex30: an improved method of assessing productive vocabulary in an L2. *System, 28*, 19-30.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*, 83–108.

Möhle, D. & Raupach, M. (1983). *Planen in der Fremdsprach*. Frankfurt: Peter Lang.

Morrison, A. (2018). *Fluency in the EFL Chilean classrooms*. British Council ELT Master's Dissertation Award: Winner. https://englishagenda.britishcouncil.org/sites/default/files/attachments/astrid_morrison_university_of_reading_dissertation.pdf

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nattinger, R. & DeCarrico, S. (1992*). Lexical Phrases and language Teaching*. Oxford: Oxford University Press.

Pawley, A. & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). Harlow: Longman.

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review*, *69(3),* 324-348.

Raupach, M. (1980). Temporal variables in first and second language speech production. In H.W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 263-270). The Hague: Mouton.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23*, 497–526.

Rossiter, M. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review/ La Revue Canadienne des Langues Vivantes, 65*, 395-412.

Rossiter, M., Derwing, T., Minintin, L. & Thomson, R. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Review/ La Revue Canadienne des Langues Vivantes, 66(4)* 583-606.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50(4)*, 696-735.

Saito, K., Ilkan, M., Magne, V., Tran, M., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low, mid and high-level second language fluency. *Applied Psycholinguistics, 39*, 593-617.

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

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64(4),* 913–951.

Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. *Language*, *212(2)*, 8-13.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics, 54(2),* 79-95.

Siegal, M. (1995). Individual differences and study abroad: Women learning Japanese in Japan. In Freed, B. (Ed.), *Second language acquisition in a study abroad context*, (225 -244). Amsterdam: John Benjamins.

Skehan, P. & Foster, P. (1997). Task type and processing conditions as influences on foreign language performance. *Language Teaching Research 1(3),* 185-211.

Skehan, P. & Foster, P. (1999). 'The influence of task structure and processing conditions on narrative retellings' *Language Learning 49(1),* 93-120.

Skehan, P. (2003). Task-based instruction. *Language Teaching 36*, 1 -14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30(4),* 510–532.

Spoken Learner Corpus Project (n.d.). Retrieved February, 20, 2020 from www.cass.lancs.ac.uk/css-projects/the-spoken-learner-corpus-slc-project/).

Stengers, H. Boers, F., Housen, A. & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics in Language Teaching, 49(4)* https://doi.org/10.1515/iral.2011.017.

Tavakoli, P. & Hunter, A-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research, 22(3),* 330-349.

Tavakoli, P. & Uchihara, T. (2019). To what extent are multiword sequences associated with oral fluency? *Language Learning, 70(2).* https://doi.org/10.1111/lang.12384

Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, *65(1),* 71-79.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in define and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54(2),* 133-150.

TEEP, (n.d.). *Test of English for Educational Purposes*. Retrieved February 20, 2020 from https://www.reading.ac.uk/ISLI/study-in-the-uk/tests/isli-test-teep.aspx

Towell, R., Hawkins, R. & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics 17(1)*, 84-115.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61(2),* 569–613.

Uchihara, T. & Clenton, J. (2018). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*. https://doi.org/10.1177/1362168818799371

Uchihara, T. & Saito, K. (2017). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *Language Learning Journal 47(1)*, 64-75.

van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg University Press, Tilburg.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14(1)*, 85–106.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zuniga, M. & Simard, D. (2019). Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1. *Journal of Psycholinguistic Research. 48,* 43-50.